Enhanced Detection of Genitourinary Cancers Using Fragmentation and Copy Number Profiles Obtained from Urinary Cell-Free DNA

Yang Han,[a,b,c,†] Xinxin Li,[a,b,†] Mingxin Zhang,[d,†] Yang Yang,[a,b,c,†] Guangzhe Ge,[a,b] Kunxiang Wang,[d] Yanqing Gong,[e,f,g] Yuan Liang,[a,b] Haitao Niu,[d,*] and Weimin Ci[a,b,c,h,*]

BACKGROUND: Recent studies have reported that examining the fragmentation profiles (FP) of plasma cell-free DNA (cfDNA) further improves the clinical sensitivity of tumor detection. We hypothesized that considering the differences of the FP of urinary cfDNA would increase the clinical sensitivity of genitourinary (GU) cancer detection.

METHODS: 177 patients with GU cancer and 94 individuals without tumors were enrolled in the discovery cohort. An independent validation dataset comprising 30 patients without tumors and 66 patients with GU cancer was also collected. We constructed an ensemble classifier, GUIDER, to detect and localize GU cancers using fragmentation and copy number profiles obtained from shallow whole-genome sequencing of urinary cfDNA.

RESULTS: Urinary cfDNA of patients with GU cancer had a higher proportion of long fragments (209–280 bp) and a lower proportion of short fragments (140–208 bp) compared to controls. The overall mean classification accuracy of the FP was 74.62%–85.39% for different algorithms, and integration of the FP and copy number alteration (CNA) features further enhanced the classification of samples from patients with GU cancer. The mean diagnostic accuracy was further improved by the ensemble classifier GUIDER, which integrated the FP and CNA profiles and resulted in a higher mean accuracy (87.52%) compared to the analysis performed without FP features (74.62%). GUIDER performed well in an independent validation dataset.

CONCLUSIONS: The lengthening and shortening of urinary cfDNA within specific size ranges were identified in patients with GU cancer. Integration of the FP should further enhance the ability to use urinary cfDNA as a molecular diagnostic tool.

Introduction

By 2020, the estimated incidence of genitourinary (GU) cancers (kidney cancer, urothelial cancer, and prostate cancer) will be over 2 million worldwide (1). Surgery can cure the majority of localized cancers without systemic therapy. The prognosis is poor, however, when the cancer metastasizes to distant sites. Therefore, the detection and localization of GU cancers is key to improving the clinical outcome. Unfortunately, the currently used biomarkers and diagnostic strategies are not clinically sensitive enough and are sometimes invasive (2–4).

Recently, the analysis of circulating cell-free DNA (cfDNA) has been increasingly used for the detection and monitoring of cancers (5–7). For GU cancers, the detection of urinary cfDNA may be a particularly attractive alternative to reveal cancer signals easily due to the direct contact of urinary flow-through with GU organs. Previous studies have shown that cancer-associated copy number alterations (CNAs) of urinary cfDNA function as promising molecular features that can be detected cost-efficiently via the collection of shallow whole-genome sequencing data (8–11). The above described classifiers performed well in the detection and monitoring of urothelial cancers; however, the accuracy of the localization of the tissue of origin for GU cancers was suboptimal. (10, 11). Integration with other molecular characteristics of urinary cfDNA to further improve the diagnostic accuracy is warranted.

However, any fundamental molecular characteristics of urinary cfDNA in patients with cancer remain unclear. In particular, previous studies on the size of

[a] Key Laboratory of Genomics and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing China; [b] China National Center for Bioinformation, Beijing China; [c] University of the Chinese Academy of Sciences, Beijing China; [d] Department of Urology, the Affiliated Hospital of Qingdao University, Qingdao China; [e] Department of Urology, Peking University First Hospital, Beijing China; [f] Institute of Urology, Peking University, Beijing China; [g] National Urological Cancer Center, Beijing China; [h] Institute of Stem Cells and Regeneration, Chinese Academy of Sciences, Beijing China.

* Address correspondence to this author at: Hospital of Qingdao university, 16 Jiangsu Road, Qingdao, China. Fax 0532-82911329; e-mail niuht0532@126.com. W.C. at Building 104, NO.1 Beichen West Road, Chaoyang District, Beijing, China. Fax 86-010-84097720; e-mail ciwm@big.ac.cn.

† These authors contributed equally to this work.

Received July 6, 2020; accepted October 23, 2020.
DOI: 10.1093/clinchem/hvaa283

circulating cfDNA, especially plasma cfDNA, in patients with cancer led to inconsistent results for different cancer types and patients (12–15). The fragmentation profiles (FP) of urinary cfDNA from patients with GU cancer have not been comprehensively evaluated. Here, we evaluated the FP of urinary cfDNA from patients with GU cancer and control individuals without tumors. We developed an ensemble classifier named GUIDER (GenitoUrinary cancer non-Invasive Diagnosis by Ensemble Recognition) to detect and classify GU cancers by integrating FP and CNAs. We found that GUIDER enhanced the detection and localization of GU cancers.

Materials and Methods

SAMPLE COLLECTION AND PROCESSING

This study was approved by the Ethics Committee of the Affiliated Hospital of Qingdao University and Peking University First Hospital, and informed consent was obtained from all participants who provided urine. Fresh urine samples from 30 healthy individuals (who had no previous history of cancer and negative screening results) and patients with GU cancers [26 clear-cell renal cell carcinoma (KIRC) and 40 urothelial (UC)] before surgery were collected in 50 mL Falcon tubes and immediately processed within 30 min. The urine supernatant and sediment were separated by centrifugation at 4000 g for 10 min at 4 °C and stored in 50 mL Falcon tubes and 1.5 mL Axygen microcentrifuge tubes, respectively, at -80 °C until use.

LIBRARY CONSTRUCTION AND SEQUENCING OF URINARY CFDNA

The urinary cfDNA was size-selected, and the library was constructed as previously described (10). Briefly, the urinary cfDNA was first size-selected by AMPure XP beads to enrich the fragments with a size in the range of approximately 100 to 300 bp. The libraries were sequenced on a HiSeq × 10 system to generate 2 × 150 bp paired-end reads according to the manufacturer's protocols.

EVALUATION OF THE FRAGMENTATION AND CNA PROFILES ACCORDING TO THE SHALLOW WHOLE-GENOME SEQUENCING DATA IN VARIABLE-LENGTH BINS

As previously described (10), we applied the varbin (9, 16, 17) algorithm to extract the copy number profiles from the shallow whole-genome sequencing (sWGS) data in variable-length bins (∼50 000 genomic bins) by using uniform expected unique read counts. The unique mapped reads were counted and normalized according to the library size and GC content using locally weighted regression (LOWESS) statistics to calculate the GC-corrected ratio for each bin. The ratios obtained for the 50 000 bins were used as feature values for further analysis.

For the FP, the size of the fragments was determined by the CollectInsertSizeMetrics tool in Picard using bam files. The fragmentation features were defined as the ratio of the number of short fragments (140–208 bp) to the number of long fragments (209–280 bp) in each of the abovementioned 50 000 bins. Then, the ratios were normalized by the z-score algorithm in R version 3.6.1 and used as FP features for further analysis.

GUIDER MODEL TRAINING AND EVALUATION

First, we selected the FP and CNA features by using the gradientBoosting (18, 19) algorithm for each class in comparison to another class. The features with importance scores > 0 were retained for model training. To avoid gender bias, we excluded the sex chromosomes from feature selection. The selected FP and/or CNA features were used to train and evaluate the GUIDER model by randomly partitioning the cohort data into training and test sets 10 times based on a 75%/25% split. The GUIDER model was built based on 5 different machine learning algorithms: Logistic (LogisticRegression), LRmul (LogisticRegression with multinomial parameter), SVM (support vector machine), SVM_line (support vector machine with linear kernel), and Ridge (RidgeRegression) (20). The final label of each cfDNA sample was determined according to the highest mean probability predicted by the 5 different algorithms.

DATA AVAILABILITY

The raw sequence data reported here have been deposited in the Genome Sequence Archive (21) in the BIG Data Center (22), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (accession numbers PRJCA001138 and PRJCA002669).

Results

DATA GENERATION AND ANALYSIS

In this study, 177 patients with GU cancer, including 79 patients with urothelial carcinoma (UC), 54 patients with clear-cell renal cell carcinoma (KIRC), 44 patients with prostate adenocarcinoma (PRAD), and 94 control individuals without tumors, were included in the discovery cohort (Fig. 1) (10). An independent validation dataset comprising 30 patients without tumors, 40 patients with UC, and 26 patients with KIRC was also collected (Supplemental Table 1). Urinary samples were collected after admission and before surgery, and to ensure DNA integrity, cfDNA was extracted immediately

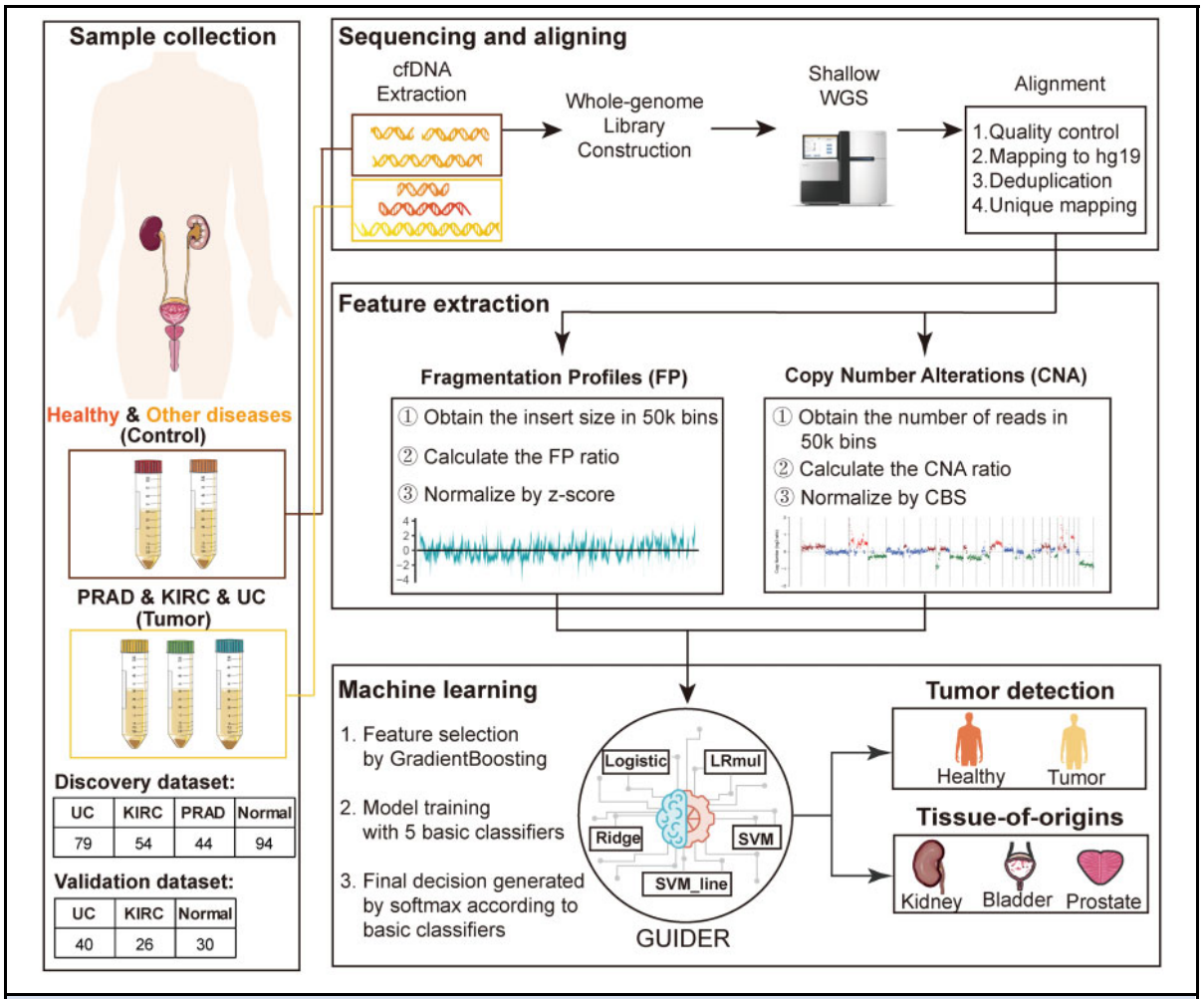


Fig. 1. Workflow chart of data generation and GUIDER classifier development for noninvasive detection and localization of genitourinary cancers. The classifier was named GUIDER (GenitoUrinary cancer non-Invasive Diagnosis by Ensemble Recognition). Five machine learning algorithms are implemented in GUIDER: Logistic (LogisticRegression); LRmul (LogisticRegression with multinomial parameters); SVM (support vector machine); SVM_line (support vector machine with a linear kernel); Ridge (RidgeRegression).

after urinary collection, which was followed by library construction and sWGS. The fragmentation profiles and CNA profiles were evaluated using the sWGS data described earlier. Then, we developed an approach called 'GenitoUrinary cancer non-Invasive Diagnosis by Ensemble Recognition' (GUIDER) (Fig. 1) to detect and localize GU cancer in patients using the fragmentation profiles and CNA profiles of urinary cfDNA.

ABERRANT URINARY CFDNA FRAGMENTATION PROFILES IN PATIENTS WITH GU CANCERS

We generated a catalog of genome-wide cfDNA fragmentation features from the discovery dataset obtained from 177 GU cancers and 94 nontumor controls. In general, the most prominent peak was observed at 142 bp, with an approximately 10 bp periodicity in the size distribution plot of each subject. The size distribution of the cfDNA fragments within the size ranges of 140–208 bp and 209–280 bp in patients with cancer differed compared to those in control individuals (Fig. 2A, Supplemental Fig. 1). Notably, we found that patients with cancer had a higher proportion of long fragments (209–280 bp) but a lower proportion of short fragments (140–208 bp) compared to control individuals. Thus, we defined the FP feature as the ratio of short cfDNA fragments (140–208 bp) to long cfDNA fragments (209–280 bp).

Next, to examine the differences in the fragment profiles in a position-dependent manner throughout the

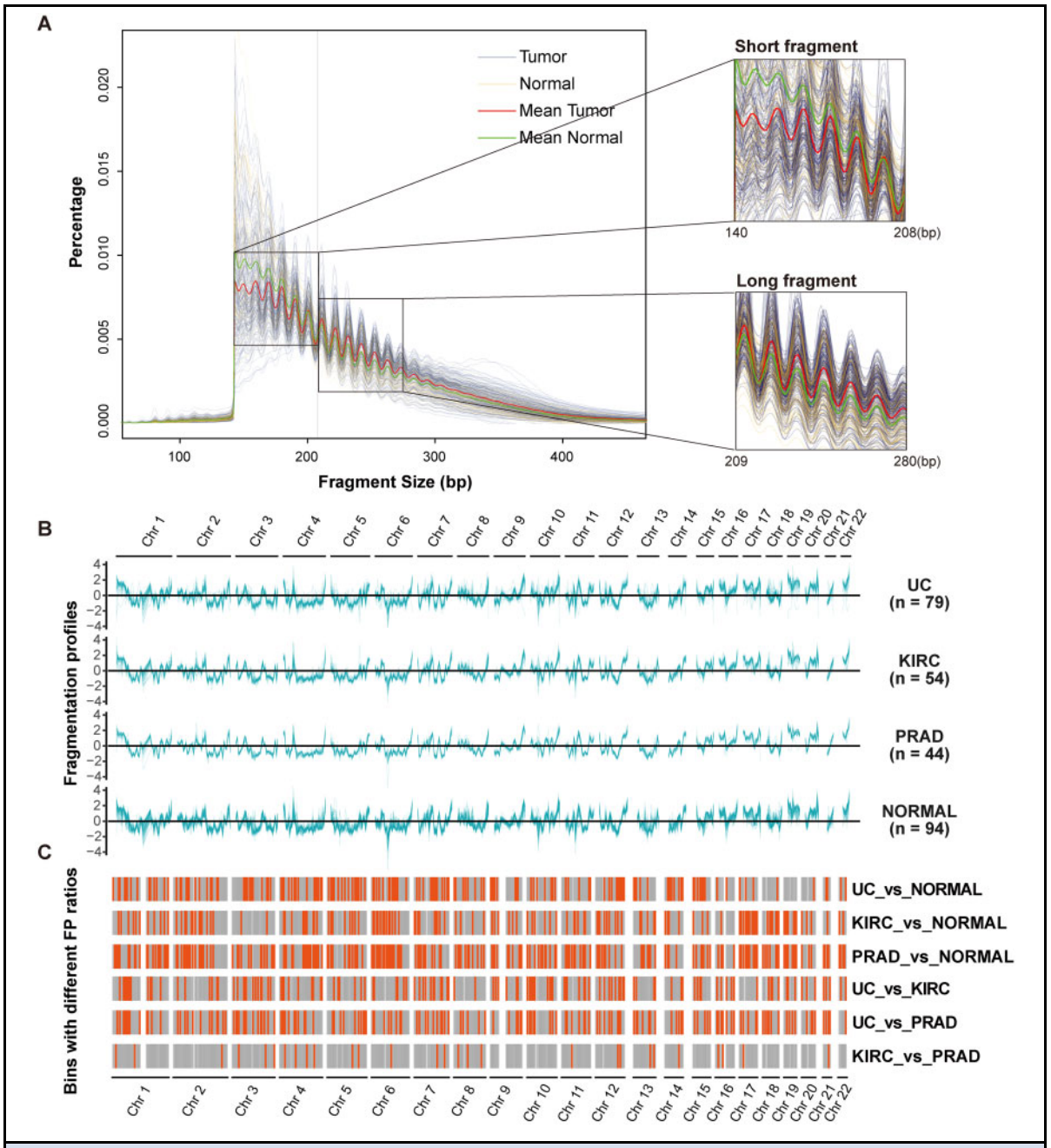


Fig. 2. Fragmentation profiles of urinary cfDNA. (A) Overall size distribution of urinary cfDNA between the tumor (UC, KIRC, and PRAD) and normal groups. (B) Fragmentation profiles of urinary cfDNA using 5 MB bins normalized by the z-score algorithm are depicted for each tumor type. (C) Pairwise comparison of the fragmentation profiles in the different tumor types and the normal group. Bins are indicated in orange if one type of sample has a fragment ratio that is significantly different from that of another type of sample (*t*-test, $P < 0.05$).

genome, we first divided the genome into nonoverlapping 5-megabase (MB) bins that were uniform in terms of the expected unique read counts, as described earlier. We mapped the fragments to their genomic origin and evaluated the FP features in each of the variable-length bins (~500 genomic bins). The analysis of the altered genomic bins revealed a median of 151 affected bins among patients with GU cancer and control individuals

(Fig. 2B), which highlighted the position-dependent alterations in the fragmentation profiles of the cfDNA.

THE DIAGNOSTIC PERFORMANCE OF THE FRAGMENTATION PROFILES OF URINARY CFDNA

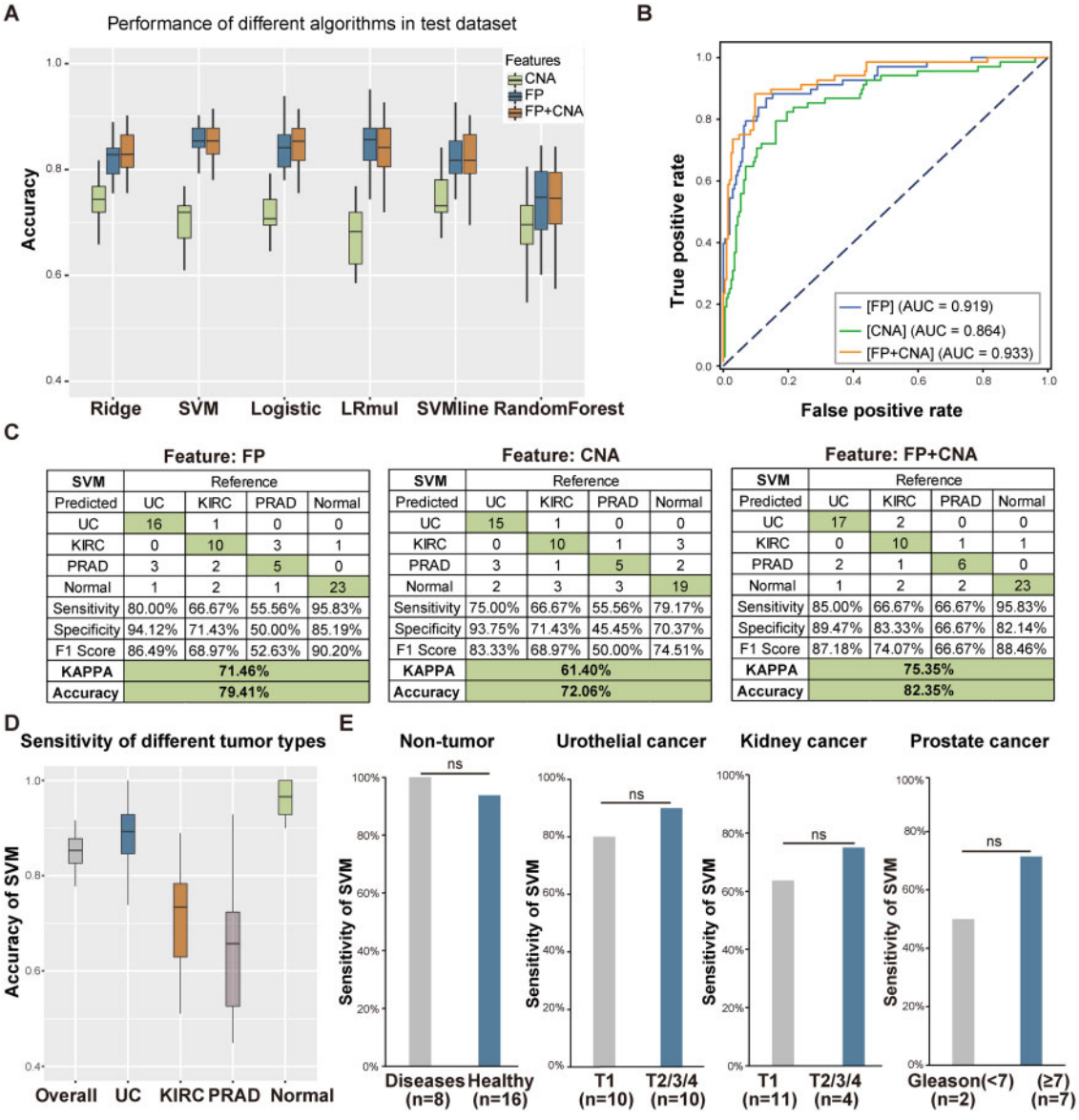
Since the fragmentation profiles revealed regional differences among the GU cancer tissues, we used machine learning methods to identify the tissue of origin of the cfDNA. First, we extracted the high resolution FP of the urinary cfDNA and divided the genome into 50 000 bins that were ∼50 kb in length each with the varbin algorithm (9, 16, 17); then, feature selection was performed with the gradientBoosting algorithm. The selected features were further trained and validated by 6 different machine learning algorithms (Fig. 3A). The overall mean classification accuracy of the FP was 74.62%–85.39% for the different algorithms for the test dataset using 10 cross-validations (Fig. 3A). Notably, the performance of the FP features was better than that of the CNA features for the tested algorithms, and integrating the FP and CNA features further slightly enhanced the identification and classification of samples (Fig. 3A), suggesting that the fragmentation pattern of cfDNA is an important predictive feature. Moreover, the SVM obtained both the best prediction performance and robustness and was selected for the detection and localization of genitourinary cancers (Fig. 3A). As shown in Fig. 3B, an increased area under the curve (AUC) was obtained for the SVM algorithm by integrating the FP and CNA features compared to that obtained by using either feature alone. The confusion matrices for 1 representative run of the SVM model for the training data set and the test samples are shown in Fig. 3C. The SVM model could localize genitourinary cancers to the tissue of origin in 85.39% of individuals, with clinical sensitivities of 96.53%, 89.31%, 73.35%, 65.75% for the nontumor, UC, KIRC, and PRAD groups for 10 randomly partitioned sets of training/test data, respectively (Fig. 3D). Moreover, the classifiers based on the SVM algorithm performed equally well for both high-grade/late-stage tumors and low-grade/early-stage UC and KIRC tumors (Fig. 3E).

DETECTION OF GU CANCERS BY AN ENSEMBLE LEARNING MODEL, GUIDER, BY COMBINING FRAGMENTATION PROFILES AND CNA PROFILES

Next, we built an ensemble learning model, GUIDER, to use multiple machine learning algorithms to obtain better predictive performance than that could be obtained from any of the constituent learning algorithms alone. The incorporation of both the FP and CNA profiles further improved the prediction performance of GUIDER. After integrating the FP and CNV profiles, GUIDER obtained the best prediction performance (Fig. 4A). A high AUC was observed (Fig. 4B and Supplemental Fig. 2). The confusion matrices of 1 representative run of the GUIDER model for the training data set and the test samples are shown in Fig. 4C. GUIDER improved the classification of the sample and obtained an AUC = 0.983, with a mean clinical sensitivity of 96.78%, 95.45%, 74.33%, 66.71% in the nontumor, UC, KIRC, and PRAD cohorts for 10 randomly partitioned sets of training/test data, respectively (Fig. 4D). Moreover, GUIDER also performed well in high-stage tumors and low-stage tumors from both healthy and nontumor disease individuals, indicating that our model shows potential for early tumor diagnosis (Fig. 4E).

However, the prediction performance and robustness of GUIDER in classifying PRAD was the worst (Fig. 4, B and C). This result may be due to the relatively lower percentage of tumor fragments in patients with prostate cancer that released cfDNA during the draining of genitourinary organs compared to that in patients with urothelial cancer and kidney cancer. Moreover, if we excluded prostate cancer from the classification, the classification accuracy was further increased for UC and KIRC (Supplemental Fig. 3A and 3B).

To further validate the performance and robustness of GUIDER in the detection and classification of UC and KIRC, we collected urine samples from 66 patients with UC, 26 patients with KIRC from another clinical center (Supplemental Table 1), and 30 individuals without tumors as controls. These samples were treated and analyzed in the same way. In the independent validation dataset, the fragment distribution of UC versus that of normal individuals and that of KIRC versus that of normal individuals were consistent with the distribution in the test dataset (Supplemental Fig. 3C). In addition, we found that the accuracy of the binary mode of GUIDER reached 94.29%, 87.50%, and 90.91% for the comparisons of UC vs normal, KIRC vs normal and UC vs KIRC (Supplemental Fig. 3D) and that the AUCs were 0.981, 0.938, and 0.955, respectively (Fig. 5B). When we used GUIDER in 3-class mode (UC, KIRC, and nontumor), the accuracy of the diagnostic model still reached 83.33%, with an AUC of 0.937 (Fig. 5A and B). Compared to that of the test dataset, the performance of the independent validation dataset was slightly decreased, but it was still acceptable. Notably, the GUIDER diagnosis model had similar high performance for tumors at a low stage and a high stage as well as in comparing individuals with diseases and healthy individuals, which indicates that our model has high potential for early tumor diagnosis and is minimally affected by the presence of other nontumor diseases (Fig. 5C).



SVM	Reference			
Predicted	UC	KIRC	PRAD	Normal
UC	16	1	0	0
KIRC	0	10	3	1
PRAD	3	2	5	0
Normal	1	2	1	23
Sensitivity	80.00%	66.67%	55.56%	95.83%
Specificity	94.12%	71.43%	50.00%	85.19%
F1 Score	86.49%	68.97%	52.63%	90.20%
KAPPA	71.46%			
Accuracy	79.41%			

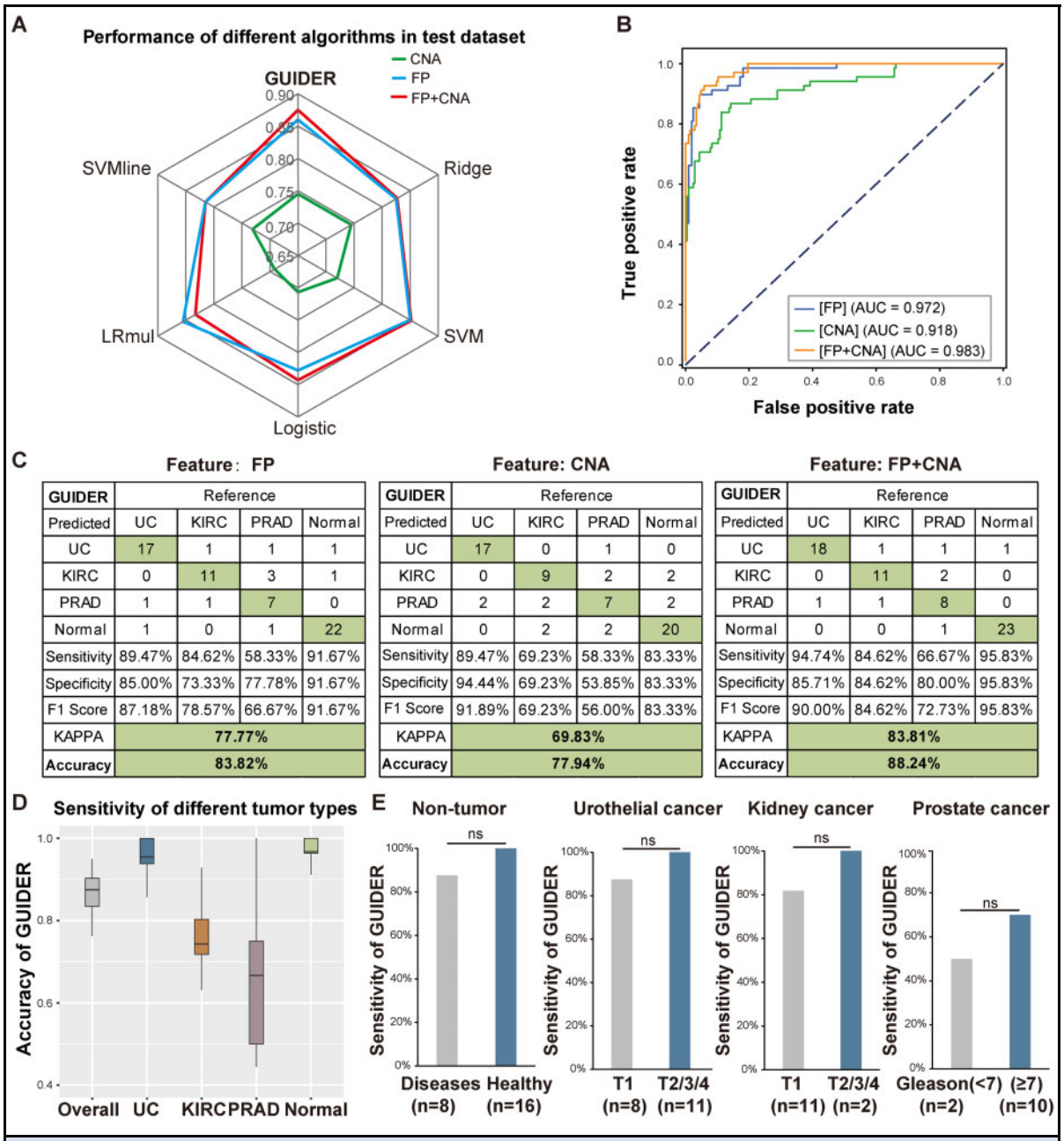
SVM	Reference			
Predicted	UC	KIRC	PRAD	Normal
UC	15	1	0	0
KIRC	0	10	1	3
PRAD	3	1	5	2
Normal	2	3	3	19
Sensitivity	75.00%	66.67%	55.56%	79.17%
Specificity	93.75%	71.43%	45.45%	70.37%
F1 Score	83.33%	68.97%	50.00%	74.51%
KAPPA	61.40%			
Accuracy	72.06%			

SVM	Reference			
Predicted	UC	KIRC	PRAD	Normal
UC	17	2	0	0
KIRC	0	10	1	1
PRAD	2	1	6	0
Normal	1	2	2	23
Sensitivity	85.00%	66.67%	66.67%	95.83%
Specificity	89.47%	83.33%	66.67%	82.14%
F1 Score	87.18%	74.07%	66.67%	88.46%
KAPPA	75.35%			
Accuracy	82.35%			

Fig. 3. Diagnostic performance of the fragmentation profiles. (A) Box plot of the classification accuracy of the 6 machine learning algorithms: Ridge, SVM, logistic, LRmul, SVMline, and random forest. The discovery cohort was randomly split into training (75%) and test sets (25%) 10 times. (B) ROC curve of the diagnostic performance of the SVM classifier. (C) Confusion matrices of one representative run of the SVM multiclass model (nontumor, KIRC, UC, and PRAD). (D) Box plot of the classification accuracy of each tumor type or normal group of the SVM multiclass model in 10 randomly partitioned sets of training/test data (75%/25%). (E) The clinical sensitivity of the SVM multiclass classifier was equally good for the diagnosis of patients with both high-stage and low-stage GU cancer and both patients with nontumor diseases and healthy individuals. Statistical significance was evaluated by Fisher's exact tests; ns indicates not significant.

Discussion

The size of plasma DNA molecules is one of the most fundamental diagnostic biomarkers (23–25). There are a number of inconsistencies in the literature, which may be due to differences in cancer types and/or the applied size detection methodologies used in different studies (15, 24, 26–28). This study was hence designed to explore the urinary cfDNA size profile of patients with GU cancer according to high-resolution WGS data. In



GUIDER	Reference			
Predicted	UC	KIRC	PRAD	Normal
UC	17	1	1	1
KIRC	0	11	3	1
PRAD	1	1	7	0
Normal	1	0	1	22
Sensitivity	89.47%	84.62%	58.33%	91.67%
Specificity	85.00%	73.33%	77.78%	91.67%
F1 Score	87.18%	78.57%	66.67%	91.67%
KAPPA	77.77%			
Accuracy	83.82%			

GUIDER	Reference			
Predicted	UC	KIRC	PRAD	Normal
UC	17	0	1	0
KIRC	0	9	2	2
PRAD	2	2	7	2
Normal	0	2	2	20
Sensitivity	89.47%	69.23%	58.33%	83.33%
Specificity	94.44%	69.23%	53.85%	83.33%
F1 Score	91.89%	69.23%	56.00%	83.33%
KAPPA	69.83%			
Accuracy	77.94%			

GUIDER	Reference			
Predicted	UC	KIRC	PRAD	Normal
UC	18	1	1	1
KIRC	0	11	2	0
PRAD	1	1	8	0
Normal	0	0	1	23
Sensitivity	94.74%	84.62%	66.67%	95.83%
Specificity	85.71%	84.62%	80.00%	95.83%
F1 Score	90.00%	84.62%	72.73%	95.83%
KAPPA	83.81%			
Accuracy	88.24%			

Fig. 4. An ensemble learning model, GUIDER (GenitoUrinary cancer non-Invasive Diagnosis by Ensemble Recognition), was used to generate multiple machine learning algorithms to obtain better predictive performance. (A) GUIDER outperformed the other 5 multiclass machine learning algorithms. (B) ROC curve of the diagnostic performance of GUIDER in the discovery dataset. (C) Confusion matrices of one representative run of GUIDER (nontumor, KIRC, UC, and PRAD). (D) Box plot of the classification accuracy of each tumor type or normal group of the SVM multiclass model in 10 randomly partitioned sets of training/test data (75%/25%). (E) The clinical sensitivity of GUIDER was equally good in the diagnosis of patients with both high-stage and low-stage GU cancer and both patients with nontumor diseases and healthy individuals. Statistical significance was evaluated by Fisher's exact tests; ns indicates not significant.

A

Independent validation dataset

GUIDER	Reference		
Predicted	UC	KIRC	Normal
UC	35	4	0
KIRC	4	17	2
Normal	1	5	28
Sensitivity	87.50%	65.38%	93.33%
Specificity	89.74%	73.91%	82.35%
F1 Score	88.61%	69.39%	87.50%
KAPPA	74.56%		
Accuracy	83.33%		

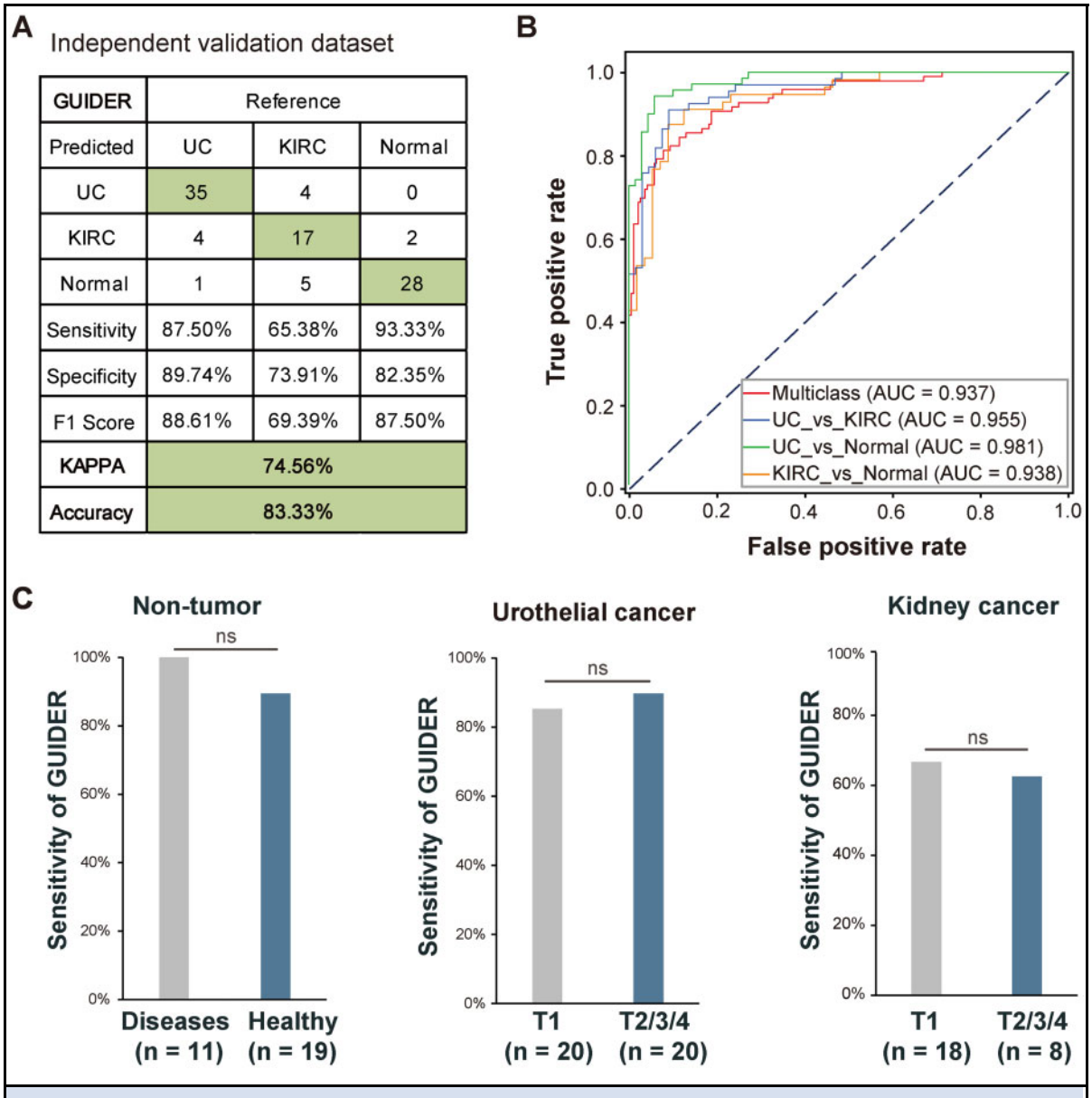


Fig. 5. Diagnostic performance of GUIDER in the independent validation dataset. (A) Confusion matrices of GUIDER in the independent validation dataset (nontumor, UC, and KIRC). (B) ROC curve of the diagnostic performance of GUIDER in the independent validation dataset. (C) The clinical sensitivity of GUIDER was equally good in the diagnosis of patients with both high-stage and low-stage GU cancer and both patients with nontumor diseases and healthy individuals. Statistical significance was evaluated by Fisher's exact tests; ns indicates not significant.

general, the most prominent peak was at 142 bp with a 10 bp periodicity in the size distribution plot, suggesting that most of the urinary cfDNA was derived from apoptotic cells. We found that the size of urinary cfDNA molecules changed differently in patients with GU cancer and control individuals without tumors within different size ranges, as shown by the shortening of the short fragments (140–208 bp) and the lengthening of the long fragments (209–280 bp). This observation is consistent with a previous report of plasma cfDNA in liver cancer with different size ranges, which showed the shortening of fragments shorter than 150 bp and the lengthening of fragments longer than 180 bp (25), and this suggested that the fragmentation of urinary cfDNA

might have different mechanisms compared with those of the fragmentation of plasma cfDNA. Similarly, a recent study also showed that the fragment size of urinary cfDNA from patients with muscle-invasive bladder cancer but not that from patients with nonmuscle invasive bladder cancer is longer than that from nontumor controls (11). Of note, the size range of urinary cfDNA in a study (11) that was determined using 41 bp X 2 paired-end sequencing appeared to be shorter than that shown in our study, which was determined using 150 bp X 2 paired-end sequencing. Differences in cfDNA purification and sequencing library preparation may also have contributed to the differences. Moreover, it will definitely worthwhile to try other different definition of size range according to the fragmentation profiles (Fig. 2), such as including the reads smaller than 142 bp for the short fragment, and including the reads longer than 280 bp for the long fragment.

Meanwhile, we found that the accuracy of detecting patients with prostate cancer was still suboptimal. Integrating other features, such as point mutations, protein analytes, and DNA methylation, may further improve the performance of the test in patients with prostate cancer. Consistently, our previous study showed that the detection accuracy of patients with prostate cancer was enhanced by using urine sediment DNA methylomes and CNA profiles (29). Moreover, we must take sex differences into account when analyzing the prostate cancer data. When performing intertumor-type comparisons, we excluded sex chromosomes from feature selection. Previous studies, however, have shown that multiple genomic and epigenomic alterations of the sex chromosome are also important for the occurrence of prostate cancer (30, 31). Therefore, by comparing and contrasting the omics features of urinary cfDNA and urine sediment, it is possible to build a more accurate and robust prostate cancer prediction model by taking sex chromosome differences into account in further studies.

However, the exact mechanism resulting in the differences in the sizes of cfDNA fragments originating from tumors and normal tissue remains elusive. It has been postulated that the length and distribution of cfDNAs are closely related to the structure of the nucleosome (32, 33). In addition, the change in the DNA methylation level caused by tumorigenesis may change the stability and accessibility of nucleosomes, resulting in differences in the fragment lengths of cfDNA (34–39). Mechanistically, a recent study has shown that the fragmentation pattern of plasma cfDNA is generated by deoxyribonuclease 1 (DNASE1), deoxyribonuclease 1-like 3 (DNASE1L3), and DNA fragmentation factor subunit beta (DFFB) (40). Further understanding of the size and other biological properties contributing to urinary cfDNA fragmentation profiles could further enhance the development of new diagnostic approaches.

In summary, urinary cfDNA fragment size analysis via paired-end sequencing and ensemble machine learning approaches enhanced the noninvasive detection and localization of GU cancer. Combining fragmentation profiles and CNA profiles further enhanced the performance of GUIDER. Additionally, FP and CNA profiles can simultaneously be measured by sWGS with considerable cost efficiency. With the gradual decrease in the cost of next-generation sequencing, FP and CNA detection in urinary cfDNA by using sWGS may be implemented as a routine test for the screening and management of patients with GU cancers.

Nonstandard Abbreviations: cfDNA, cell-free DNA; ctDNA, circulating tumor DNA; GU, genitourinary; FP, fragmentation profiles; CNA, copy number alterations; KIRC, kidney renal clear cell carcinoma; UC, urothelial carcinoma; PRAD, prostate adenocarcinoma; UCB, urothelial carcinoma of the bladder; sWGS, shallow whole genome sequencing; MB, megabase; ROC, receiver operating characteristic; AUC, area under the curve; Logistic, LogisticRegression; Ridge, RidgeRegression; LRmul, LogisticRegression with multinomial parameters; SVM, support vector machine; SVM_line, support vector machine with a linear kernel; GUIDER, GenitoUrinary cancer non-Invasive Diagnosis by Ensemble Recognition; LOWESS, locally weighted regression

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.*

W. Ci and H. Niu conceptualized the project. Y. Han, X. Li, and Y. Yang conducted the experiments and analyzed the sWGS data. M. Zhang and K. Wang collected the urine samples and performed the pathological review. G. Ge, Y. Gong, and Y. Liang provided technical support. W. Ci, Y. Han, and X. Li wrote the manuscript. All coauthors have critically reviewed the manuscript.

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: This work was supported by the CAS Strategic Priority Research Program (XDA16010102 to W. Ci), the National Key R&D Program of China (2018YFC2000100, 2019YFA0110900 to W. Ci), CAS (QYZDB-SSW-SMC039 to W. Ci), the National Natural Science Foundation of China (81672541 to W. Ci), the K.C. Wong Education Foundation (to W. Ci).
Expert Testimony: None declared.

Patents: None declared.
Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

References

1. Bagheri MH, Ahlman MA, Lindenberg L, Turkbey B, Lin J, Cahid Civelek A, et al. Advances in medical imaging for the diagnosis and management of common genitourinary cancers. Urol Oncol 2017;35:473–91.
2. Linehan WM, Ricketts CJ. Decade in review-kidney cancer: discoveries, therapies and opportunities. Nat Rev Urol 2014;11:614–6.
3. Babjuk M, Bohle A, Burger M, Capoun O, Cohen D, Comperat EM, et al. EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: Update 2016. Eur Urol 2017;71:447–61.
4. Margulis V, Shariat SF, Matin SF, Kamat AM, Zigeuner R, Kikuchi E, et al. Outcomes of radical nephroureterectomy: a series from the upper tract urothelial carcinoma collaboration. Cancer 2009;115:1224–33.
5. Chaudhuri AA, Chabon JJ, Lovejoy AF, Newman AM, Stehr H, Azad TD, et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. Cancer Discov 2017;7:1394–403.
6. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 2018;563:579–83.
7. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018;359:926–30.
8. Molparia B, Nichani E, Torkamani A. Assessment of circulating copy number variant detection for cancer screening. PLoS One 2017;12:e0180647.
9. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90–4.
10. Ge G, Peng D, Guan B, Zhou Y, Gong Y, Shi Y, et al. Urothelial carcinoma detection based on copy number profiles of urinary cell-free DNA by shallow whole-genome sequencing. Clin Chem 2020;66:188–98.
11. Cheng THT, Jiang P, Teoh JYC, Heung MMS, Tam JCW, Sun X, et al. Noninvasive detection of bladder cancer by shallow-depth genome-wide bisulfite sequencing of urinary cell-free DNA for methylation and copy number profiling. Clin Chem 2019;65:927–36.
12. Chan KC, Leung SF, Yeung SW, Chan AT, Lo YM. Persistent aberrations in circulating DNA integrity after radiotherapy are associated with poor prognosis in nasopharyngeal carcinoma patients. Clin Cancer Res 2008;14:4141–5.
13. Gao YJ, He YJ, Yang ZL, Shao HY, Zuo Y, Bai Y, et al. Increased integrity of circulating cell-free DNA in plasma of patients with acute leukemia. Clin Chem Lab Med 2010;48:1651–6.
14. Umetani N, Kim J, Hiramatsu S, Reber HA, Hines OJ, Bilchik AJ, Hoon DS. Increased integrity of free circulating DNA in sera of patients with colorectal or periampullary cancer: direct quantitative PCR for ALU repeats. Clin Chem 2006;52:1062–9.
15. Wang BG, Huang HY, Chen YC, Bristow RE, Kassauei K, Cheng CC, et al. Increased plasma DNA integrity in cancer patients. Cancer Res 2003;63:3966–8.
16. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, et al. Genome-wide copy number analysis of single cells. Nat Protoc 2012;7:1024–41.
17. Ulz P, Belic J, Graf R, Auer M, Lafer I, Fischereder K, et al. Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. Nat Commun 2016;7:12008.
18. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38:367–78.
19. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232.
20. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.
21. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome sequence archive. Genomics Proteomics Bioinformatics 2017;15:14–8.
22. Members BIGDC. Database resources of the big data center in 2018. Nucleic Acids Res 2018;46:D14–D20.
23. Jiang P, Lo YMD. The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. Trends Genet 2016;32:360–71.
24. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 2019;570:385–9.
25. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci USA 2018;115:E10925–E33.
26. Umetani N, Giuliano AE, Hiramatsu SH, Amersi F, Nakagawa T, Martino S, Hoon DS. Prediction of breast tumor progression by integrity of free circulating DNA in serum. J Clin Oncol 2006;24:4270–6.
27. Vizza E, Corrado G, De Angeli M, Carosi M, Mancini E, Baiocco E, et al. Serum DNA integrity index as a potential molecular biomarker in endometrial cancer. J Exp Clin Cancer Res 2018;37:16.
28. Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, Baker DN, et al. Fragment length of circulating tumor DNA. PLoS Genet 2016;12:e1006162.
29. Xu Z, Ge G, Guan B, Lei Z, Hao X, Zhou Y, et al. Noninvasive detection and localization of genitourinary cancers using urinary sediment DNA methylomes and copy number profiles. Eur Urol 2020;77:288–90.
30. Prensner JR, Feng FY. "Lincing" the y chromosome to prostate cancer: Ttty15 takes center stage. Eur Urol 2019;76:327–8.
31. Xiao G, Yao J, Kong D, Ye C, Chen R, Li L, et al. The long noncoding RNA TTTY15, which is located on the Y chromosome, promotes prostate cancer progression by sponging let-7. Eur Urol 2019;76:315–26.
32. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. BMC Genomics 2015;16:S1.
33. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell 2016;164:57–68.
34. Lee JY, Lee TH. Effects of DNA methylation on the structure of nucleosomes. J Am Chem Soc 2012;134:173–5.
35. Choy JS, Wei S, Lee JY, Tan S, Chu S, Lee TH. DNA methylation increases nucleosome compaction and rigidity. J Am Chem Soc 2010;132:1782–3.
36. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 2007;8:286–98.
37. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.
38. Jensen TJ, Kim SK, Zhu Z, Chin C, Gebhard C, Lu T, et al. Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. Genome Biol 2015;16:78.
39. Sun K, Jiang P, Wong AIC, Cheng YKY, Cheng SH, Zhang H, et al. Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. Proc Natl Acad Sci USA 2018;115:E5106–E14.
40. Han DSC, Ni M, Chan RWY, Chan VWH, Lui KO, Chiu RWK, Lo YMD. The biology of cell-free DNA fragmentation and the roles of DNASE1, DNASE1l3, and dffb. Am J Hum Genet 2020;106:202–14.