medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization

Van Thien Chi Nguyen[1,2#], Trong Hieu Nguyen[1,2#], Nhu Nhat Tan Doan[1,2], Thi Mong Quynh Pham[1,2], Giang Thi Huong Nguyen[1,2], Thanh Dat Nguyen[1,2], Thuy Thi Thu Tran[1,2], Duy Long Vo[3], Thanh Hai Phan[4], Thanh Xuan Jasmine[4], Van Chu Nguyen[5,6], Huu Thinh Nguyen[3], Trieu Vu Nguyen[7], Thi Hue Hanh Nguyen[1,2], Le Anh Khoa Huynh[1,8], Trung Hieu Tran[1,2], Quang Thong Dang[3], Thuy Nguyen Doan[3], Anh Minh Tran[3], Viet Hai Nguyen[3], Vu Tuan Anh Nguyen[3], Le Minh Quoc Ho[3], Quang Dat Tran[3], Thi Thu Thuy Pham[4], Tan Dat Ho[4], Bao Toan Nguyen[4], Thanh Nhan Vo Nguyen[4], Thanh Dang Nguyen[4], Dung Thai Bieu Phu[4], Boi Hoan Huu Phan[4], Thi Loan Vo[4], Thi Huong Thoang Nai[4], Thuy Trang Tran[4], My Hoang Truong[4], Ngan Chau Tran[4], Trung Kien Le[3], Thanh Huong Thi Tran[5,6], Minh Long Duong[5,6], Hoai Phuong Thi Bach[5,6], Van Vu Kim[5,6], The Anh Pham[5,6], Duc Huy Tran[3], Trinh Ngoc An Le[3], Truong Vinh Ngoc Pham[3], Minh Triet Le[3], Dac Ho Vo[1,2], Thi Minh Thu Tran[1,2], Minh Nguyen Nguyen[1,2], Thi Tuong Vi Van[1,2], Anh Nhu Nguyen[1,2], Thi Trang Tran[1,2], Vu Uyen Tran[1,2], Minh Phong Le[1,2], Thi Thanh Do[1,2], Thi Van Phan[1,2], Luu Hong Dang Nguyen[1,2], Duy Sinh Nguyen[1,2], Van Thinh Cao[9], Thanh Thuy Thi Do[2], Dinh Kiet Truong[2], Hung Sang Tang[1,2], Hoa Giang[1,2], Hoai-Nghia Nguyen[1,2], Minh Duy Phan[1,2,*] , Le Son Tran[1,2,*]

[1]Gene Solutions, Ho Chi Minh City, Vietnam
[2]Medical Genetics Institute, Ho Chi Minh City, Vietnam
[3]University Medical Center, Ho Chi Minh City, Vietnam
[4]MEDIC Medical Center, Ho Chi Minh City, Vietnam
[5]National Cancer Hospital, Hanoi, Vietnam
[6]Hanoi Medical University, Hanoi, Vietnam
[7]Thu Duc City Hospital, Ho Chi Minh City, Vietnam
[8]Department of Biostatistics, Virginia Commonwealth University, School of Medicine, Richmond, VA, USA
[9]Pham Ngoc Thach University of Medicine, Ho Chi Minh City, Vietnam

# Van Thien Chi Nguyen and Trong-Hieu Nguyen contributed equally to this study.

*Correspondence: pmduy@yahoo.com; leson1808@gmail.com

Key words: liquid biopsy, multimodal analysis, methylation, fragment length, cell-free DNA, circulating tumor DNA, multicancer early detection, tissue of origin, machine learning, graph convolutional neutral networks.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

The non-invasive approach for early cancer detection promises a screening assay accessible for everyone. However, the delivery of this promise is limited due mostly to the high sequencing cost associated with available assays. Here, we developed a multimodal assay called SPOT-MAS (Screening for the Presence Of Tumor by Methylation And Size) to simultaneously profile methylomics, fragmentomics, copy number, and end motifs in a single workflow using targeted and shallow genome-wide sequencing of cell-free DNA. We applied SPOT-MAS to 738 nonmetastatic patients with breast, colorectal, gastric, lung and liver cancer, and 1,550 healthy controls. SPOT-MAS detected the five cancer types with a sensitivity of 72.4% and specificity of 97.0%, with AUC of 0.95 (95% CI 0.93-0.96). For tumor-of-origin, a graph convolutional neural network was adopted and could achieve an accuracy of 0.7. In conclusion, our study demonstrates comparable performance to other early cancer detection assays while requiring significantly lower sequencing depth, making it economically feasible for population-wide screening.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

The incidence of cancer-related morbidity and mortality is rapidly increasing globally, and accounted for nearly one fifth of all deaths in 2020 (*1*). High-cost treatment is a significant financial burden for cancer patients, with almost 286 billion dollars in 2021 and an increase of 8.2% to 581 billion dollars in 2030. In Vietnam, GLOBOCAN 2020 reported over 182,500 newly diagnosed cases and 122,690 cancer-related deaths (*1*). Among these, liver (14.5%), lung (14.4%), breast (11.8%), gastric (9.8%), and colorectal cancer (9%) are the five most common types. Up to 80% of cancer patients in Vietnam were diagnosed at stage III or stage IV, resulting in a high rate of 1-year mortality (25%) and a low 5-year survival rate compared to other countries (*2*). Diagnostic delays are associated with a lower chance of survival, greater treatment-associated problems, and higher costs (*3*). Cancer detection at earlier stages can improve the opportunity to control cancer progression, increase the patient survival rate, and lower medical expenses (*4*).

Most current early cancer screening assays have limitations such as invasiveness, low accessibility, and high false positive rates when used sequentially, resulting in overdiagnosis and overtreatment. Multi-cancer early detection (MCED) tests can potentially overcome these challenges by simultaneously detecting multiple cancer types from a single test (*5*). Liquid biopsy, an emerging non-invasive approach for MCED, can capture a wide range of tumor features, including cell free DNA (cfDNA), circulating tumor DNA (ctDNA), exosomes, proteins, mRNA, and metabolites (*6, 7*). Among them, ctDNA has become a promising biomarker for detecting early-stage cancers because it is a carrier of genetic and epigenetic modifications from cancer-derived DNA (*8*). Indeed, ctDNA detection has demonstrated several advantages in non-invasive diagnostic, prognostic, and monitoring of cancer patients during and after treatment (*9, 10*). Furthermore, ctDNA carrying tumor-specific alterations could be used to identify the corresponding unknown primary cancer and tumor localization.

In recent years, there has been considerable interest in exploring the potential of ctDNA alterations for early detection of cancer and localization of the tissue of origin (TOO) (*11, 12*). One such approach is the PanSeer test, which uses 477 differentially methylated regions (DMRs) in ctDNA to detect five different types of cancer up to four years prior to conventional diagnosis (*13*). However, this assay is limited in its ability to determine the TOO, as it only uses methylation regions common to multiple cancers. The DELFI assay employs a genome-wide analysis of ctDNA fragment profiles to increase sensitivity in early detection, but also lacks accuracy in classifying the source of tumor-derived cfDNA (*14*). Recently, the Galleri test has emerged as a multi-cancer detection assay that analyses more than 100,000 methylation regions in the genome to detect over 50 cancer types and localize the tumor site (*15*). This approach requires a large-scale target capture panel at very high-depth sequencing (with a depth coverage of 30X), incurring high sequencing costs and limitting the accessibility of this test to the wider population.

Despite their great potential, there remain several challenges that these assays must solve to deliver accessible and reliable clinical adoption for the large population, including the low fraction of ctDNA in the blood of early stage cancer patients, the heterogeneity of ctDNA signatures from diverse cancer types, subtypes and stages (*16*), and the high sequencing depth required. To address these challenges, recent studies have focused on multi-analyte approach - combining genomic and nongenomic features such as methylomics and fragmentomics to

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

increase the detection of ctDNA and accuracy for TOO identification (*16-18*). Advances in multimodal analysis approaches have led to the development of powerful screening tests that enable high sensitivity and cost-effectiveness. For example, CancerSEEK uses a combined approach of protein biomarkers and genetic alterations to detect and locate the presence of eight types of cancers (*19*). In this assay, cancer-associated serum proteins play a complementary role in tumor localization as cfDNA mutations are not tissue specific. However, detecting both protein and genetic biomarkers are time-consuming and costly. Thus, the development of future MCED tests should endeavor to deliver a screening approach with high sensitivity, specificity, and TOO identification at cost-effective price to provide better clinical outcomes and treatment opportunities for all cancer patients.

In an effort to address the challenges of early cancer detection, we have developed a multimodal approach called SPOT-MAS (Screening for the Presence Of Tumor by DNA Methylation And Size). This assay was previously applied to cohorts of colorectal (*20*) and breast cancer (under review in Frontiers in Oncology) patients and demonstrated ability for early detection of these cancers at high sensitivity across different cancer stages and patient age groups. In this study, we expanded our multimodal approach, SPOT-MAS to comprehensively analyze methylomics, fragmentomics, DNA copy number and end motifs of cfDNA for simultaneously detecting and locating cancer from a single screening test. As proof of concept, we used 2,288 participants, including 738 nonmetastatic patients and 1,550 healthy controls, to train and fully validate this approach on five commonly diagnosed cancers, including breast, gastric, lung, colorectal, and liver cancer. These cancer types accounted for more than 54% of new cancer cases and 57% of cancer death worldwide as well as the most diagnosed cancer types in developing counties (*1, 2*). By using targeted sequencing, shallow whole genome sequencing (with a depth coverage of 0.55X) and innovative machine learning algorithms, we could analyze a large multi-feature datasets of cfDNA for multi-cancer early detection and tumor localization with high sensitivity and cost-effectiveness. Our assay achieved sensitivities of early detection of 72.4% among the five cancer types at specificity of 97.0%, with AUC of 0.95 . Moreover, we could identify the tissue of origin with an accuracy of 0.7 in independent validation cohort using graph convolutional neural network. Thus, SPOT-MAS has the potential to become a universal, simple, and cost-effective approach for early multi-cancer detection in large populations.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results

Clinical characteristics of cancer and healthy participants.

This study recruited 738 patients with five common cancer types, including breast cancer (n=223), CRC (n=159), gastric cancer (n=98), liver cancer (n=122), lung cancer (n=136) and 1,550 healthy participants (Table S1). Cancer patients were diagnosed by either imaging and/or histology analysis, depending on cancer type. All cancer patients were treatment-naïve at the time of blood collection. Healthy participants had no history of cancer at the time of sample collection and remained cancer-free at the 6- and 12-month follow-ups. Cancer patients and healthy participants were randomly assigned to the discovery and validation cohorts (Table 1 and Table S2). The discovery cohort was used to profile multiple cancer- and tissue-specific signatures and to construct machnine learning algorithm while the validation cohort was used solely to external evaluation of the performance of machine learning models.

The discovery cohort comprised of 499 cancer patients (156 breast, 106 CRC, 67 gastric, 77 liver and 93 lung, Table S1) and 1,076 healthy participants. The cancer group had a median age of 58 (range 25 to 97, Table 1) and consisted of 279 females and 220 males. The discovery healthy group consisted of 599 females and 477 males, with a median age of 47 (range 18 to 84, Table 1). In the discovery cohort, gender ratios were similar between cancer and healthy control groups, whereas cancer patients were older than controls ($p< 0.0001$, Mann-Whitney test, Table 1). Of the cancer patients, 10.4% were at stage I, 33.9% were at stage II, and 30.1% were at non-metastatic stage IIIA. Staging information was not available for 25.7% of cancer patients, who were confirmed by specialized clinicians to have non-metastatic tumors (Table 1).

The validation cohort consisted of 239 cancer patients (67 breast, 53 CRC, 31 gastric, 45 liver and 43 lung, Table S1) and 474 healthy participants (Table 1). Consistent with the discovery cohort, the gender distribution was comparable between the cancer and healthy control groups, and the cancer group was older than the control group, with a median age of 59 and 48 years old, respectively ($p<0.0001$, Mann-Whitney test,Table 1). The percentage of cancer patients with each stage was similar to that of the discovery cohort, with 9.6% at stage I, 28.9% at stage II and 32.2% at stage IIIA. Staging information was unavailable for 29.3% of non-metastatic cancer patients (Table 1).

The multimodal SPOT-MAS assay for multi-cancer and tissue of origin detection

In our recent study of SPOT-MAS, we have demonstrated that the integration of ctDNA methylation and fragmentomic features can significantly improve the early detection of colorectal cancer (*20*). Here, we expanded the breadth of ctDNA analyses by adding two sets of features including DNA copy number and end motif into SPOT-MAS to maximize cancer detection rate and identify TOO. Briefly, a novel and cost-effective workflow of SPOT-MAS was developed involving three main steps (Figure 1). In step 1, cfDNA was isolated from peripheral blood and subjected to bisulfite conversion and adapter ligation to create a single whole-genome bisulfite library of cfDNA. From this library, in step 2, a hybridization reaction was performed to collect the target capture fraction (450 cancer specific regions), then the whole-genome fraction was retrieved by collecting the 'flow-through' and hybridizing with probes specific for adapter sequences of DNA library. Both the target capture fraction and whole-genome fraction were sequenced to the depth of ~52X and 0.55X, respectively (Table

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

S3). Data pre-processing was performed to generate five different sets of cfDNA features, including methylation changes at target regions (TM), genome-wide methylation (GWM), fragment length patterns (Flen), copy number aberrations (CNA) and end motif (EM). In step 3, these features were used as inputs for a two-stage model to obtain prediction outcomes. Stage 1 of our model comprised of a stacked ensemble machine learning model for binary classification of cancer versus healthy. Then the samples predicted as cancer were passed to stage 2 where graph convolution neural network (GCNN) was adopted to predict TOO (Figure 1).

Identification of differentially methylated regions (DMRs) in cancer patients from target capture fraction

DNA methylation is an important epigenetic signature responsible for major changes in regulating expression of cancer associated genes by impacting the binding of transcription factors to regulatory sites and the structure of chromatin (*21, 22*). Of the 450 target regions associated with cancer that were selected from public data (*13, 23*), 402 regions were identified as differentially methylated regions (DMRs) in cancer patients when compared to healthy participants from the discovery cohort (Wilcoxon rank-sum test, p-values < 0.05, Figure 2A and Table S4). Of those, 339 (84.3%) regions were identified as hypermethylated ($logFC > 0$), and 63 (15.7%) regions as hypomethylated in cancer samples ($logFC < 0$, Figure 2A). We next examined the genomic location of the 402 DMRs and found 100, 108, 107 and 87 DMRs that were mapped to promoter, exon, intron and intergenic regions, respectively (Figure 2B). To understand the relationship between the differences in methylation regions and biological pathways, we performed pathway enrichment analysis using g:Profiler on hypermethylated DMRs. We detected 36 enriched pathways, including 14 from Kyoto Encyclopedia of Genes and Genomes (KEGG) and 22 from WikiPathway (WP) (Figure 2C and Table S5). These significant pathways were known to regulate tumorigenesis of breast, gastric, hepatocellular, and colorectal cancer. Therefore, the methylation changes in the targeted regions, particularly the hypermethylated DMRs, mostly occur early in tumorigenesis and are crucial for distinguishing early-stage cancer patients from healthy individuals.

Genome-wide methylation changes in cfDNA of cancer patients

In addition to site-specific hypermethylation, hypomethylation is a significant genome-wide change that has been identified in many types of cancers (*24-26*). To investigate the methylation changes at genome-wide level, bisulfite sequencing reads from the whole-genome fraction were mapped to the human genome, split into bins of 1Mb (2,734 bins across the genome), and the reads from each bin were used to calculate methylation ratio. As expected, we observed a left-ward shift in the distribution of methylation ratio in cancer samples compared to healthy controls, indicating global hypomethylation in the cancer genome ($p<0.0001$, two-sample Kolmogorov-Smirnov test, Figure 3A). Of these bins, we identified 1,715 (62.7%) bins as significantly hypomethylated in cancer, located across 22 autosomes of the genome (Figure 3B, Wilcoxon rank sum test with Benjamini-Hochberg adjusting p-value <0.05). In contrast, there were only 10 bins identified as hypermethylated and mapped to chromosome 1, 2, 3, 5, 6, 7 and 12 in the cancer genome (Figure 3B). Therefore, our data confirmed the widespread hypomethylation across the genome and this would potentially serve to distinguish cancer patients from healthy controls.

Increase DNA copy number aberrations (CNAs) in cfDNA of cancer patients

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Somatic copy number aberrations (CNAs) in the cancer genome are associated with the initiation and progression of numerous cancers by altering transcriptional levels of both oncogenes and tumor suppressor genes (*27*). Recent studies have shown that CNAs detection could identify and quantify the fraction of ctDNA in plasma cfDNA (*28-30*). To examine CNAs at genome-wide scale, we used 1Mb bin to determine the percentage of bins that showed significant copy number gains or losses between cancer and control group. We identified 729 bins (27.1%) with a significant gain and 976 bins (36.3%) with a significant loss in copy number across 22 chromosomes of the cancer genome (Benjamini-Hochberg adjusting p-value <0.05, Wilcoxon rank sum test, Figure 4A). We noted that chromosome 8 had the highest proportion of bins with CNA gains, while chromosome 22 showed the highest proportion of bins with CNA losses (Figure 4B).

It is thought that the abnormal hypomethylation at genome-wide level is linked with somatic copy number aberration (CNA), resulting in genome instability, which is an important tumorigenic event (*31-33*). Indeed, our data showed a significant increase in levels of CNA in hypomethylated bins compared to bins with unchanged methylation (p=0.024, Figure S1A). Consistently, bins with CNA gains showed significant decreases in methylation as compared to those with CNA losses or unchanged CNA (p<0.01, Figure S1B). In summary, SPOT-MAS enables comprehensive profiling of both global differences in methylation and somatic CNA as individual feature types, as well as exploring their functional links during cancer initiation and development, rendering them ideal biomarkers for cancer detection.

Fragment length analysis captured patterns of ctDNA in plasma

Several studies have shown that the fragmentation pattern of cfDNA is a non-random event mediated by apoptotic-dependent caspases and ctDNA fragments tend to be shorter than non-cancer cfDNA (*14, 34-37*). One novel technical aspect of SPOT-MAS is the use of bisulfite sequencing data not only for methylation but also for fragment length analysis. Certain studies showed evidence of DNA degradation followed bisulfite treatment, possibly due to high temperature and low pH conditions of the bisulfite conversion procedure, while other showed that bisulfite sequencing affects large genomic DNA but not small size cfDNA (*38-41*). Therefore, to demonstrate the use of bisulfite treated cfDNA for fragment length analysis, we randomly selected 3 healthy controls and 9 cancer samples to perform pair-wise comparison between bisulfite and non-bisulfite sequencing results. We observed a strong correlation between fragment length profile of non-bisulfite and bisulfite sequencing (Pearson correlation, $R^2 > 0.9$, p<0.0001, Figure S2A) for all 12 tested samples, indicating the feasibility of using bisulfite sequencing data for cfDNA fragment length analysis. Indeed, the fragment size distributions of bisulfite-treated cfDNA in both cancer patients and control subjects showed a peak at 167 bp (Figure 5A), corresponding to the length of DNA wrapped around histone (~ 147 bp) plus linker regions (~ 2x10 bp), which was in good agreement with previous studies using non-bisulfite cfDNA (*14, 42*). Importantly, our results showed that cfDNA of cancer patients was more fragmented than that of healthy participants, with a higher frequency of fragments ≤ 150 bp and a lower frequency of fragment > 150 bp (Figure 5A).

To examine whether the fragment length variation in cancer-derived cfDNA and non-cancer cfDNA could be position-dependent (*14*), we calculated the ratios of short (≤ 150bp) to long fragments (> 150 bp) across the genome in cancer patients and healthy controls. The mean ratio of short to long fragments in cancer patients was 0.29 (range 0.28 to 0.42), which was higher

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

than the mean ratio of 0.27 (range 0.26 to 0.39) for healthy controls (Figure 5B). The changes of mean ratio were across 22 autosomes of the genome. Our results indicate that the SPOT-MAS technology can effectively capture differences in fragmentation patterns between cancer and healthy participants across the entire genome, making them potential biomarkers for the detection of circulating tumor DNA in plasma.

Profile of 4-mer end motifs reflecting differences between cancer and healthy cfDNA

Associated with differences in fragment length is the differences in the DNA motifs at the end of each fragment as the consequences of differential cleavage between DNA in cancer cells and normal cells during apoptosis (*42, 43*). Here, we calculated the frequencies of 256 4-mer end motifs (EMs) of cfDNA fragments and compared them between cancer patients and healthy participants. Consistent with the fragment length features, we also confirmed the correlation of EM frequency between bisulfite and non-bisulfite sequencing results of 12 randomly selected samples, suggesting that EM profiles were reserved in bisulfite treated cfDNA (Figure S2B). Of the 256 4-mer EMs, we detected 78 motifs with increased frequencies and 106 motifs with decreased frequencies between cancer and healthy controls (Figure 6A and Table S6).

Interestingly, EMs beginning with cytosine (C) exhibited the highest number of EMs with significant changes of frequency in cancer samples (Figure 6A). Figure 6B shows the top ten EMs exhibiting significant differences. Specifically, the frequencies of five motifs (CAAA, TAGA, CAGA, CAAG, and CAAT) were found to be significantly increased, while the frequencies of another five motifs (CGCT, CGCC, CGCA, GCCT, and CGTT) were significantly decreased in cancer patients (Figure 6B). Therefore, the differences in end motif frequency identified by SPOT-MAS between cancer patients and healthy participants may serve as a promising target for the identification of ctDNA.

SPOT-MAS assay combining different features of cfDNA to enhance the accuracy of cancer detection

In order to increase the sensitivity of early cancer detection while avoiding the high cost of deep sequencing, a screening test should survey a wide range of ctDNA signatures (*16*). Therefore, we utilized multiple ctDNA signatures to construct classification models for distinguishing cancer patients from healthy individuals. To expand the feature space, we generated four additional features based on fragment length, including short, long, total fragment count, and short-to-long ratio, resulting in nine input feature groups (Figure 7A). For each feature group, we tested three different algorithms, including random forest (RF), logistic regression (LR), or extreme gradient boosting (XGB), to tune hyperparameters and select the optimal algorithms (Figure 7A). To evaluate the performance of these single-feature models, we performed 20-fold cross-validation on the discovery dataset and calculated "Area Under the Curve" (AUC) of the "Receiver Operating Characteristic" (ROC) curve. Among the nine features, EM-based model showed the highest AUC of 0.90 (95% CI: 0.89-0.92, Figure 7B) while the SHORT-based model had the lowest AUC of 0.71 (95% CI: 0.69-0.74, Figure 7B).

To assess whether combining features could improve classification, we used two strategies to construct multi-feature models. In the first strategy, all nine feature groups were concatenated into a single data frame before being fed into the RF, LR, or XGB algorithms. Of the three algorithms, the XGB model exhibited the best performance with an AUC of 0.88 (95% CI: 0.87-0.90, Figure 7B). However, this AUC is still lower than that of the EM-based model (0.88

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

versus 0.90, Figure 7B). In the second strategy, we constructed an ensemble stacking model using logistic regression to combine the prediction results of the single-feature models. We conducted an exhaustive search approach to evaluate the performance of 511 possible combinations. The stacking ensemble model based on combining eight features, including TM, GW, CNA, FLEN, LONG, TOTAL, RATIO and EM, exhibited the best performance and outperformed the single-feature models (Table S7), with an AUC of 0.93 (95% CI: 0.92-0.95, Figure 7B and Figure S3). In the independent validation cohort, we obtained similar results, where the ensemble model also outperformed single-feature models, with an AUC of 0.95 (95% CI: 0.93-0.96, Figure 7C).

In order to ensure cost-effectiveness and minimize psychological impact of cancer screening tests in a large population, high specificity is a crucial requirement. Accordingly, we established the cutoff value for each constructed model based on a minimum specificity threshold of 95%. Of the nine single-feature models, EM and GWM models exhibited the highest sensitivities, at 59.5% and 60.9%, respectively. The stacking ensemble model achieved a sensitivity of 73.8% and a specificity of 95.1% with a cutoff value of 0.546 in the discovery cohort (Figure 7D), and a mean sensitivity of 72.4% and a specificity of 97.0% in the validation cohort (Figure 7E). Stratification of samples by cancer types revealed that the ensemble model performed most accurately in predicting liver cancer (89.6% sensitivity), followed by CRC (82.1%sensitivity m), lung cancer (78.5% sensitivity) and gastric cancer (71.6% sensitivity) (Figure 7F, Table S8). Breast cancer had the lowest detection rate of 58.3% (91/156 patients). Importantly, the performance of our ensemble model remained consistent in the validation cohort, with liver cancer again showing the highest sensitivity (91.1%), followed by lung cancer (83.7%), CRC (83.0%), gastric cancer (61.3%), and breast cancer (49.3%) (Figure 7G, Table S8).

Influence of clinical features on model prediction

Upon stratifying our dataset by gender, we found that there was no significant difference in the prediction of healthy status between males and females (Figure S4A and S4C). However, in the case of cancer prediction, our model demonstrated higher accuracy in males than females in both the discovery and validation cohorts (Figure S4A and S4C). Notably, when breast cancer samples were removed from our analysis, there was no difference in the detection rates between male and female patients (Figure S4B and S4D), suggesting that the observed gender bias may be attributed to the high proportion of breast cancer patients (all females) in our cohort, who exhibited the lowest detection rate among the five cancer groups.

We next evaluated the potential confounding effect of age on our prediction model by examining the correlation between the model prediction scores and the participants' ages. The results revealed no significant correlation, suggesting that age differences are unlikely to affect the accuracy of our model (Figure S4E and S4F). With regards to cancer burden (ie. tumor size), our model performed better for cancers with higher burden, as reflected by the higher cancer scores assigned to these cases (Figure S4G and S4H). Specifically, patients with tumor diameter $\geq$ 3.5 cm were more likely to be detected than those with a diameter < 3.5 cm (Figure S4G and S4H). Similarly, cancer stages also influence the performance of our stacking ensemble model, showing increasing detection accuracy as the stages get more advanced. In the discovery cohort, the model's accuracy was highest for stage IIIA cancers, with an AUC of 0.95 (95% CI 0.93-0.97), and lowest for stage I cancer, with an AUC of 0.90 (95% CI 0.86-

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

0.95) (Figure S4I and S4J). Consistently, our model performance was lower with an AUC of 0.94 (95% CI 0.89-0.98) and 0.93 (95% CI 0.90-0.96) for stage I and II cancer, respectively, increasing to 0.98 (95% CI 0.97-0.99) for stage IIIA in the validation cohort (Figure S4K and S4L). These results demonstrated that our ensemble model can detect cancers at all stages found in our cohorts, despite a slightly lower performance in early stages (stage I and II) compared to non-mestatatic stage (IIIA).

SPOT-MAS enables prediction of cancer types

The ability to predict the tissue origin of ctDNA is critical for early cancer detection as this can guide subsequent diagnostic tests and treatment. Previous studies have attempted to use either fragment length or methylation landscapes to achieve this goal (*5, 14, 44*). In this study, we demonstrated the ability of SPOT-MAS to identify the TOO using low-depth bisulfite sequencing to generate multiple sets of cfDNA features. We first concatenated the nine sets of cfDNA features into a single data frame and focused our analysis on 499 cancer patients with five cancer types in the discovery cohort. We then constructed a Random Forest (RF) and two neural network models (convolutional neural network and graph convolutional neural network) to predict the TOO and used 10-fold cross-validation to estimate and compare the performance of these models (Figure 8A and Figure S5A). The Graph Convolutional Neural Network (GCNN) was chosen due to its superior performance and stability (Figure S5B and S5C and Table S9).

We then used the GNNExplainer tool to measure the importance of different cfDNA features. Our results showed that breast cancer had the highest number of features with an important score >0.9 (497 features), while lung cancer had the lowest number of important features (126 features) (Figure 8B). Colorectal, gastric, and liver cancers had 363, 309, and 204 important features, respectively (Figure 8B and Table S10). Genome-wide methylation and copy number aberration were the most important features for differentiating breast, colorectal, CRC, gastric and liver cancer from other cancer types, while the end motif had the lowest contribution to distinguish cancer types (Figure 8C). Visualization of the 3D GCNN showed that this set of discriminative features could segregate the five different cancer types (Figure 8D), highlighting the benefits of a multimodal approach for predicting TOO.

The median accuracy for TOO identification among the five cancer types by the GCNN-based multi-feature model was 0.73 (range 0.54 to 0.87) in the discovery cohort (Figure 8E). The accuracy in the discovery cohort was highest for breast (0.87) and liver cancer (0.82) and lowest for gastric cancer (0.54). In the validation cohort, we obtained a slightly lower accuracy with a median of 0.70 (range 0.55 to 0.78). The accuracies for individual cancer types were 0.78 for breast, 0.76 for liver, 0.66 for colorectal, 0.63 for lung and 0.55 for gastric cancer (Figure 8F). Among the 5 cancer types, breast cancer showed the highest TOO accuracy, possibly due to the highest number of important features detected by the model. In contrast, CRC and gastric cancer exhibited the lowest TOO accuracy with high misprediction rates between these two cancer types (0.11 and 0.19 for CRC versus gastric and gastric versus CRC, respectively). Together, our study highlights the benefits of integrating multimodal analysis with the GCNN model to capture the broad landscape of tissue-specific markers in different cancer types.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Discussion

In an era marked by a global rise in cancer-related morbidity and mortality, the development of liquid biopsy screening tests that can detect and localize cancer at an early stage holds tremendous potential to revolutionize cancer diagnosis and therapy. Despite this, challenges in test performance and cost must still be overcome, due mostly to the limited abundance of ctDNA and its inherent variability. To address these, published liquid biopsy assays to date demanded a very high-depth sequencing (*15*), or a combination of protein and genetic biomarkers (*19*), resulting in an elevated price of test. In the current study, we present the SPOT-MAS assay as a single workflow with comparable performance to current tests while requiring a much lower sequencing depth (Table S11). SPOT-MAS achieved a sensitivity of 72.4 % at a specificity of 97.0 % for detecting five common cancer types using shallow depth sequencing. Furthermore, it can predict the tissue of origin with an accuracy of 70%.

The SPOT-MAS assay allows comprehensive investigation of multiple biomarkers in cfDNA, including targeted methylation (TM), genome-wide methylation (GWM), copy number aberration (CNA), end motif (EM) and fragment length profiles (Flen). In TM analysis, out of 450 TM regions chosen from previous publications (*13, 23*), we identified 402 regions as significant differentially methylated regions (DMRs) in cancer patients (Figure 2A). These DMRs were enriched for regulatory regions of well-known cancer-related gene families such as PAX family genes, TBX family genes, FOX family genes and HOX family genes, and some have previously been reported as biomarkers for noninvasive cancer diagnosis, such as *SEPT9* and *SHOX2 (45, 46).* In addition to the targeted hypermethylation regions, our study also showed widespread hypomethylation patterns across 22 autosomes of cancer patients (Figure 3), a hallmark of cancer (*47*). Importantly, we demonstrated that the same bisulfite sequencing data could be used to identify somatic CNA (Figure 4), cancer-associated fragment length (Figure 5) and end motifs (Figure 6), highlighting the advantage of SPOT-MAS in capturing the broad landscape of ctDNA signatures without high cost deep sequencing. For cancer-associated fragment length, we pre-processed this data into five different feature tables to better reflect the information embedded within the data. Overall, nine feature tables are available for model construction.

The involvement and orthogonal links of the above features in the transcriptional regulation of cancer-associated genes during carcinogenesis prompted us to examine whether the combination of multiple cancer-specific signatures in cfDNA could improve the efficiency of cancer detection (*48, 49*). We first determined the performance of models constructed using individual type of cfDNA features. Next, by performing exhaustive searches for all possible combinations of single-feature models, we identified that the stacking ensemble of seven features could achieve the AUC of 0.95 (95% CI: 0.93-0.96, Figure 8C and Figure S3), which is superior to all single-feature models. Among the five cancer types, breast cancer showed the lowest detection rate of 58.3% and 49.3% in the discovery and validation cohort, respectively. Variations in detection rates among different cancer types have been previously reported (*5, 19, 44*). Consistently, it has been reported that the detection of breast cancer, particularly in early stages, is challenging due to the low levels of ctDNA shedding and heterogeneity of molecular subtypes of breast tumors (*5*). In contrast, we obtained the highest detection rate for liver cancer patients with the sensitivity of 89.6% and 91.1% in the discovery and validation cohort, respectively. Our finding is in good agreement with the literature showing that liver tumors shed high amounts of ctDNA (*50*). This result demonstrated the advantage of a

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

multimodal approach to enhance ctDNA detection in plasma. We also conducted a survey of liquid biopsy assays to put our SPOT-MAS into the context of current state-of-the art in the field. Table S11 showed that SPOT-MAS is using the lowest sequencing depth approach (with a depth coverage of ~0.55X) and making up for this by integrating the greatest number of cfDNA features to achieve comparable performance to other assays.

For TOO identification, our results showed that the graph convolutional neural network (GCNN) performed the best among the models tested (Figure S5 and Figure 8). GCNN has the ability to explore the similarity and mutual representation among samples, therefore achieving great success in multi-class classification tasks (*51, 52*). Unlike the reference-based deconvolution approaches (*53, 54*), our GCNN approach is independent of a reference methylation atlas, which was developed from tissue or cell type specific methylation markers and thus may introduce bias due to discordance between the methylomes of tissue gDNA and plasma cfDNA (*16, 55*). Although the methylation changes were reported as most predictive for TOO in previous studies (*53, 54*), our results showed the contribution of each of the 9 features for TOO identification (Figure 8C). In addition to GWM, fragment ratio (RATIO) and CNA are the major contributors to the discrimination of different tissue types. This finding provided additional evidence that the multimodal approach capturing the breadth of tissue-specific signatures could improve the accuracy of TOO identification (*5*). Our GCNN model achieved an accuracy of 0.70 for TOO prediction in validation cohort. This was comparable to the performance of CancerLocator, which was based a probabilistic distribution model of tissue specific methylation markers (*56*). Recently, Liu et al. (*5*) developed a methylation atlas based method, which achieved a higher accuracy of 93% for locating 50 types of cancer. However, this approach is based on deep genome-wide sequencing with high depth coverage of 30X (Table S11), thus might not be a cost effective approach for cancer screening in large populations, especially in low-income countries.

There are several limitations in our study. First, despite using a large dataset of 738 cancer samples, there was an unequal distribution of samples among cancer types, with breast cancer accounting for 30.2% (223/738, Table S2) of the total samples and gastric cancer having a much smaller representation (13.3%, Table S2). As a result, our models may have been influenced by this imbalance, potentially introducing bias in the training and evaluation process. Therefore, future studies should consider incorporating more samples to better estimate the overall performance of the SPOT-MAS test. Second, tumor staging information was not available for 26.8% of cancer patients (198/738) in our study. This is due to the patients' decision to select different hospitals for diagnostics and treatment, leading to missing histopathological information at the hospitals where they were originally recruited. However, all cancer patients recruited in this study were confirmed to have non-metastatic tumors. Third, the cancer patients in both the discovery and validation cohort were older than the healthy participants. Age differences could be a confounding variable of methylation and could affect the model performance (*57, 58*). However, we observed no significant association between the participants' age and model prediction scores (Figure S4). Fourth, the ability of SPOT-MAS to differentiate cancer patients from those with benign lesions has not been examined in this study. Fifth, this study only focused on the top 5 common cancer types, thus the current version of SPOT-MAS might misidentify cancer patients of other types, resulting in lower sensitivity to real world application. Lastly, this was a retrospective cohort study and may be biased by the nature of this study design. In an interim 6-month report of a prospective study named K-

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

DETEK, we were encouraged by the preliminary data demonstrated the ability of SPOT-MAS to detect cancer patients who exhibited no symptoms at the time of testing (*59*). Despite these promising results, the performance of SPOT-MAS as an early cancer screening test remains to be fully validated in a large, multi-center prospective study with 1 to 2 years of follow up.

In conclusion, we have developed the SPOT-MAS assay to comprehensively profile methylomic, fragmentomic, copy number aberrations, and motif end signatures of plasma cfDNA. Our large-scale case-control study demonstrated that SPOT-MAS, with its unique combination of multimodal analysis of cfDNA signatures and innovative machine-learning algorithms, can successfully detect and localize multiple types of cancer at a low-cost sequencing. These findings provided important supporting evidence for the incorporation of SPOT-MAS into clinical settings as a complementary cancer screening method for at-risk populations.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Materials and Methods

Patient enrollment

This study recruited 738 cancer patients (223 breast cancer, 159 CRC, 122 liver cancer, 136 lung cancer, 98 gastric cancer) and 1550 healthy subjects. All cancer patients were confirmed to have one of the five cancers analyzed in this study. Cancer stages were determined following guidelines from the American Joint Committee on Cancer and the International Union for Cancer Control (*60*). Individuals were considered healthy if they had no history of cancer at the time of enrollment and follow-up interviews were conducted by specialized physicians to confirm noncancer status at 6 and 12 months after enrollment. Study subjects were recruited from the University Medical Center, Thu Duc City Hospital, University of Medicine and Pharmacy, Medic Medical Center and Medical Genetics Institute in Ho Chi Minh city, Vietnam, National Cancer Hospital and Hanoi Medical University in Hanoi from May 2019 to December 2022.

Written informed consent was obtained from each participant in accordance with the Declaration of Helsinki. This study was approved by the Ethics Committee of the Medic Medical Center, University of Medicine and Pharmacy and Medical Genetics Institute, Ho Chi Minh city, Vietnam. All cancer patients were treatment-naïve at the time of blood sample collection.

Isolation of cfDNA

10 mL of blood was collected from each participant in a Cell-Free DNA BCT tube (Streck, USA). Plasma was collected from blood samples after centrifugation with two rounds (2,000 × g for 10 min and then 16,000 × g for 10 min). The plasma fraction was aliquoted for long-term storage at -80 ºC. Cell free DNA (cfDNA) was extracted from 1 mL plasma aliquots using the MagMAX Cell-Free DNA Isolation kit (ThermoFisher, USA), according to the manufacturer's instructions. Extracted cfDNA was quantified by the QuantiFluor dsDNA system (Promega, USA).

Bisulfite conversion and library preparation

According to the manufacturer's instructions, bisulfite conversion and cfDNA purification were prepared by EZ DNA Methylation-Gold Kit (Zymo research, D5006, USA). DNA library was prepared from bisulfite-converted DNA samples using xGen™ Methyl-Seq DNA Library Prep Kit (Integrated DNA Technologies, 10009824, USA) with Adaptase™ technology, according to the manufacture's instructions. The QuantiFlour dsDNA system (Promega, USA) was used to analyse the concentration of DNA.

Target region capture, whole genome hybridization & sequencing

DNA from library products were pooled equally, hybridized and captured using The XGen hybridization and wash kit (Integrated DNA Technologies, 1072281, USA), together with our customized panel of xGen Lockdown Probes including 450 regions across 18,000 CpG sites (Integrated DNA Technologies, USA). The construction of panel was built as previously described (*13, 20, 23*). After hybridization, the flow-through product was concentrated using SpeedVac (N-Biotek, NB-503CIR, Korea) at 65◦C. The samples were then added with the

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

hybridization master mixture (hybridization buffer, hybridization enhancer and H_2O) and denatured. Biotinylated P5 and P7 probes (P5-biotin: /5Biosg/AATGATACGGCGACCACCGA, P7-biotin: /5Biosg/CAAGCAGA AGACGGCATACGAGAT) on streptavidin magnetic beads (Invitrogen, CA, USA) were hybridized with the single-stranded DNA. The captured DNA products were amplified by a PCR reaction with free P5 and P7 primers (P5 primer: AATGATACGGCGACCACCGA, P7 primer: CAAGCAGAAGACGGCATACGA). The concentrations of DNA libraries were determined using the QuantiFluor dsDNA system (Promega, USA). Both target and flow-through fraction were sequenced on the DNBSEQ-G400 DNA system (MGI Tech, Shenzhen, China) with 100-bp paired-end reads at a sequencing depth of 20 million reads per fraction. Data was demultiplexed by bcl2fastq (Illumina, CA, USA). FASTQ files were then examined using FastQC v. 0.11.9 and MultiQC v. 1.12.

Targeted methylation analysis (TM)

All paired-end reads were processed by Trimmomatic v 0.32 with the option HEADCROP. The trimmed reads were then aligned by Bismark v. 0.22.3. Deduplication and sorting of BAM files were conducted using Samtools v. 1.15. Reads falling into our 450 target regions were filtered using Bedtools v. 2.28. Methylation calling was performed using Bismark methylation extractor (*20*). Briefly, methylation ratio was measured for each target region:

$$\text{Methylation ratio} = \frac{methylated\ cyto\sin e\ (C)}{methylated\ C + unmethylated\ C}$$

Methylation fold change from cancer to control was calculated for each target region. For analyzing differential methylated regions, significance level was set at $p \leq 0.05$, corresponding to a -log10 adjusted p-value ≥1.301 (Benjamini-Hochberg correction).

Genome-wide methylation analysis (GWM)

The integrated bioinformatics pipeline Methy-pipen was used to analyse GWM. We carried out the trimming step using Trimmomatic, removing adapter sequence and low-quality bases at fragment ends (*20*). The methylation ratio for each bin was calculated as following equation.

$$\text{Methylation ratio} = \frac{methylated\ cyto\sin e\ (C)}{methylated\ C + unmethylated\ C}$$

Mean methylation ratio was calculated for each bin and subsequently used to plot GWM density curves. To identify bins with significant methylation changes between cancer and control group, methylation ratio in each bin of cancer samples were compared with corresponding values in control samples using Wilcoxon rank sum test. Bins with adjusted p-value (Benjamini-Hochberg correction) ≤ 0.05 were considered significant. Those with $\log_2$ fold change (cancer vs control) > 0 were categorized as hypermethylated bins. Those with $\log_2$ fold change (cancer vs control) < 0 were categorized as hypomethylated bins.

Copy number aberration analysis (CNA)

CNA analysis was performed using the R-package QDNAseq (*35*). We also used 1-Mb segmetation strategy to analyse CNA. We excluded bins that felt into the low mappability and Duke blacklist regions(*25*). The number of reads mapped to each bin was measured by the function “binReadCounts”, and GC-content correction was conducted by the functions

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

"estimateCorrection" and "correctBins". The final CNA feature was derived by bin-wise normalizing and outlier smoothing with the functions "normalizeBins" and "smoothOutlierBins". This process resulted in a feature vector of a length of 2691 bins.

To identify significant DNA gain or loss between cancer and control group, CNA values in each bin of cancer samples were compared with corresponding values in control samples using Wilcoxon rank sum test. Bins with adjusted p-value (Benjamini-Hochberg correction) ≤ 0.05 were considered significant. Those with $\log_2$ fold change (cancer vs control) > 0 were categorized as significant increase. Those with $\log_2$ fold change (cancer vs control) < 0 were categorized as significant decrease.

Fragment length analysis (FLEN, SHORT, LONG, TOTAL, RATIO)

We used an in-house python script to convert the.bsalign files into BAM files and collected the fragment length from 100 to 250 bp, resulting in 151 possible fragment lengths for further analysis. The fragment frequency in each length (%) was measured by getting the proportion of reads with that length to the total read count in the range of 100 to 250 bp. Fragment length (bp) against fragment frequency (%) was plotted to obtain a FLEN distribution curve.

We divided the whole genome into 588 non-overlapping bins of 5Mb (5 million bases) long and then extracted the read counts regarding these bins. Short fragments have lengths from 100 to 150 bp and long fragments have lengths from 151 to 250 bp. The ratio of short and long fragments was calculated by dividing the number of each fragment. All the short, long and total read counts for each sample in 588 bins were normalized using z-score normalization. The short, long and total normalized read counts and short/long ratios were chosen as features analyzed (SHORT, LONG, TOTAL, RATIO).

End motif analysis (EM)

Adaptase^{TM} technology (Integrated DNA Technologies, USA) was used during library preparation to ligate adapters to ssDNA fragments in a template-independent reaction (*42*). This step involved adding a random tail to the 5' end of reverse reads. Although median length of the tail was 8 bp and thus allowed trimming to obtain information for other analysis, the random-length tails did not allow exact determination of the 5' end of the reverse reads. Therefore, EM features were determined based on the genomic coordinate of the 5' end of the forward reads. We determined the first 4-mer sequence based on the human reference genome hg19. In 256 possible 4-mer motifs, the frequency of each motif was calculated by dividing the number of reads carrying that motif by the total number of reads, generating an EM feature vector of a length of 256 for each sample.

Construction of machine learning models

All samples in the discovery cohort were used for model training to classify if a sample is cancerous or not. For every feature type (TM, GWM, CNA, FLEN, SHORT, LONG, TOTAL, RATIO and EM), three machine learning algorithms, including Logistic regression (LR), Random Forest (RF) and Extreme Gradient Boosting (XGB), were applied. By using the "GridSearchCV" function in the scikit-learn (v.1.0.2), model hyperparameters with the best performance were chosen with 'CV' parameter (cross-validation) set to 5. The best hyperparameters for each algorithm were found using function 'best_params_' implemented in GridSearchCV. Subsequently, feature selection was performed for each algorithm as

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

follows: (1) for LR, the "penalty" parameter with 'l1' (LASSO regression), 'l2' (Ridge regression) and 'none' (no penalty) were examined to select the setting with the best performance; (2) for RF and XGB, a "SelectFromModel" function with the 'threshold' was set at 0.0001 to get all features. Then, the three algorithms (LR, RF, XGB) trained with the best hyperparameters and selected features were validated using k-fold cross validation approach on the dataset of training cohort with k-fold set to 20-fold, and 'scoring' parameter set to 'roc_auc'. This split the data into 20 groups, in which 19 groups were model-fitted and the remaining group was tested, which resulted in 20 'roc_auc' scores. The average of these scores was used to obtain the prediction performance of each model. The model with the highest 'roc_auc' average score was chosen (either LR, RF or XGB). Ensembled models were constructed by combining probability scores of nine single-feature base models (TM, GWM, CNA, FLEN, SHORT, LONG, TOTAL, RATIO, EM) with different combination using LR, resulting in one probability score for every sample. An extensive search was performed to evaluate the performance of all possible combinations (n= 511) and the combination with highest AUC was selected as the final model. The model cut-off was set at the threshold specificity of >95%. This combination model performance was evaluated on an independent validation dataset to examine the model classification power.

In addition the stacking ensemble, another combinatory strategy was examined. Instead of combining nine base models , we generated a single dataframe consisting of raw data of all nine features. The model hyperparameters tuning and features selecting were followed the same strategy as described above. After choosing the best algorithm, the model performance was also evaluated using the same external validation dataset.

Construction of models for TOO

Strategy 1: Random Forest (RF) model

A single data frame of nine features in discovery cohort was used to train the Random Forest (RF) to classify 5 cancer types. By using the "GridSearchCV" function in the scikit-learn (v.1.0.2), model hyperparameters with the best performance were chosen with 'CV' parameter (cross-validation) set to 3 and "class_weight" parameter set to "balanced". The best hyperparameters were found by function 'best_params_'. Then, the model was validated using k-fold cross validation approach on the training cohort with k-fold set to 10-fold and its performance was evaluated on the validation cohort.

Strategy 2: Deep neural network (DNN) model

Backpropagation trained the H_2O deep neural network (DNN) (multi-layer feedforward artificial neural network) (H_2O package, version 3.36.1.2) with stochastic gradient descent. The random grid search was selected as previously described (*20*).

Strategy 3: Graph Convolutional Neural Network (GCNN) model

The model training utilized an input graph formed from a discovery dataset and a validation dataset as transudative setting (*13*) comprising patients diagnosed with five types of cancer: breast, colorectal (CRC), gastric, liver, and lung. The discovery dataset contains a set of sample-label pairs $\mathcal{J} = \{(X_i, Y_i) | i = 1, \dots, N\}$ where X_i represents the ith sample and Y_i represents ith label, and N is the number of sample-label pairs. For each X_i in the discovery dataset, a node's feature vector $f = \{F_0, \dots, F_d\} \in \mathbb{R}^d$ is constructed by combining groups of

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

features, where F_i is the ith feature, d is the number of features. The same procedure was applied for the independent validation dataset. To construct an interaction graph between cancer nodes, we employed the k-nearest neighbors' algorithm. An interaction graph defined as $G = (V, E)$ where $V = \{X_i | i = 1, \dots, N\}$ is a node set formed by the discovery samples, and $E = \{e_{ij}\}$ is an edge set, where e_{ij} denotes an edge. Given N nodes in the node set, i.e. $|V| = N$, a graph topology $A \in \mathbb{R}^{N \times N}$ is defined by:

$$A_{ij} = \begin{cases} 1, e_{ij} \in E \text{ and } d_{ij} < \delta \\ 0, otherwise \end{cases}$$

where d_{ij} is the Euclidean distance of node i and j, and δ is set to 0.8.

In accordance with (*23*), a Graph Convolutional Neural Network (GCN) was constructed for the purpose of tissue of origin classification. The network comprised three message-passing layers, each with a hidden size of 44 and a head number of 4. Tissue of origin classification was approached as a node classification problem, wherein the model assigned each node to one of five cancer types: breast, colorectal, gastric, liver or lung cancer. Focal loss was employed for multi-class classification optimization and the Adam optimizer was utilized for gradient-based optimization. A 10-fold cross-validation approach was implemented on the discovery dataset; nine groups were used for model training and one group for evaluation. The optimal model was selected based on its ability to achieve the highest accuracy on the validation set during 10-fold cross-validation. This model was subsequently applied to an independent validation dataset consisting of 239 cancer patients across five cancer types to obtain the performance of tissue of origin classification.

Given the predictions of trained model and the graph topology, we estimated the feature importance score by the GNN Explainer [4]. The feature was considered important if it satisfied:

$$F_i > \delta_f$$

where F_i is the important score of ith feature estimated by the GNN Explainer, δ_f is the chosen cut-off and was set to 0.9.

Statistical analysis

This study used either the Wilcoxon Rank Sum test or t-test to find statistically significant differences between cancer and control. The Kolmogorov-Smirnov test was used to decide whether two cohorts have the same statistical distribution. The Benjamini-Hochberg correction was used to correct p-value for multiple comparisons (with a corrected p-value cutoff $\alpha \leq 0.05$). DeLong's test was used to compare the differences between AUCs. All statistical analyses were performed using R (4.1.0) packages, including ggplot2, pROC, and caret. 95% confident interval (95% CI) was presented in a bracket next to a value accordingly.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

References

1. H. Sung *et al.*, Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **71**, 209-249 (2021).
2. T. Pham *et al.*, Cancers in Vietnam-Burden and Control Efforts: A Narrative Scoping Review. *Cancer Control* **26**, 1073274819863802 (2019).
3. N. Hawkes, Cancer survival data emphasise importance of early diagnosis. *BMJ* **364**, l408 (2019).
4. Z. Kakushadze, R. Raghubanshi, W. Yu, Estimating Cost Savings from Early Cancer Diagnosis. *Data*. 2017 (10.3390/data2030030).
5. M. C. Liu, G. R. Oxnard, E. A. Klein, C. Swanton, M. V. Seiden, Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol* **31**, 745-759 (2020).
6. J. Li *et al.*, Clinical applications of liquid biopsy as prognostic and predictive biomarkers in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. *J Exp Clin Cancer Res* **37**, 213 (2018).
7. H. T. Nguyen *et al.*, Ultra-Deep Sequencing of Plasma-Circulating DNA for the Detection of Tumor- Derived Mutations in Patients with Nonmetastatic Colorectal Cancer. *Cancer Invest* **40**, 354-365 (2022).
8. Q. Gao *et al.*, Circulating cell-free DNA for cancer early detection. *The Innovation* **3**, 100259 (2022).
9. J. Pascual *et al.*, ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group. *Ann Oncol* **33**, 750-768 (2022).
10. H. T. Nguyen *et al.*, Evaluation of a Liquid Biopsy Protocol using Ultra-Deep Massive Parallel Sequencing for Detecting and Quantifying Circulation Tumor DNA in Colorectal Cancer Patients. *Cancer Invest* **38**, 85-93 (2020).
11. N. Constantin, A. A. Sina, D. Korbie, M. Trau, Opportunities for Early Cancer Detection: The Rise of ctDNA Methylation-Based Pan-Cancer Screening Technologies. *Epigenomes* **6**, (2022).
12. T. H. Phan *et al.*, Circulating DNA methylation profile improves the accuracy of serum biomarkers for the detection of nonmetastatic hepatocellular carcinoma. *Future Oncol* **18**, 4399-4413 (2022).
13. X. Chen *et al.*, Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nature Communications* **11**, 3475 (2020).
14. S. Cristiano *et al.*, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385-389 (2019).
15. A. Jamshidi *et al.*, Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell* **40**, 1537-1549.e1512 (2022).
16. T. Moser, S. Kühberger, I. Lazzeri, G. Vlachos, E. Heitzer, Bridging biological cfDNA features and machine learning approaches. *Trends Genet* **39**, 285-307 (2023).
17. Y. R. Im, D. W. Y. Tsui, L. A. Diaz, Jr., J. C. M. Wan, Next-Generation Liquid Biopsies: Embracing Data Science in Oncology. *Trends Cancer* **7**, 283-292 (2021).
18. Q. Zhou *et al.*, Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proceedings of the National Academy of Sciences* **119**, e2209852119 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

19. J. D. Cohen *et al.*, Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926-930 (2018).
20. H. T. Nguyen *et al.*, Multimodal analysis of ctDNA methylation and fragmentomic profiles enhances detection of nonmetastatic colorectal cancer. *Future Oncol* **18**, 3895-3912 (2022).
21. Y. Yin *et al.*, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, (2017).
22. D. Buitrago *et al.*, Impact of DNA methylation on 3D genome structure. *Nature Communications* **12**, 3243 (2021).
23. H.-N. Nguyen *et al.*, Liquid biopsy uncovers distinct patterns of DNA methylation and copy number changes in NSCLC patients with different EGFR-TKI resistant mutations. *Scientific Reports* **11**, 16436 (2021).
24. P. M. Das, R. Singal, DNA methylation and cancer. *J Clin Oncol* **22**, 4632-4642 (2004).
25. M. Ehrlich, DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400-5413 (2002).
26. M. J. Hoffmann, W. A. Schulz, Causes and consequences of DNA hypomethylation in human cancer. *Biochem Cell Biol* **83**, 296-321 (2005).
27. X. Shao *et al.*, Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med Genet* **20**, 175 (2019).
28. S. Baldacchino, G. Grech, Somatic copy number aberrations in metastatic patients: The promise of liquid biopsies. *Semin Cancer Biol* **60**, 302-310 (2020).
29. S. Knuutila *et al.*, DNA copy number losses in human neoplasms. *Am J Pathol* **155**, 683-694 (1999).
30. A. Dereli-Öz, G. Versini, T. D. Halazonetis, Studies of genomic copy number changes in human cancers reveal signatures of DNA replication stress. *Mol Oncol* **5**, 308-314 (2011).
31. K. Brennan, J. M. Flanagan, Is there a link between genome-wide hypomethylation in blood and cancer risk? *Cancer Prev Res (Phila)* **5**, 1345-1357 (2012).
32. W. Zhang *et al.*, Global DNA Hypomethylation in Epithelial Ovarian Cancer: Passive Demethylation and Association with Genomic Instability. *Cancers (Basel)* **12**, (2020).
33. K. C. A. Chan *et al.*, Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proceedings of the National Academy of Sciences* **110**, 18761-18768 (2013).
34. H. R. Underhill *et al.*, Fragment Length of Circulating Tumor DNA. *PLOS Genetics* **12**, e1006162 (2016).
35. F. Mouliere *et al.*, Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med* **10**, (2018).
36. Y. M. D. Lo, D. S. C. Han, P. Jiang, R. W. K. Chiu, Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* **372**, eaaw3616 (2021).
37. V. C. Nguyen *et al.*, Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma. *BMC Cancer* **23**, 233 (2023).
38. A. M. Raizis, F. Schmitt, J. P. Jost, A bisulfite method of 5-methylcytosine mapping that minimizes template degradation. *Anal Biochem* **226**, 161-166 (1995).
39. K. Tanaka, A. Okamoto, Degradation of DNA by bisulfite treatment. *Bioorg Med Chem Lett* **17**, 1912-1915 (2007).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

40. S. Kint, W. De Spiegelaere, J. De Kesel, L. Vandekerckhove, W. Van Criekinge, Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. *PLoS One* **13**, e0199091 (2018).
41. M. Ehrich, S. Zoll, S. Sur, D. van den Boom, A new method for accurate assessment of DNA quality after bisulfite treatment. *Nucleic Acids Res* **35**, e29 (2007).
42. P. Jiang *et al.*, Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discovery* **10**, 664-673 (2020).
43. C. Jin *et al.*, Characterization of fragment sizes, copy number aberrations and 4-mer end motifs in cell-free DNA of hepatocellular carcinoma for enhanced liquid biopsy-based cancer detection. *Mol Oncol* **15**, 2377-2389 (2021).
44. E. A. Klein *et al.*, Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol* **32**, 1167-1177 (2021).
45. P. Ilse, S. Biesterfeld, N. Pomjanski, C. Wrobel, M. Schramm, Analysis of SHOX2 Methylation as an Aid to Cytology in Lung Cancer Diagnosis. *Cancer Genomics - Proteomics* **11**, 251 (2014).
46. J. D. Warren *et al.*, Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med* **9**, 133 (2011).
47. P. A. Jones, H. Ohtani, A. Chakravarthy, D. D. De Carvalho, Epigenetic therapy in immune-oncology. *Nat Rev Cancer* **19**, 151-161 (2019).
48. P. Ulz *et al.*, Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun* **10**, 4666 (2019).
49. M. Ivanov, A. Baranova, T. Butler, P. Spellman, V. Mileyko, Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16 Suppl 13**, S1 (2015).
50. C. Caggiano *et al.*, Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE. *Nature Communications* **12**, 2717 (2021).
51. C. Yin *et al.*, Molecular Subtyping of Cancer Based on Robust Graph Neural Network and Multi-Omics Data Integration. *Front Genet* **13**, 884028 (2022).
52. Y. Huang, A. C. S. Chung, Disease prediction with edge-variational graph convolutional networks. *Med Image Anal* **77**, 102375 (2022).
53. J. Moss *et al.*, Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nature Communications* **9**, 5068 (2018).
54. N. Loyfer *et al.*, A DNA methylation atlas of normal human cell types. *Nature* **613**, 355-364 (2023).
55. X. Zhou *et al.*, Tumor fractions deciphered from circulating cell-free DNA methylation for cancer early diagnosis. *Nature Communications* **13**, 7694 (2022).
56. S. Kang *et al.*, CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* **18**, 53 (2017).
57. I. Yusipov *et al.*, Age-related DNA methylation changes are sex-specific: a comprehensive assessment. *Aging (Albany NY)* **12**, 24057-24080 (2020).
58. A. E. Field *et al.*, DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Mol Cell* **71**, 882-895 (2018).
59. T. H. H. Nguyen *et al.*, Clinical validation of a ctDNA-Based Assay for Multi-Cancer Detection: An Interim Report from a Vietnamese Longitudinal Prospective Cohort Study of 2795 Participants. *Cancer Investigation* **41**, 232-248 (2023).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

60. S. B. Edge, C. C. Compton, The American Joint Committee on Cancer: the 7th Edition of the AJCC Cancer Staging Manual and the Future of TNM. *Annals of Surgical Oncology* **17**, 1471-1474 (2010).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Acknowledgments

We thank all participants who agreed to take part in this study, and all the clinics and hospitals who assisted in patient consultation and sample collection.

Funding: The study was funded by Gene Solutions

Disclosure statement: The authors including LST, HNN, HG, MDP, HHN and DSN hold equity in Gene Solutions. The funder Gene Solutions provided support in the form of salaries for authors are inventors on the patent application (USPTO 17930705). We also confirm that this does not alter our adherence to Cancer Investigation policies on sharing data and materials.

Author contribution:
Conceptualization: DLV, THP, TXJ, VCN, HTN, TVN, MDP, HG, HNN, LST
Patient consultancy and screening: DLV, THP, TXJ, VCN, HTN, TVN, QTD, TND, AMT, VHN, TAVN, QDT, TTTP, TDH, BTN, TNVN, TDN, DTBP, BHHP, TLV, THTN, TTT, MHT, NCT, TKL, THTT, MLD, HPTB, VVK, TAP, DHT, TNAL, TVNP, MTL, DSN, VTC, TTTD, HST
Formal analysis: VTCN, THN, NNTD, TMQP, TDN, THHN, LAKH, THT, DHV, TMTT, MNN, TTVV, ANN, TTT, VUT, MPL, TTD, TVP, LHDN
Supervision: DKT
Writing-original draft: VTCN, GTHN, TTTT, LST
Writing-review and editing: VTCN, GTHN, TTTT, MDP, LST

Competing interests: The authors declare no conflict of interest.

Data and materials availability: Sequencing data will be deposited in a public portal database (NCBI SRA) upon acceptance and are available on request from the corresponding author, LST. The data are not publicly available due to ethical restrictions.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Figures and Tables

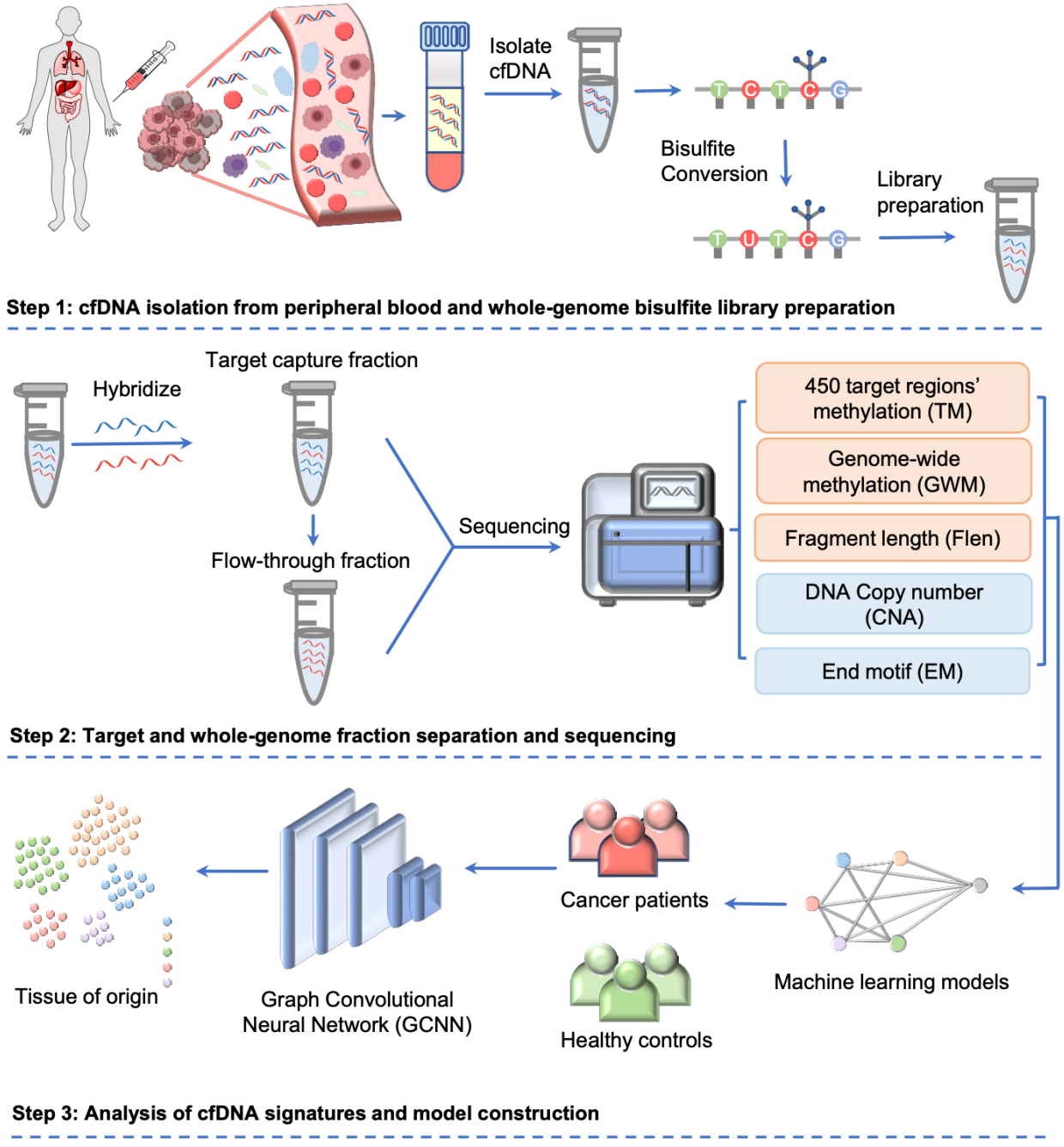

Figure 1. Workflow of SPOT-MAS assay for multi-cancer detection and localization. There are three main steps in the SPOT-MAS assay. Firstly, cfDNA is isolated from peripheral blood, then treated with bisulfite conversion and adapter ligation to make whole-genome bisulfite cfDNA library. Secondly, whole-genome bisulfite cfDNA library is subjected to hybridization by probes specific for 450 target regions to collect the target capture fraction. The whole-genome fraction was retrieved by collecting the 'flow-through' and hybridized with probes specific for adapter sequences of DNA library. Both the target capture and whole-genome fractions were subjected to massive parallel sequencing and the resulting data were pre-processed into five different features of cfDNA: Target methylation (TM), genome-wide methylation (GWM), fragment length profile (Flen), DNA copy number (CNA) and end motif

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

(EM). Finally, machine learning models and graph convolutional neural networks are adopted for classification of cancer status and identification tissue of origin.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

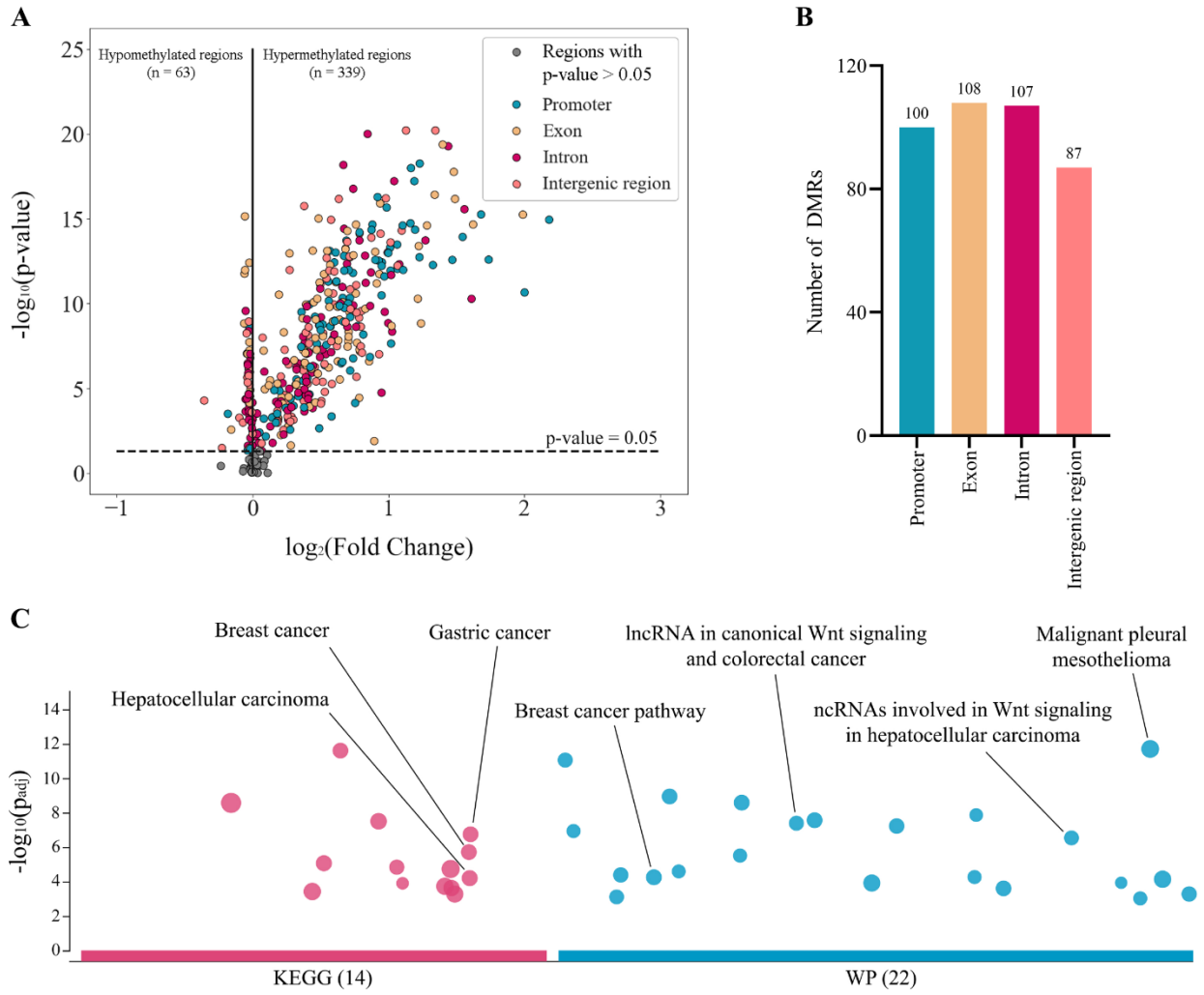

Figure 2. Analysis of targeted methylation in cfDNA. (A) Volcano plot shows $\log_2$ fold change (logFC) and significance ($-\log_{10}$ Benjamini-Hochberg adjusted p-value from Wilcoxon rank-sum test) of 450 target regions when comparing 499 cancer patients and 1,076 healthy controls in the discovery cohort. There are 402 DMRs (p-value < 0.05), color-coded by genomic locations. (B) Number of DMRs in the four genomic locations. (C) KEGG and WP pathway enrichment analysis using g:Profiler for genes associated with the DMRs. A total of 36 pathways are enriched, suggesting a link between differences in methylation regions and tumorigenesis.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

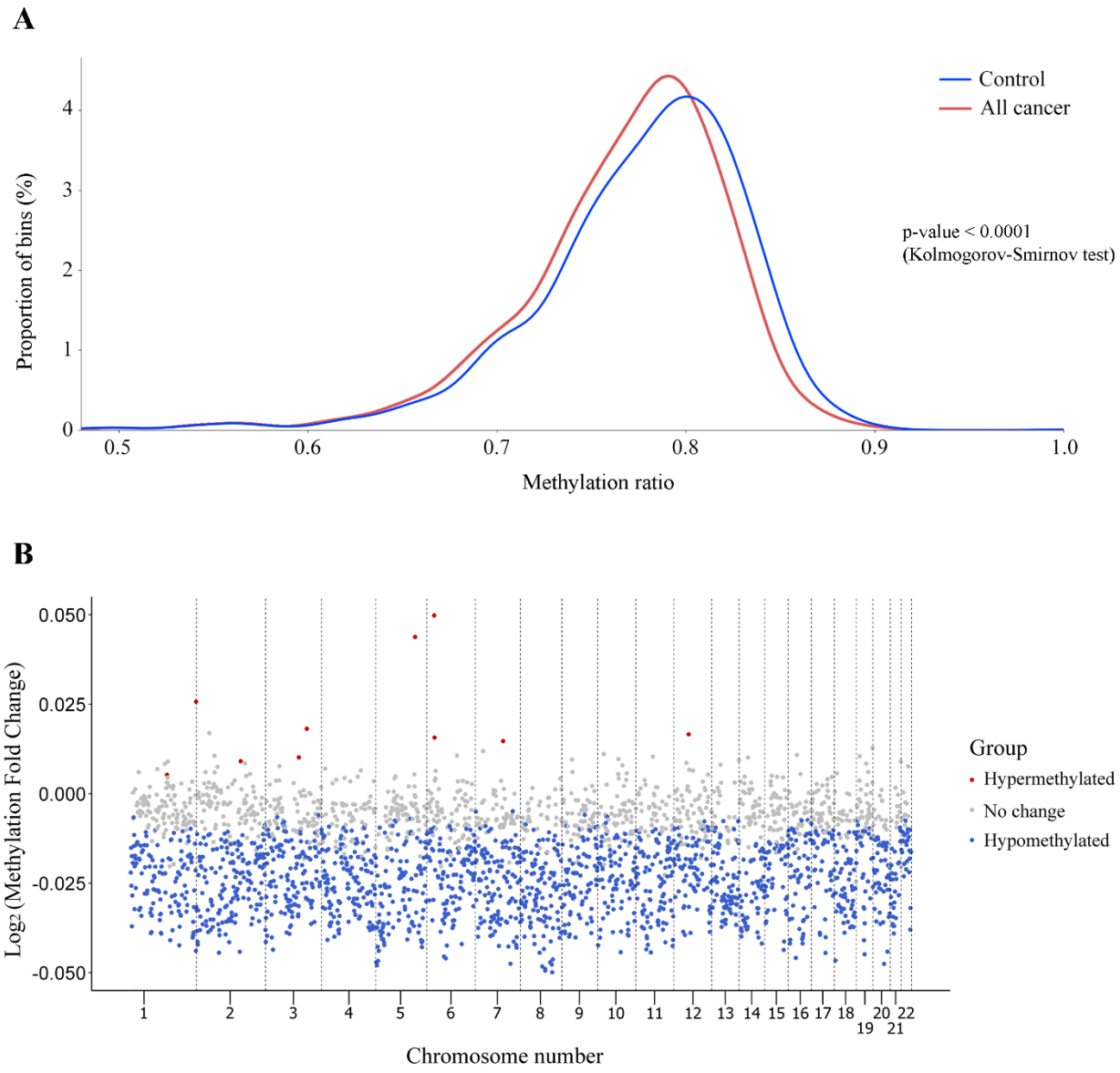

Figure 3. Genome-wide methylation changes in cfDNA of cancer patients. (A) Density plot showing the distribution of genome-wide methylation ratio for all cancer patients (red curve, n= 499) and healthy participants (blue curve, n= 1,076). The left-ward shift in cancer samples indicates global hypomethylation in the cancer genome ($p < 0.0001$, two-sample Kolmogorov-Smirnov test). (B) Log_2 fold change of methylation ratio between cancer patients and healthy participants in each bin across 22 chromosomes. Each dot indicates a bin, identified as hypermethylated (red), hypomethylated (blue), or no significant change in methylation (grey).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

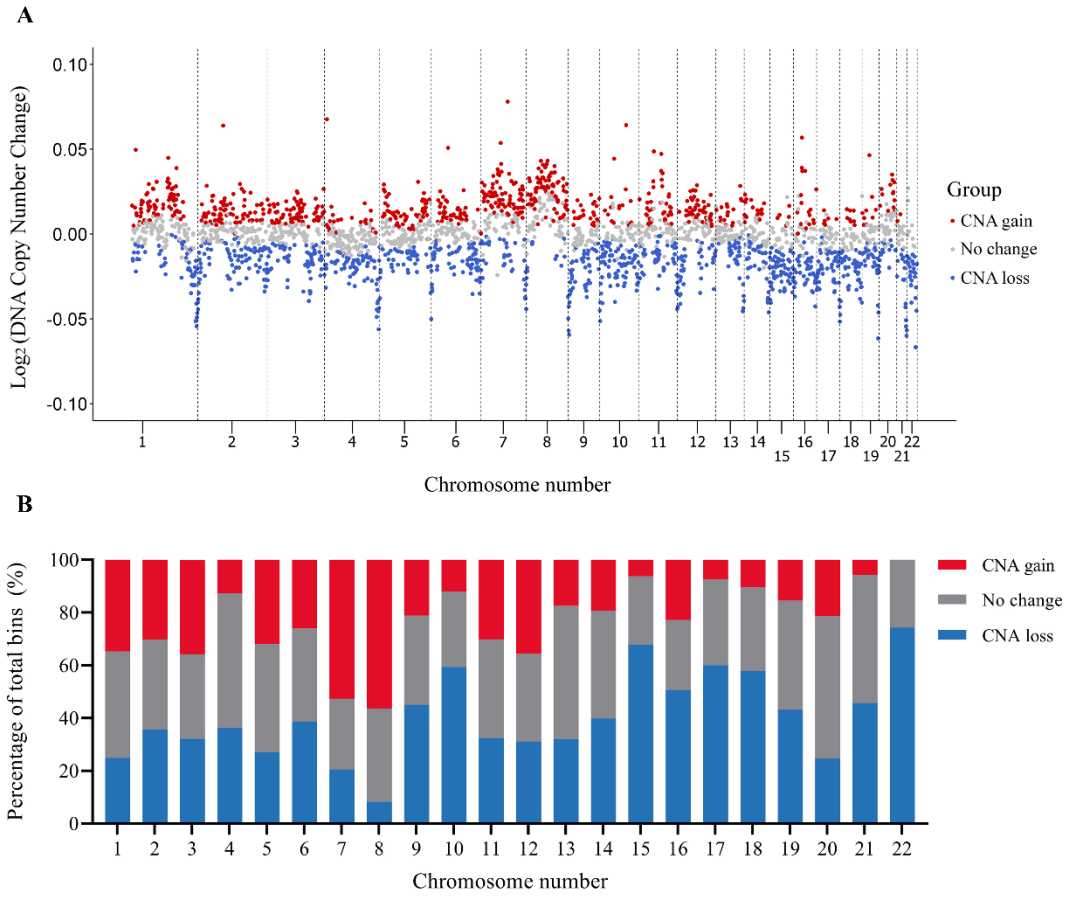

Figure 4. Analysis of copy number aberration (CNA) in cfDNA. (A) Log_2 fold change of DNA copy number in each bin across 22 autosomes between 499 cancer patients and 1,076 healthy participants in the discovery cohort. Each dot represents a bin identified as gain (red), loss (blue) or no change (grey) in copy number. (B) Proportions of different CNA bins in each autosomes.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

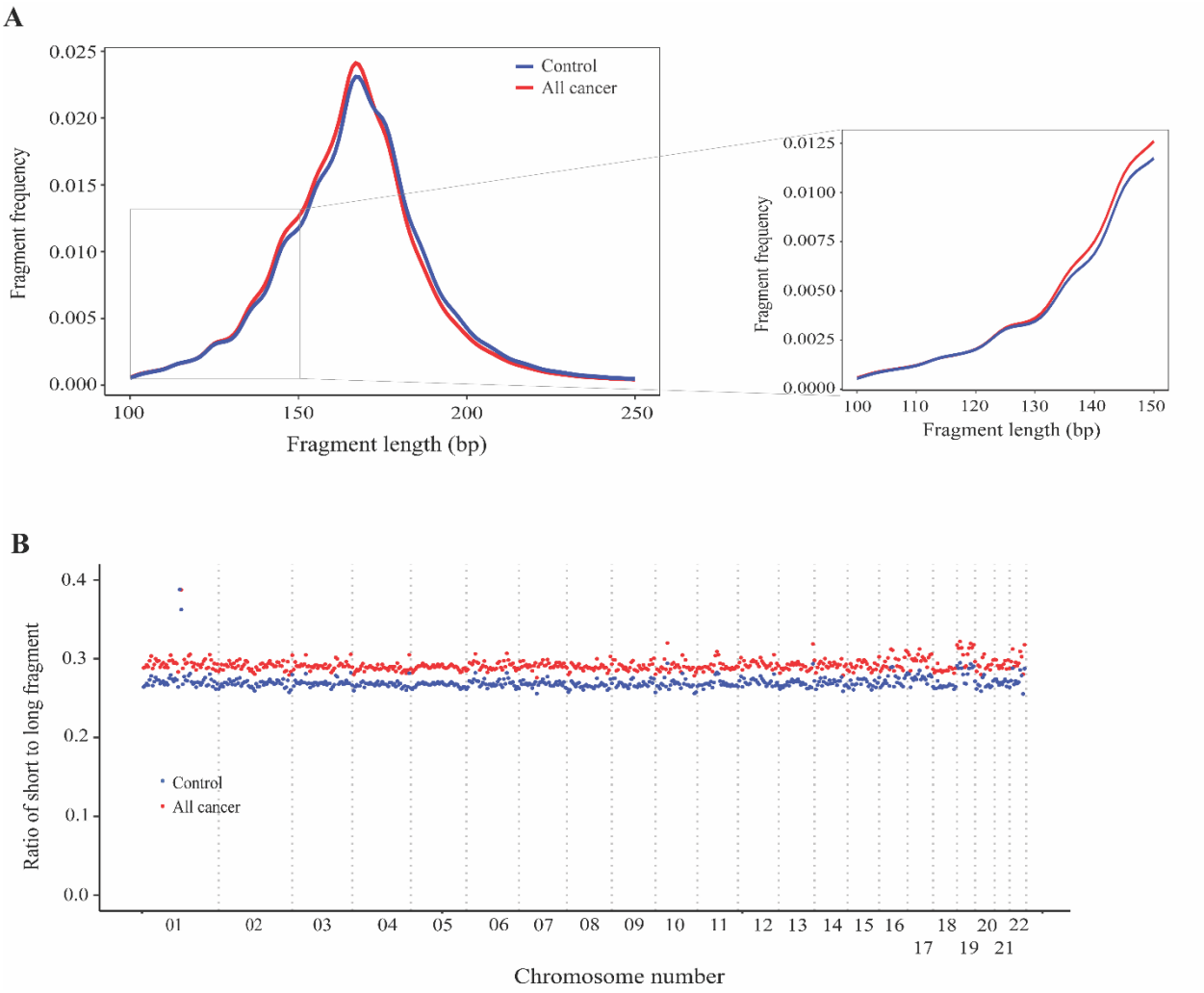

Figure 5. Analysis of fragment length patterns of ctDNA in plasma. (A) Density plot of fragment length between cancer patients (red, n=499) and healthy participants (blue, n=1,076) in the discovery cohort. Inset corresponds to an x-axis expansion of short fragment (<150 bp). (B) Ratio of short to long fragments across 22 autosomes. Each dot indicates a mean ratio for each bin in cancer patients (red) and healthy participants (blue).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

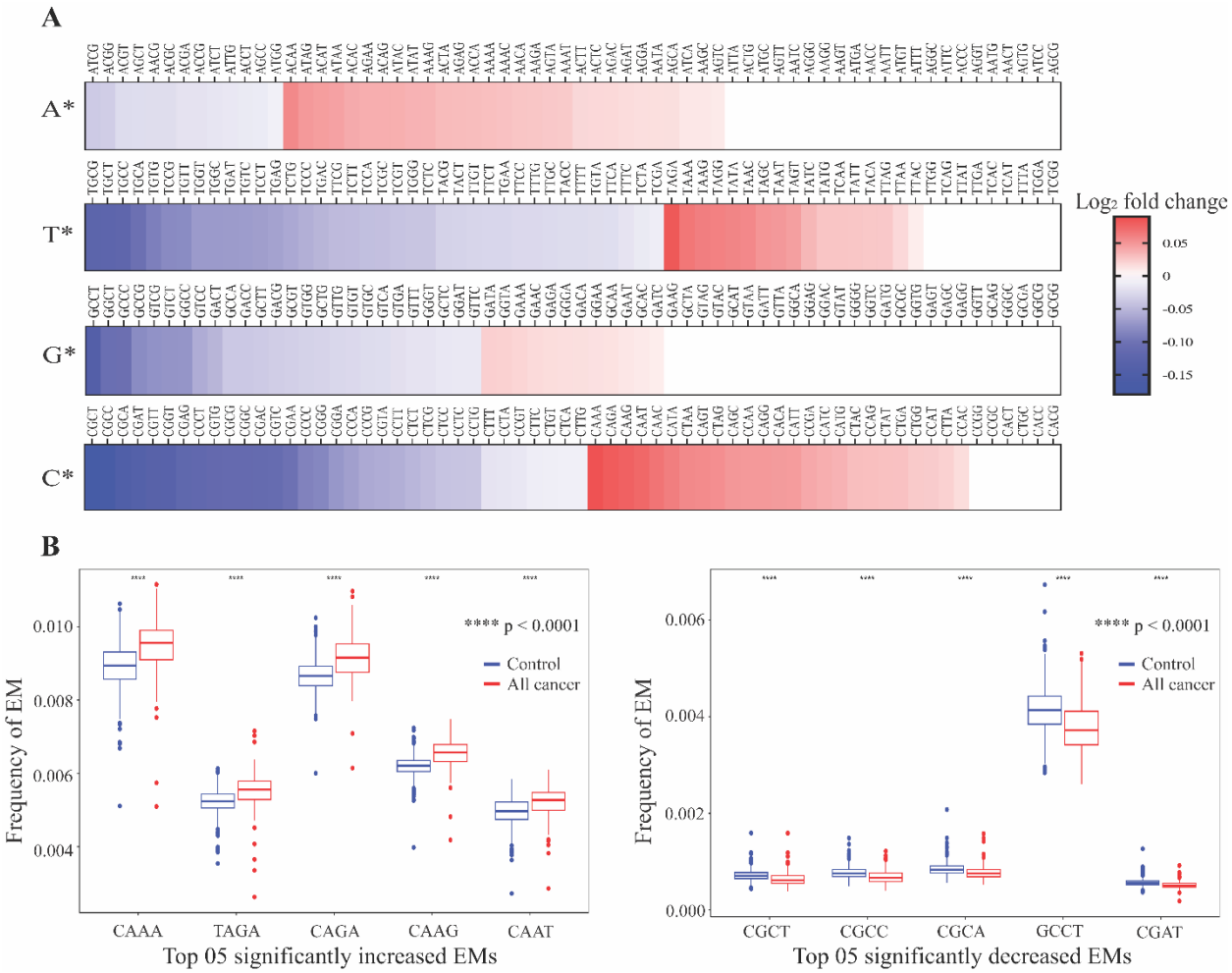

Figure 6. Differences in 4-mer end motif between cancer and healthy cfDNA. (A) Heatmap shows $\log_2$ fold change of 256 4-mer end motifs in cancer patients (n=499) compared to healthy controls (n=1,076). (B) Box plots showing the top ten motifs with significant differences in frequency between cancer patients (red) and healthy controls (blue) using Wilcoxon rank-sum test with Bonferroni-adjusted p-value < 0.0001.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

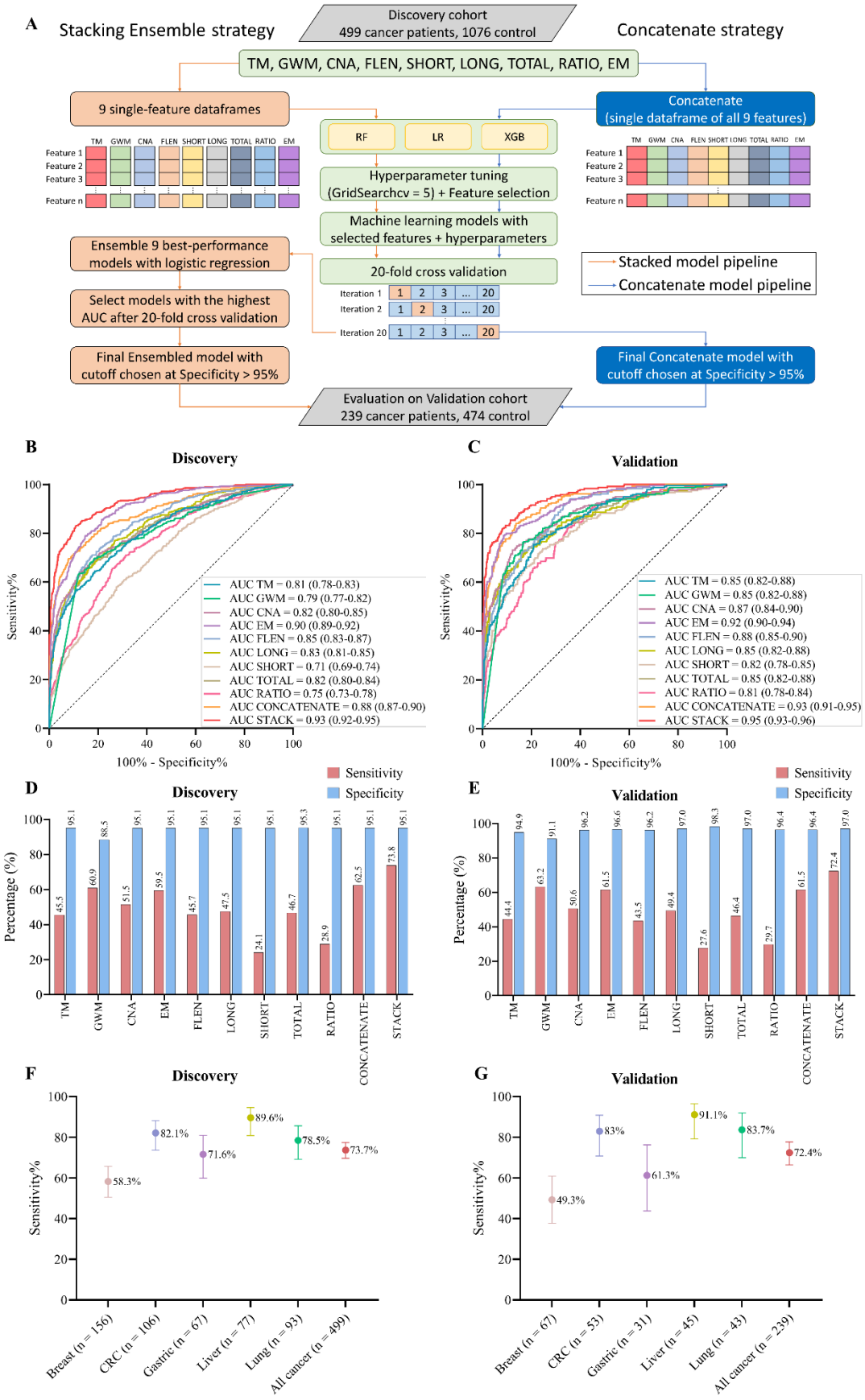

Figure 7. Model construction and performance validation for SPOT-MAS. (A) Two model construction strategies for cancer detection. (B) and (C) ROC curves comparing the

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

performance of single-feature models, and two combination models (concatenate and ensemble stacking) in the discovery (B) and validation cohorts (C). (D) and (E) Bar charts showing the specificity and sensitivity of single-feature models and two combination models (concatenate and ensemble stacking) in the discovery (D) and validation cohorts (E). (F) and (G) Dot plots showing the sensitivity of SPOT-MAS assay in detection of 5 different cancer types in the discovery (F) and validation cohorts (G). The points and error bars represent the average sensitivity over 20 runs and 95% confidence intervals. Feature abbreviations as follows: TM – target methylation density, GWM – genome-wide methylation density, CNA – copy number aberration, EM – 4-mer end motif, FLEN – fragment length distribution, LONG – long fragment count, SHORT – short fragment count, TOTAL – all fragment count, RATIO – ratio of short/long fragment.

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

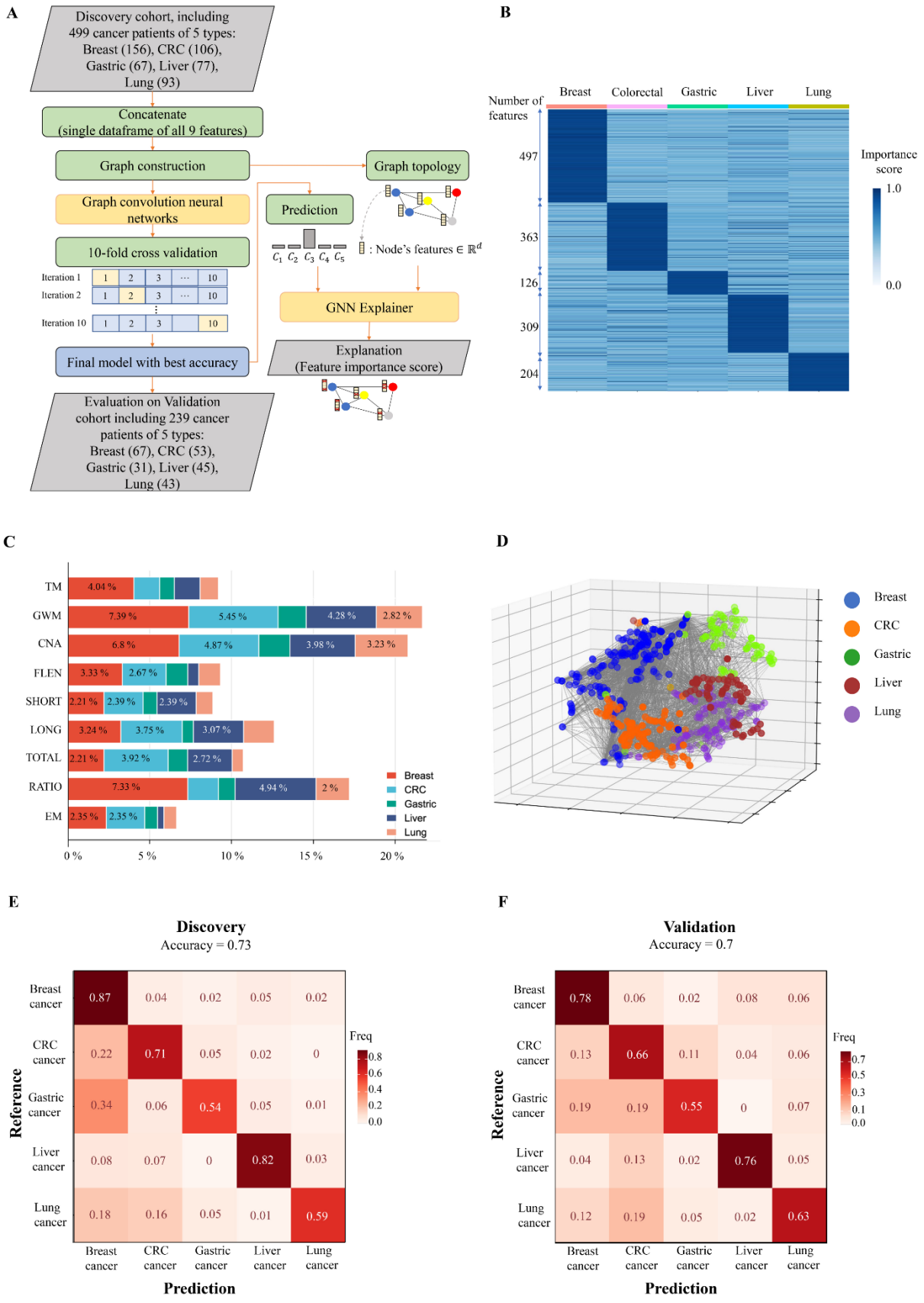

Figure 8. The performance of SPOT-MAS assay in prediction of the tissue of origin. (A) Model construction strategy to predict tissue of origin by combining nine sets of cfDNA

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

features using graph convolutional neural networks. (B) Heatmap shows feature important scores of five cancer types. (C) Bar chart indicates the contribution of important features for classifying five different cancers. (D) Three dimensions graph represents the classification of five cancer types. (E) and (F) Cross-tables show agreement between the prediction (x-axis) and the reference (y-axis) to predict tissue of origin in the discovery cohort (E) and validation cohort (F).

medRxiv preprint doi: https://doi.org/10.1101/2023.04.12.23288460; this version posted April 17, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 1. Summary of clinical features of 738 cancer patients and 1,550 healthy controls in discovery and validation cohorts.

Clinical features		Discovery cohort (N=1,575)					Validation cohort (N=713)				
		Cancer (N = 499)		Healthy (N = 1,076)		p-value (Cancer vs Healthy)	Cancer (N = 239)		Healthy (N = 474)		p-value (Cancer vs Healthy)
		N	Percentage	N	Percentage		N	Percentage	N	Percentage	
Gender	Female	279	55.9%	599	55.7%	0.9281[#]	126	52.72%	270	56.1%	0.2818[#]
	Male	220	44.1%	477	44.3%		113	47.28%	204	43.9%	
Age	Median	58		47		< 0.0001 [##]	59		48		< 0.0001 [##]
	Min	25		18			28		19		
	Max	97		84			92		85		
Stage	I	52	10.4%				23	9.6%			0.4947[#]
	II	169	33.9%				69	28.9%			
	IIIA	150	30.1%				77	32.2%			
	Non-metastasis with unknown staging information	128	25.7%				70	29.3%			

[#] P-values from Chi-square test; [##] P-values from Mann-Whitney test