UCseek: ultrasensitive early detection and recurrence monitoring of urothelial carcinoma by shallow-depth genome-wide bisulfite sequencing of urinary sediment DNA

Ping Wang,[a,b,c,i] Yue Shi,[b,i] Jianye Zhang,[a,d,i] Jianzhong Shou,[e,i] Mingxin Zhang,[f] Daojia Zou,[b,c] Yuan Liang,[b] Juan Li,[b,c] Yezhen Tan,[b,c] Mei Zhang,[b] Xingang Bi,[e,****] Liqun Zhou,[a,d,g,***] Weimin Ci,[b,c,h,**] and Xuesong Li[a,d,g,*]

[a]Department of Urology, Peking University First Hospital, Beijing, 100034, China
[b]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing, 100101, China
[c]University of Chinese Academy of Sciences, Beijing, 100049, China
[d]Institute of Urology, Peking University, Beijing, 100034, China
[e]Department of Urology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China
[f]Department of Urology, The Affiliated Hospital of Qingdao University, Qingdao, 266003, China
[g]National Urological Cancer Center, Beijing Key Laboratory of Urogenital Diseases (Male) Molecular Diagnosis and Treatment Center, Beijing, 100034, China
[h]Institute for Stem Cell and Regeneration, Chinese Academy of Sciences, Beijing, 100101, China

eBioMedicine
2023;89: 104437
Published Online xxx
https://doi.org/10.1016/j.ebiom.2023.104437

Summary

Background Current methods for the detection and surveillance of urothelial carcinomas (UCs) are often invasive, costly, and not effective for low-grade, early-stage, and minimal residual disease (MRD) tumors. We aimed to develop and validate a model from urine sediments to predict different grade and stage UCs with low cost and high accuracy.

Methods We collected 167 samples, including 90 tumors and 77 individuals without tumors, as a discovery cohort. We assessed copy number variations and methylation values for them and constructed a diagnostic classifier to detect UC, UCseek, by using an individual read-based method and support vector machine. The performance of UCseek was validated in an independent cohort derived from three hospitals (n = 206) and a relapse cohort (n = 42) for monitoring recurrence.

Findings We constructed UCseek, which could predict UCs with high sensitivity (92.7%), high specificity (90.7%), and high accuracy (91.7%) in the independent validation set. The accuracy of UCseek in low-grade and early-stage patients reached 91.8% and 94.3%, respectively. Notably, UCseek retained great performance at ultralow sequencing depths (0.3X-0.5X). It also demonstrated a powerful ability to monitor recurrence in a surveillance cohort compared with cystoscopy (90.91% vs. 59.09%).

Interpretation We optimized an improved approach named UCseek for the noninvasive diagnosis and monitoring of UCs in both low- and high-grade tumors and in early- and advanced-stage tumors, even at ultralow sequencing depths, which may reduce the burden of cystoscopy and blind second surgery.

Funding A full list of funding bodies that contributed to this study can be found in the Acknowledgments section.

Keywords: Urothelial carcinoma; Molecular diagnostics; Tumor markers; Diagnosis; Relapse monitoring; Machine learning

*Corresponding author.
**Corresponding author.
***Corresponding author.
****Corresponding author.
E-mail addresses: pineneedle@sina.com (X. Li), ciwm@big.ac.cn (W. Ci), zhoulqmail@sina.com (L. Zhou), bixingang@csco.org.cn (X. Bi).
[i]These authors contributed equally to this work.

Research in context

Evidence before this study

Urothelial carcinoma (UC) is considered the type of cancer with the highest cost/patient ratio due to its high recurrence and postsurgical monitoring. Certain FDA-approved tools, such as urine cytology, UroVysion, and cystoscopy, are used to detect UC but have some disadvantages of being inaccurate, invasive, and expensive.

At present, many studies have focused on the diagnostic models of UC, most of which are based on a single feature, such as methylation alterations or copy number variations, with suboptimal specificity or sensitivity, especially for early, low-grade and minimal residual UCs. Therefore, there is an urgent need for a sensitive, noninvasive, convenient and affordable technology to replace these conventional methods to some extent.

Added value of this study

In this study, we optimized an augmentation method named "UCseek" and further validated it in a retrospective multicenter cohort. The results showed that the sensitivity and specificity of UCseek in the independent validation set were 92.7% and 90.7%, respectively, which showed potential clinical feasibility. Furthermore, we demonstrate that UCseek also has good diagnostic performance for low-grade and early-stage UC. Meanwhile, UCseek performed well at ultralow WGBS sequencing depth (0.3X-0.5X), which was not observed in previous liquid biopsy studies. In addition, we demonstrate the high accuracy of this method for monitoring UC recurrence in a multitime-point surveillance cohort, presenting the high sensitivity of this method for minimal residual disease.

Implications of all the available evidence

The results of the study proved that the tumor markers we screened were tumor specific, and UCseek proved to have good diagnostic performance in early-stage, low-grade patients and was validated in a multicenter independent validation set. Therefore, this method demonstrates its potential in clinical application for noninvasive routine diagnosis and recurrence monitoring of patients, which can further reduce the burden on patients.

Introduction

Urothelial carcinoma (UC) is a malignancy of the urinary tract originating in the bladder, renal pelvis, and/or ureter (upper tract urothelial carcinoma, UTUC).[1] The high recurrence rate and ongoing invasive monitoring requirements, such as cystoscopy and ureteroscopy, are the key contributors to the economic and human toll of this disease.[2] Another commonly used tool for diagnosing and monitoring bladder cancer is urine cytology,[3] which is highly specific but lacks sensitivity (25%–35%), especially for low-grade tumors (4%–15%)[1,3,4] and UTUC.[5,6] Several different noninvasive urine-based tests approved by the FDA, such as NMP22 and UroVysion (FISH), also have low sensitivity for low-grade or small tumors.[7,8] Therefore, it is necessary to develop sensitive, specific, noninvasive, convenient, and cost-effective assays to complement the current clinical practice of UCs, particularly for low-grade and early-stage tumors.[9]

A rapidly increasing number of genomic and epigenomic studies have demonstrated the potential of urinary DNA for the diagnosis and surveillance of UCs. For example, Uroseek detects 11 gene mutations and 39 chromosomal abnormalities but needs to be combined with urine cytology to increase sensitivity.[10] EpiCheck uses a set of 15 methylation markers to detect recurrence in muscle-invasive bladder cancer patients, but the overall sensitivity is only 68.2%.[11] UroCAD was built by incorporating all the autosomal chromosomal CNVs for UCs and has low sensitivity (65.6%) in low-grade tumors. Previously, we found that the genomic heterogeneity of CNV profiles was less prevalent between UTUC and bladder cancer.[12] Recently, our epigenomic study showed that UTUCs and bladder cancers showed higher similarities in terms of DNA methylation profiles than SNP profiles.[12–14] Thus, we hypothesized that noninvasive biomarkers for UCs based on CNV and DNA methylation will perform better in both bladder cancers and UTUCs. Indeed, we developed UCdetector[12] for UCs and GUseek for genitourinary cancer.[15] However, further validation is needed in multicenter and large-scale cohorts.

In this study, we optimized an enhanced approach named ‘UCseek’ and further validated it in a retrospective multicenter cohort. Furthermore, UCseek outperforms our previous UC diagnostic method, namely, "GUseek", in both predictive performance and robustness. Compared with GUseek, the performance of UCseek was evaluated in a validation dataset, including low-grade and early-stage UCs and ultralow WGBS sequencing depths (0.3X-0.5X), as well as in a cohort of relapsed patients.

Methods

Study design and participants

These samples were obtained from randomly voided midstream urine in the hospital (Peking University First Hospital, the Affiliated Hospital of Qingdao University and Cancer Hospital Chinese Academy of Medical Sciences) with written informed consent obtained from all patients. All of the participants were aged ≥18 years. All urine samples were collected for testing prior to standard-of-care treatments, such as transurethral resection of bladder tumor (TURBT), cystectomy, cystoscopy, and other procedures. The pathology diagnosis results were reviewed by at least two independent pathologists. In the discovery cohort, according to the Chinese guidelines for the diagnosis and treatment of urothelial carcinoma of the bladder, the incidence of UC was approximately 3 times higher in men than in women. Therefore, based on this information, UC samples with a similar sex ratio in the discovery cohort were collected. The nontumor samples were collected randomly. In the independent validation cohort, all the samples were randomly collected. Sample size calculation was performed according to the accuracy and tolerance of UC diagnosis. Since no data were available to extrapolate the accuracy of UCseek for UC diagnosis, we expected to find a sensitivity of 90% and a specificity of 90% according to previous studies analyzing the biomarker in UC. With a 95% confidence interval (CI) and 10% precision, a minimum of 70 patients were required in the current study.

Then, we collected 167 urine samples obtained from Peking University First Hospital, including samples from 90 tumor patients and 77 control individuals without tumors from the cohort study "CAS Precision Medicine Initiative" with shared medical data and cancer-free urological diseases such as renal or ureteral calculus and ureterostenosis in the discovery cohort. To maintain the excellent performance of the UCseek model developed from the discovery cohort, we calculated the number of samples required for an independent validation set according to the sample size estimation method. In the independent validation cohort, samples were collected from three different centers, Peking University First Hospital (n = 54), the Affiliated Hospital of Qingdao University (n = 30) and Cancer Hospital Chinese Academy of Medical Sciences (n = 55), as well as 67 volunteers, which included 109 tumor samples and 97 control individuals without tumors (individuals from the cohort study "CAS Precision Medicine Initiative" with shared medical data and cancer-free urological diseases such as renal or ureteral calculus and ureterostenosis). Additionally, 54 tissue samples were collected, including 45 UC and 9 normal tissue samples. Moreover, we downloaded SNP6.0 array copy number data from the TCGA database to verify the reliability of the copy number marker. In addition, we followed up 42 patients at multiple time points in an independent validation set. Eight of the patients were newly diagnosed patients, and the remaining 34 patients were followed up. Urine samples from 12 patients were collected from two follow-up visits as a multi-item point monitoring dataset.

DNA extraction and WGBS library construction

Urine samples were collected using a cell preservation solution kit and shipped to the laboratory within 48 h. Urine (50 ml) was centrifuged to obtain the sediment, and DNA was extracted using Qiagen Cat#: 51306. DNA purity and quantity were examined with Qubit 4.0. Subsequently, we performed WGBS-seq library construction by using Ultra II End Repair/dA-Tailing and Ultra II Ligation modules (NEB Cat#: E7546 and E7595 L) and an EZ DNA Methylation-Gold kit (Zymo Cat#: D5005). Finally, all the libraries were sequenced on a NovaSeq 6000 system to generate 2X150 bp paired-end reads according to the manufacturer's protocols.

Tumor-specific methylation analysis at the block level

A total of 147,888 blocks of tightly coupled CpG sites, called methylation haplotype blocks (MHBs), were generated in a previous study[16] by combining WGBS data. Meanwhile, we identified ∼24,0000 UC-specific MHBs in tumor tissues from UC patients (n = 4). We performed feature selection using the combined ≈380,000 MHBs. The methylation haplotype load (MHL) for each MHB in urine sediment was then calculated. For each MHB, if the proportion of missing values (NAs) in the sample was greater than 10%, this MHB region was discarded; otherwise, the NA in the sample was replaced by the average MHL value of the MHB. We then calculated the difference for each MHB between the control and tumor groups using a *t test* to reduce the number of markers, followed by LASSO analysis to obtain the most significantly different methylation markers.

Identification of CNV markers for discriminating between UC and noncancer urine sediments

As previously described, we applied the varbin algorithm to extract the copy number profiles from WGBS data in variable-length bins (∼50000 genomic bins) by using uniform expected unique read counts.[17–19] Unique reads were counted and normalized using the LOWESS statistical method, and finally, the GC-corrected ratio of each bin was obtained. Afterward, the same feature screening method as the methylation marker was used to obtain the most significantly different copy number variation markers.

Construction and validation of the methylation model

For the methylation model, we calculated the methylation scores based on the tumor-specific methylation markers selected above. In brief, we classified each read of the samples into either a tumor-derived DNA class (abbreviated as T class) or a nontumor urine-derived DNA class (abbreviated as N class) and then predicted tumor-derived methylation scores through the following process. Step 1: (1) We fitted the distributions of those markers in the nontumor group and tumor group according to Equation (1). (2) According to Equation

(2), we calculated the probability that each read belongs to the nontumor or tumor groups. (3) According to the probability of all reads, we constructed a maximum likelihood function using Equation (3) to estimate the tumor score, called the methylation score. Step 2: In the discovery set, the methylation scores were used for model building by randomly partitioning the data into training and test sets based on a 70%/30% split to obtain the Youden index more than 100 times, and finally, we selected the methylation scores corresponding to the best Youden index as the threshold.[20] Compared with GUseek, the methylation model we adopted is based on the status of each read of the sample to deconvolute the patient's tumor score, while GUseek is based on the status of a sample. Such a modeling approach can effectively expand the ability to extract signals from significantly differentially methylated sites.

$$\begin{cases} f\left(marker^{kT}\right) = Beta\left(\eta^{kT}, \rho^{kT}\right) \\ f\left(marker^{kN}\right) = Beta\left(\eta^{kN}, \rho^{kN}\right) \end{cases} \qquad \text{Equation (1)}$$

$$\begin{cases} p(r|m)^{N} = \prod_{j} p\left(r_j \middle| Beta\left(\eta^{N}, \rho^{N}\right)\right) \\ p(r|m)^{T} = \prod_{j} p\left(r_j \middle| Beta\left(\eta^{T}, \rho^{T}\right)\right) \end{cases} \qquad \text{Equation (2)}$$

$$\begin{cases} P\left(r^{i} \middle| \theta, m\right) = \theta * p\left(r^{i} \middle| m^{T}\right) + (1-\theta) * p\left(r^{i} \middle| m^{N}\right) \\ log \prod_{i}^{I} P\left(r^{i} \middle| \theta, m\right) = \sum_{i}^{I} \log\left(P\left(r^{i} \middle| \theta, m\right)\right. \end{cases}$$

Equation (3)

where $marker^k$ is a 1×k vector ($k = c(marker^1, marker^2, \ldots, marker^k)$) representing k methylation differential regions, j ($j = c(CpG^1, CpG^2, \ldots, CpG^j)$) indicates the methylation status of the CpG site on each read (0 or 1 means unmethylated or methylated), i ($i \in c(read^1, read^2, \ldots, read^i, read^N)$) represents one of the reads from all I reads, and θ indicates the score that this read originates from the tumor.

Construction and validation of the CNV model

For the CNV model, the scores of CNV classifiers were obtained by using support vector machine (SVM) analysis.[21] Then, thresholds for the copy number model were determined by the same strategy as the methylation model.

Construction and evaluation of the classifier, termed UCseek

For UCseek, we integrated the scores of the methylation model and the CNV model according to Equation (4). The final UCseek scores were obtained based on the weight w of the two models. The optimal w is 0.45, determined by maximum likelihood estimation in this study.[22] Unlike GUseek, which uses markers screened for methylation and copy number to normalize them into a model for remodeling, our UCseek is constructed from two different types of information separately to form an integrated system through an appropriate weight for tumor diagnosis and monitoring. Subsequently, the performance of the methylation model, CNV model, and UCseek was evaluated in the test set and independent validation set, receiver operating characteristic (ROC) curves were plotted, and the area under the ROC curve (AUC) values were calculated.[23]

$$UCseek = w * cnv\ score + (1-w) * Methylation\ score$$

Equation (4)

Validation of the selected methylation markers and copy number markers in tumor and normal tissues

First, the methylation model obtained from urine samples was used to classify 54 tissues, and the performance of the methylation model on tissues was evaluated. Subsequently, for the verification of the copy number marker, we used the SNP6.0 microarray data in TCGA. We assigned the CNV segment value to the candidate CNV region as the copy number information of these CNV regions to determine whether candidate regions are also differentiated among tissue samples.

Statistical analysis

The statistical analysis and data visualization used in this study were performed by using R packages, including but not limited to the following R packages: "limma", "e1071", "ROCR", "tidyverse", "ggplot", "ggpubr", "survminer", "devtools", "dplyr", "survival", "ComplexHeatmap", "pheatmap", and "caret". All hypothesis tests were two-sided, and $P < 0.05$ was considered statistically significant.

Ethics

The study was approved by the Ethics Committee of Peking University First Hospital (No. 2015(977)), the Affiliated Hospital of Qingdao University (No. QYFY-KYLL944311920) and Cancer Hospital Chinese Academy of Medical Sciences (No. NCC2017-YZ-013).

Data availability

The original sequencing data of the discovery set, independent validation set, and urothelial carcinoma tissue data in this study have been deposited in the Genome Sequence Archive in the National Genomics Data Center, Beijing Institute of Genomics (China National Center for Bioinformation) of CAS[24] under accession Nos. PRJCA001203, HRA002284 and HRA001562.

Role of funding source
The funders had no role in the study design, data collection, data analysis, interpretation, or writing of the report.

Results

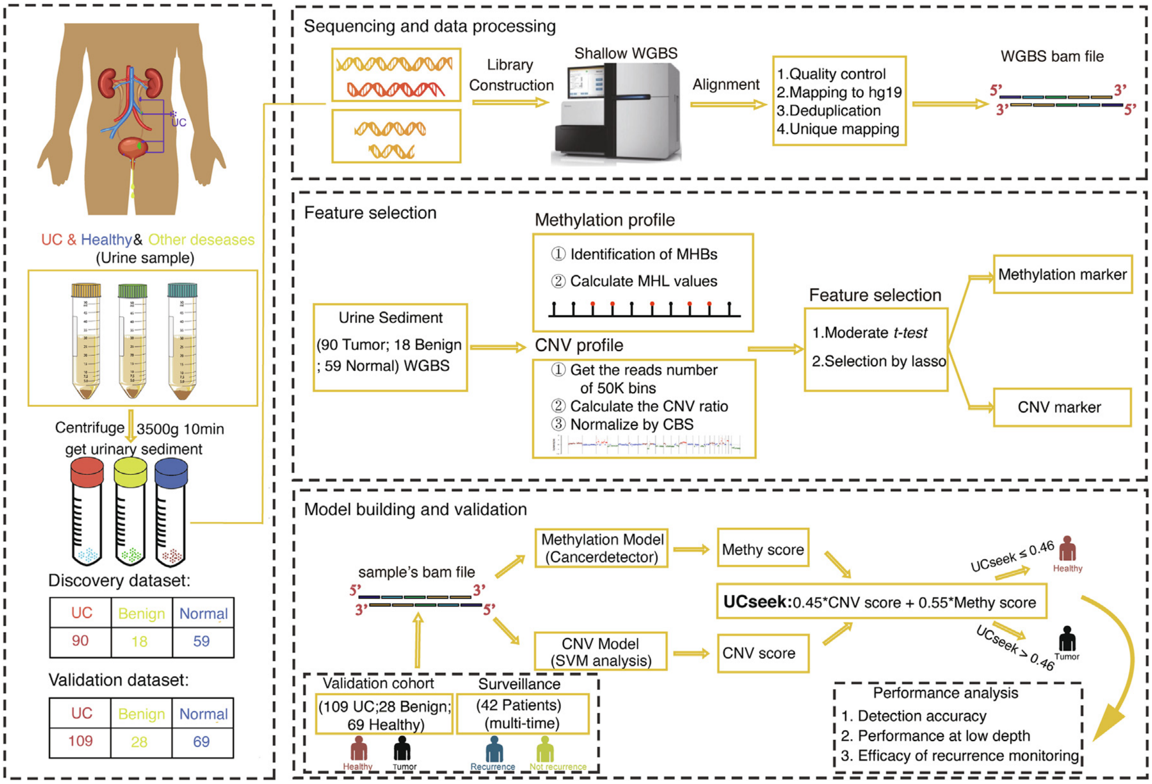
Data generation and analysis

In this study, the discovery cohort included 90 UC patients (36 low-grade and 54 high-grade tumors) and 77 control individuals without tumors (Table S1). The independent validation cohort included 109 UC patients (13 low-grade and 96 high-grade tumors) and 97 control individuals without tumors. The study included the demographic and clinical characteristics of the participants in detail (Table S1). Ultimately, based on these data, we developed a model, UCseek, that combined methylation information and copy number information to diagnose and monitor UC patients (Fig. 1).

The methylation model has excellent diagnostic sensitivity for UCs

First, we evaluated the tumor-specific methylation pattern using MHL scores for signal amplification in detecting tumor signals in plasma DNA even at shallow depths.[16,25] We identified 60 significantly differentially methylated MHBs. Among them, differences in 27 MHBs were also found between tumor and normal tissues in the WGBS data (Fig. 2a). Hypergeometric testing revealed that the differentially methylated MHBs derived from tissues were significantly enriched in those of the urine sediments ($p = 2.62 \times 10^{-12}$). These findings suggest that WGBS of urine sediment DNA can detect tumor-derived DNA methylation events.

Next, we constructed the methylation model using a probabilistic approach-based method called Cancer-Detector[25] and defined the methylation score to predict the source of the reads derived from the 60 differentially methylated MHBs of urine sediments between UC patients and nontumor controls in the discovery cohort. The methylation score was significantly higher in patients with UCs but lower in patients without tumors (Fig. 2b). Then, to classify UC patients from nontumor controls, we set an optimal cutoff value for the methylation score greater than 0.55 according to the Youden index (Fig. 2c and Fig. S1a). The AUC value of the methylation model was observed in both the training and test datasets of the discovery cohort (Fig. 2d). Additionally, it performed equally well for predicting UCs from tissue WGBS data (Fig. S1b and Fig. S1c),



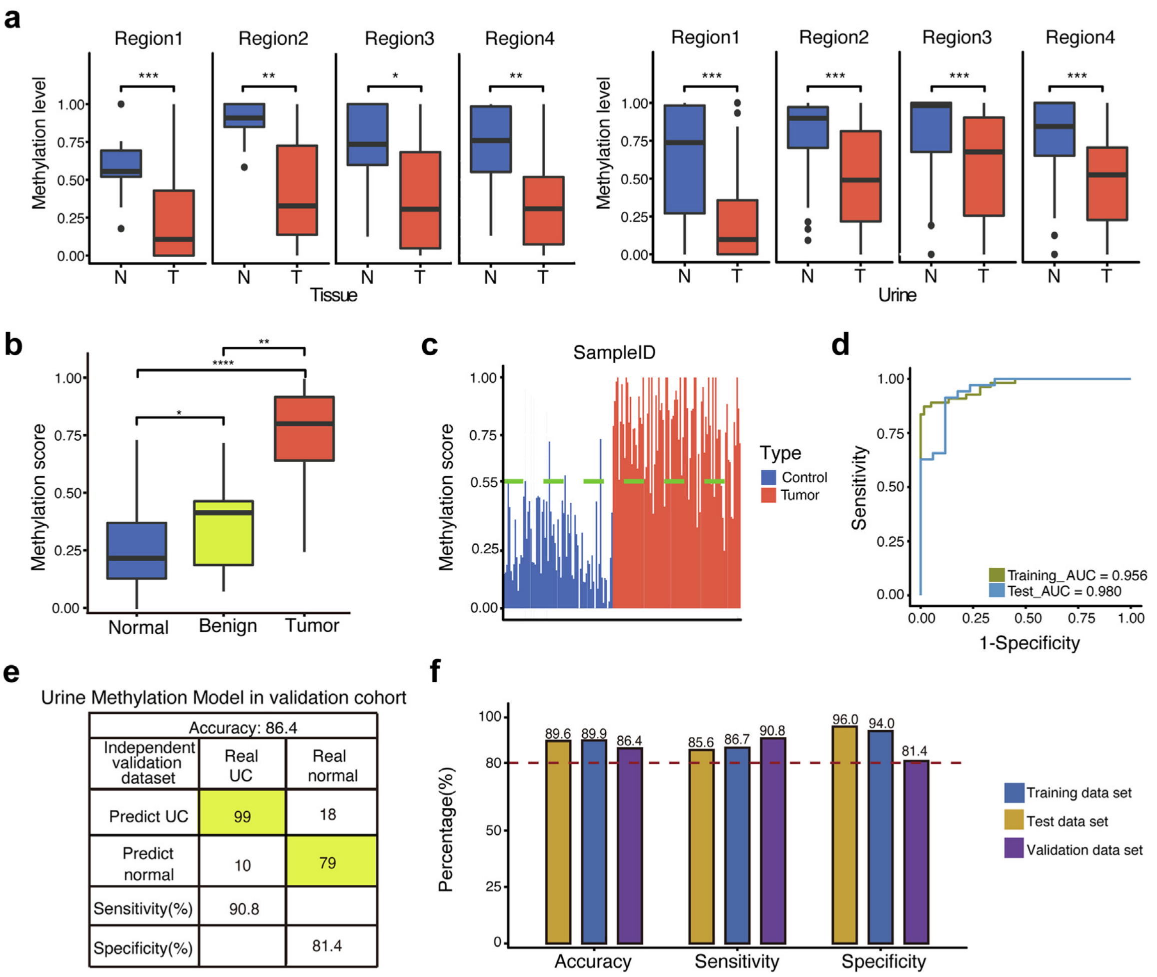
Discovery dataset:

UC	Benign	Normal
90	18	59

Validation dataset:

UC	Benign	Normal
109	28	69

Fig. 1: **Workflow of data generation and UCseek classifier development for the noninvasive detection and monitoring of UCs**. Evaluation of methylation and copy number variation (CNV) profiles in urinary sediments using shallow whole-genome bisulfite sequencing (sWGBS) markers screened based on the differences between cancer patients and noncancer people (benign disease patients and healthy individuals). Construction and integration of methylation and CNV models to develop an enhanced classifier termed 'UCseek'.



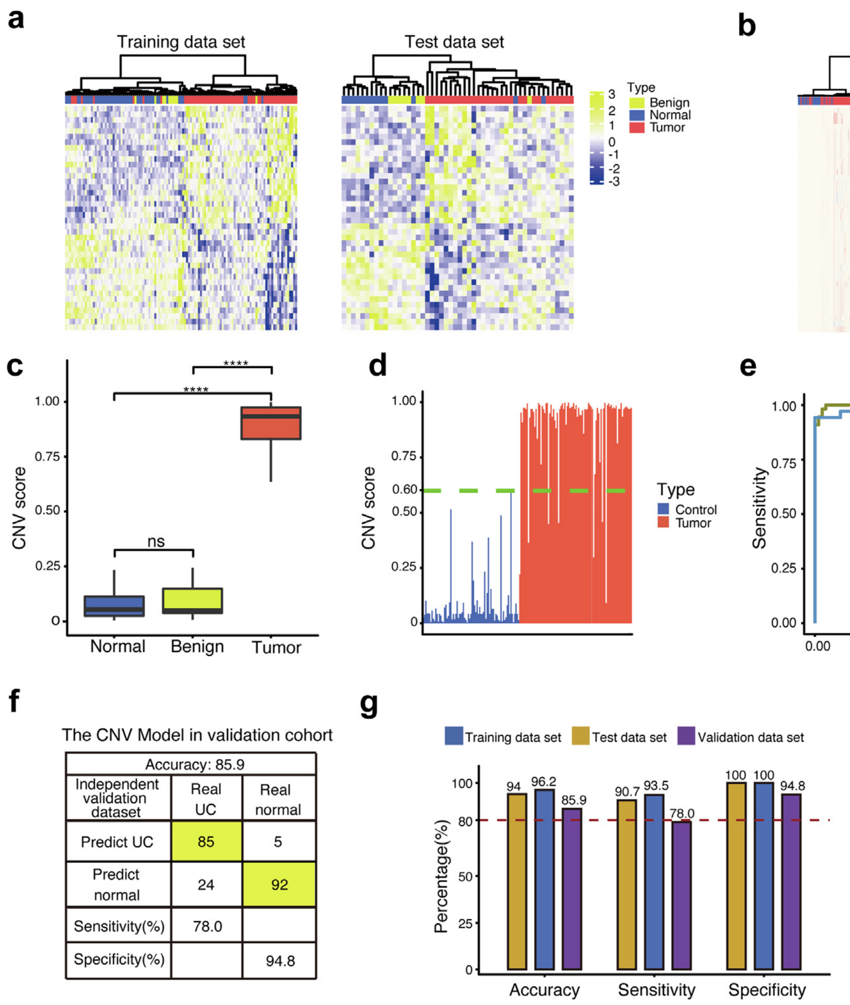
Accuracy: 86.4		
Independent validation dataset	Real UC	Real normal
Predict UC	99	18
Predict normal	10	79
Sensitivity(%)	90.8	
Specificity(%)		81.4

***Fig. 2:* The diagnostic performance of the methylation model showed high sensitivity.** (a) Box plot showing the distribution of methylation scores in primary tissue (left panel) (n_T = 45, n_N = 9) and urine (right panel) (n_T = 90, n_N = 67) of methylation markers screened from urine sediment WGBS data between the normal and tumor groups (where N represents the control group and T represents the tumor group). (b) Box plot showing the significant differences in scores in the methylation model found among healthy individuals (n_{normal} = 59), benign patients (n_{benign} = 18), and tumor patients (n_{tumor} = 90) in the discovery set. (c) Manhattan plot showing the significantly different distribution of the methylation scores of the methylation model between control and tumor samples in the discovery set (n = 167). The green line represents the threshold of the methylation model. The control group contained healthy individuals and benign patients. (d) ROC curves and the corresponding AUC values of the methylation model in the training set and test set. (e) Confusion matrix of the methylation model in the independent validation set. (f) Bar plot showing the sensitivity, specificity, and accuracy of the methylation model in the training set, test set, and independent validation set (n = 206); the red line indicates 80% accuracy. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001 by Wilcoxon tests.

which further suggested that WGBS of urine sediment DNA can detect tumor-derived DNA methylation events. Importantly, the methylation model achieved a sensitivity of 90.8% and a specificity of 81.4% in the independent validation cohort (Fig. 2e). Therefore, the methylation model showed moderate specificity (Fig. 2f) but excellent sensitivity.

The CNV model has excellent specificity in the detection of UCs

Considering that CNVs are relatively rare and less extensive in normal-appearing urothelium than in tumors, to further improve the diagnostic potential of urine sediment DNA with WGBS data, we focused on CNVs that can be accurately evaluated with WGBS data.[15] Similarly, we identified 40 regions with the most significant differences in copy number patterns between UC and nontumor urine samples. Next, we performed unsupervised clustering with selected CNV markers. As expected, these markers could distinguish UC samples from normal samples in both training and test data (Fig. 3a), as well as the tumor tissue from normal tissue in the TCGA bladder cancer dataset (Fig. 3b), suggesting that the CNV markers



The CNV Model in validation cohort

Accuracy: 85.9

Independent validation dataset	Real UC	Real normal
Predict UC	85	5
Predict normal	24	92
Sensitivity(%)	78.0	
Specificity(%)		94.8

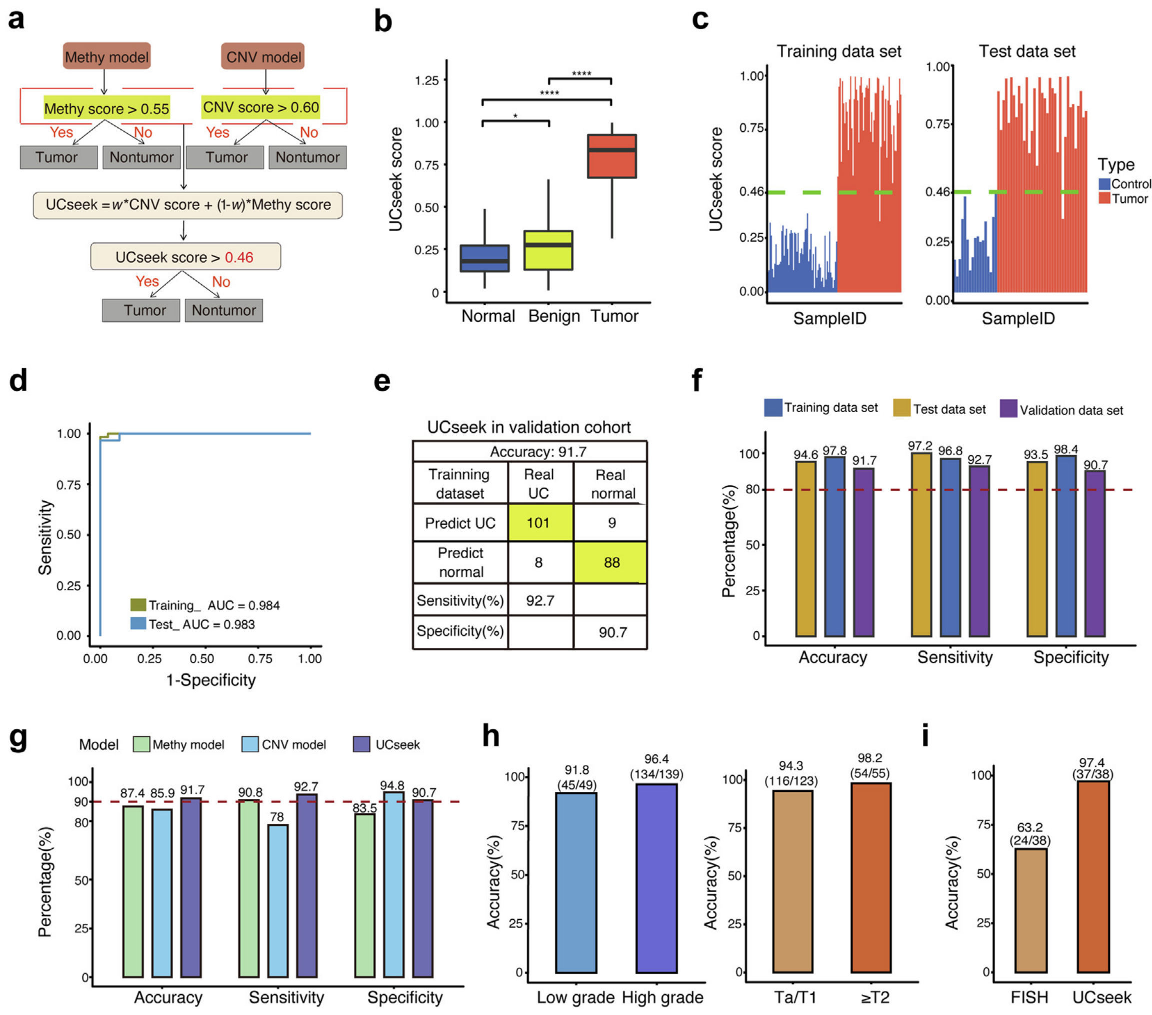
Fig. 3: **The diagnostic performance of the CNV model showed high specificity.** (a) Unsupervised clustering results of the copy number markers in the training set (left panel) and test set (right panel) screened by urine sediment WGBS data (*n* = 167). (b) Unsupervised clustering of copy number markers in the TCGA validation dataset (*n* = 448), showing the potential classification efficacy of these screened markers on UC tissues. (c) Box plot showing the significant differences in scores of the CNV model among healthy individuals, benign patients, and tumor patients in the discovery set (*n* = 167). (d) Manhattan plot showing the significantly different distribution of the CNV scores of the CNV model between control and tumor samples in the discovery set (*n* = 167). The green line represents the threshold of the CNV model. The control group contained healthy individuals and benign patients. (e) ROC curves and the corresponding AUC values of the CNV model in the training set and test set. (f) Confusion matrix of the CNV model in the independent validation set. (g) Bar plot showing the sensitivity, specificity, and accuracy values of the CNV model in the training set, test set, and independent validation set (*n* = 206). The red line represents 80% accuracy. **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001, ns indicates not significant by Wilcoxon tests.

identified in urine sediment DNA are specific markers for UCs.

Next, we constructed the CNV model based on a support vector machine (SVM) machine learning algorithm to perform the classification task with the 40 selected CNV markers of urine sediment. Similarly, the CNV score of urine sediment DNA was significantly higher in UC patients than in nontumor controls (Fig. 3c). An optimal cutoff value greater than 0.60 was set according to the Youden index (Fig. 3d). Notably, high AUC values were observed in both the training and test datasets of the discovery cohort (Fig. 3e). The CNV model achieved a sensitivity of 78% and a specificity of 94.8% in the independent validation cohort (Fig. 3f). As expected, the CNV model showed excellent diagnostic specificity in both the discovery cohort and validation cohort (Fig. 3g). Collectively, this CNV model has excellent specificity but moderate sensitivity in detecting UCs.

Construction of an ensemble model, UCseek, by combining the DNA methylation model and CNV model

Considering the high sensitivity of the DNA methylation model and the high specificity of the CNV model described above, we integrated them to obtain a new diagnostic model, referred to as UCseek (Fig. 4a). The UCseek scores were significantly different between tumor patients and nontumor controls (Fig. 4b). Similarly, we set an optimal cutoff value for the UCseek score greater than 0.46 by the Youden index (Fig. 4c). UCseek exhibited excellent diagnostic ability in the discovery cohort (Fig. 4d) and in the independent validation cohort with high sensitivity (92.7%) and specificity (90.7%) (Fig. 4e and f). The ensemble model UCseek outperformed the methylation model and the CNV model (Fig. 4g). Moreover, UCseek performed equally well for both high-grade or late-stage tumors and low-grade or early-stage tumors (Fig. 4h). Notably, we performed FISH tests and urine cytology tests in 38 cases



UCseek in validation cohort

Accuracy: 91.7		
Trainning dataset	Real UC	Real normal
Predict UC	101	9
Predict normal	8	88
Sensitivity(%)	92.7	
Specificity(%)		90.7

Fig. 4: **UCseek has better performance than the methylation model or the CNV model.** (a) The UCseek diagnostic model obtained by integrating the methylation model and the CNV model. The optimal threshold of UCseek is 0.46 according to the Youden index. (b) Box plot showing the significant differences in UCseek scores among healthy individuals, benign patients, and tumor patients in the discovery set (n = 167). (c) Manhattan plot showing the scores of UCseek between the control and tumor groups in the training dataset and test dataset. The green line represents the UCseek threshold. The control group contained healthy individuals and benign patients. (d) ROC curves and the corresponding AUC values of UCseek in the training set and test set. (e) Confusion matrix of UCseek in the independent validation set. (f) Bar plot showing the sensitivity, specificity and accuracy of UCseek in the training set, test set, and independent validation set (n = 206); the red line indicates an accuracy of 80%. (g) Bar plot showing the accuracy, sensitivity, and specificity of the three models in the independent validation set (n = 206), where the red line represents the historical 90% accuracy. (h) Bar plot showing the diagnostic performance of UCseek in low- (n = 49) and high-grade (n = 139) (left) and early- (n = 123) and late-stage patients (n = 55) (right). (i). Histogram of the accuracy of UCseek and FISH detection in urine samples (n = 38). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$ by Wilcoxon tests.

and 34 cases (Table S2) of the discovery and validation cohorts, and the results showed that UCseek significantly outperformed the FISH test (Fig. 4i) and urine cytology (Fig. S2a).

Moreover, our previous study showed that UCs can be detected based on the CNV profiles of urinary cfDNA.[12] Thus, we hypothesized that the UCseek model should also perform well with whole urine DNA (cellular and/or cell-free DNA), which can be purified easily and reliably. In fact, ten of the samples with FISH results were whole urine. We compared the performance of UCseek based on whole urine DNA and FISH in these UC patients. Notably, the accuracy of the UCseek model predictions exceeded that of FISH (90% vs. 70%) (Fig. S2b). Moreover, we found that the UCseek scores were significantly correlated with advanced grade (Fig. S2c), but there was no obvious difference in sex, age, or stage (Fig. S2d–f). Collectively, the UCseek model based on urinary cellular and/or cell-free DNA not only facilitated the detection of early-stage UCs but also offered practical advantages for urine collection, such as more frequent use and home use.

The subsampling results showed that UCseek is an ultrasensitive and robust UC detection method

To evaluate the performance of UCseek at low sequencing depths, we randomly subsampled raw data with ratios of 10%, 30%, and 50% for the samples from the independent validation cohort. We found that the Pearson correlation values of the scores UCseek, the methylation model and the CNV model were very high between the original 3X-5X depths and the subsampled depths (Fig. 5a and Figs. S3a and S3b). UCseek maintained high accuracy at a depth down to 0.9X-1.5X, whereas it decreased moderately at a depth of 0.3X-0.5X (Fig. 5b). In particular, we found that the specificity but not the sensitivity of UCseek decreased significantly at the 0.3X-0.5X depth (Fig. 5c). The results suggested that the CNV model may perform suboptimally at ultralow depths. Although the specificity of the CNV model decreased significantly at a depth of 0.3X-0.5X (Figs S3c and S3d), the methylation model maintained a relatively good performance (Figs. S3c and S3d). Furthermore, UCseek can sensitively detect low-stage and early-stage tumors even at low sequencing coverage (Fig. 5d and e). Therefore, UCseek holds great potential to detect cancer early and cost effectively.

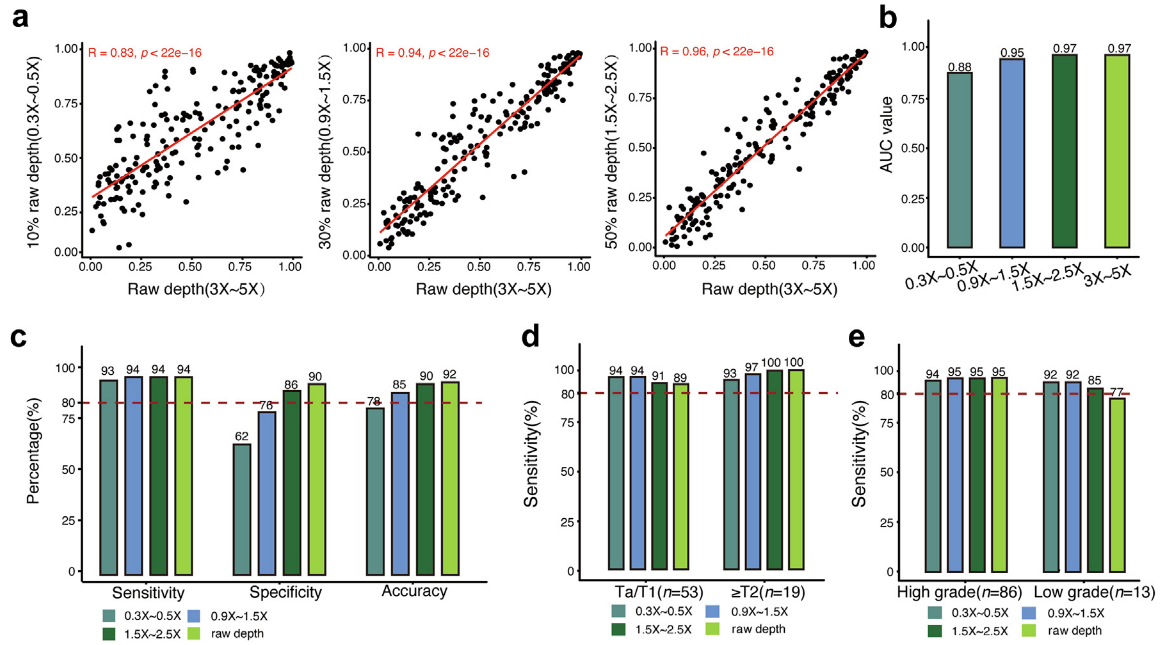


Fig. 5: **UCseek showed high concordance and classification performance with the original results at low-depth sequencing.** (a) Scatter plots illustrating the Pearson's correlation coefficient of the scores of UCseek between different sequencing depths and the original depths. Depth simulation was performed for 206 samples in the independent validation set, with 3 different depths: 0.3X~0.5X, 0.9X~1.5X, and 1.5X~2.5X. The data at these three different depths are consistent with the prediction results of the original depth data in UCseek, where the X-axis represents the original depth sequencing data and the Y-axis represents the different depth sequencing data. The concordance was calculated by the Pearson correlation coefficient. (b) AUC values of UCseek for the independent validation set at different depths (n = 206). (c) Bar plots showing the performance of UCseek in the independent validation set at different simulation depths, including sensitivity, specificity and accuracy. (d) The accuracy of UCseek for different stages at different simulation depths in the independent validation set. (e) The accuracy of UCseek for high and low grades at different simulation depths in the independent validation set (n = 206).

UCseek has better capability than cystoscopy for monitoring recurrence and small lesions

Given that UCseek performed well at the early stage (Fig. 4h), we further collected a surveillance cohort ($n = 42$) including the first-diagnosis cohort ($n = 8$) and the postintent-to-treat cohort ($n = 34$), with urine samples and cystoscopy and/or MR imaging follow-up data (Table S3). Indeed, UCseek gave a positive prediction for all 8 patients, including Ta high-grade ($n = 2$), T1 high-grade ($n = 3$), T3a high-grade ($n = 2$) and T3b high-grade ($n = 1$). In addition, 4 out of the 8 patients were subjected to repeated transurethral resection of bladder tumor (Re-TURBT), and UCseek gave a consistent prediction with Re-TURBT in 3 patients (Fig. 6a). These results further showed the ultrasensitivity of UCseek in detecting small tumors (Fig. 6a).

Next, in the postcurative-intent therapy patients ($n = 21$), UCseek gave a consistent prediction with Re-TURBT in 20 of 21 patients. The patient with one exception, 391–11, was predicted to be negative by UCseek but positive by Re-TURBT. Notably, the patient was identified as having a recurrent tumor 22 months later by Re-TURBT (Fig. 6b). More importantly, in the post curative-intent therapy patients ($n = 13$), UCseek

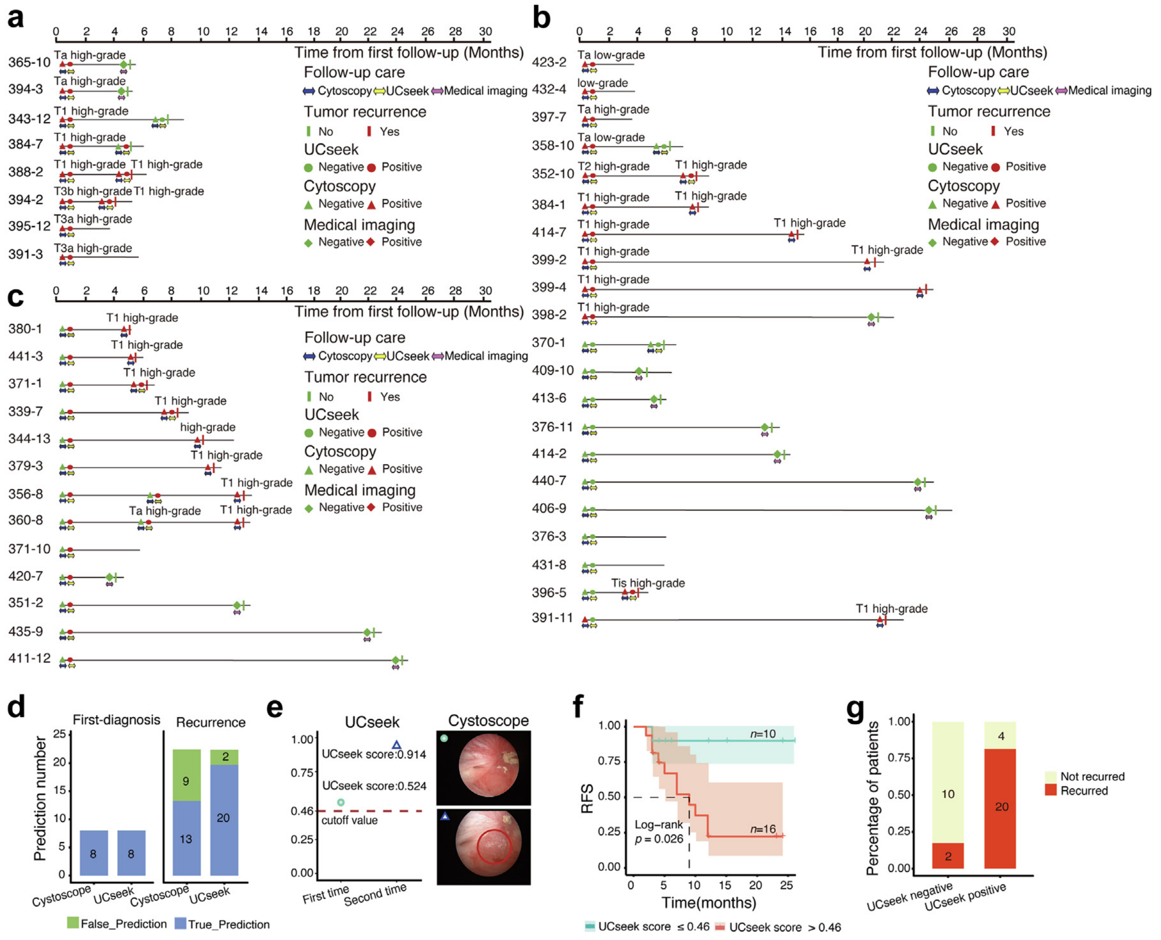


Fig. 6: **UCseek assay results and timing of treatment for each patient in the surveillance cohort.** (a) Patients with newly diagnosed UC. The UCseek-predicted results, cystoscopy results, and imaging results were marked, and the specimens were obtained on the same day before surgery ($n = 8$). (b) Patients with consistent cystoscopy and UCseek results ($n = 21$). The pathological information is shown in Figure a, and the specimens were obtained before surgery on the same day. (c) Patients with inconsistent cystoscopy and UCseek results ($n = 13$). The pathological information is shown in Figure A, and the specimens were obtained before surgery on the same day. (d) Histogram showing the accuracy of cystoscopy and UCseek for newly diagnosed patients (left) and for relapsed patients (right). In this recurrence cohort, eight patients were newly diagnosed at the first diagnosis, and 22 patients experienced recurrence. (e) Example of a patient with minimal tumor detected by UCseek. The cystoscopic images of one patient at two follow-up visits and the corresponding UCseek scores. This patient was missed by cystoscopy at the first follow-up visit, yet UCseek detected a tumor signal, which was subsequently confirmed by the second cystoscopy. (f) Recurrence-free survival (RFS) according to the UCseek scores in the recurrence cohort ($n = 26$). Statistical significance was evaluated by log-rank tests. (g) The bar plot shows the sensitivity of UCseek diagnosis.

gave a positive prediction but was negative by Re-TURBT, and 8 out of 13 patients were confirmed to have recurrence by Re-TURBT within 1 year of follow-up (Fig. 6c). Overall, we found that UCseek accurately detected 20 of 22 (90.91%) patients with recurrence, but cystoscopy detected only 13 of 22 (59.09%) recurrent patients (Fig. 6d), suggesting that UCseek could serve as a noninvasive and highly sensitive approach to predict the recurrence of bladder cancer by monitoring small lesions. The good performance of this model is highlighted by a patient who was detected by UCseek but was missed by ordinary cystoscopy (Fig. 6e). In this case, the lesion was not observed by cystoscopy but was later diagnosed as bladder cancer after half a year with a higher UCseek score (Fig. 6e). Consistently, Kaplan–Meier survival curves for recurrence-free survival (RFS) further showed that patients with detectable tumor DNA by UCseek recurred at a median of 6 months after definitive therapy (Fig. 6f). Of the 24 patients with tumor DNA detectable by UCseek, 20 recurred, while of the 12 patients without tumor DNA detectable by UCseek, only 2 recurred (Fig. 6g). Therefore, UCseek may serve as an effective method for the detection of early-stage, minimal, residual, and recurrent tumors, which in turn may improve disease management.

Discussion

The success of early cancer detection largely relies on (i) high-quality cancer-specific markers and (ii) computational methods for the ultrasensitive detection of tiny amounts of tumor ctDNA. In this study, we developed an integrated machine learning-based approach called UCseek to incorporate both CNV and DNA methylation markers for the ultrasensitive early detection and recurrence monitoring of UCs. In recent years, significant research effort has been focused on the development of urinary biomarkers for the noninvasive and cost-effective detection of variations in protein expression or chromosomal or DNA methylation instability with low coverage whole genome sequencing or bisulfite sequencing in either cfDNA or genomic DNA for UCs. For example, urine protein IL-1, IL-1ra, and IL-8 were found to be able to distinguish control urine from various bladder cancer stages with specificity values exceeding 0.9 in an independent cohort of 80 urine samples upon ELISA validation.[26] For urine tumor DNA, an assay named UroCAD identified UCs by detecting CNVs from urine-exfoliated cells by whole genome sequencing with an overall sensitivity of 80.4% and specificity of 94.9% in the external validation cohort of 95 participants but showed slightly low sensitivity (63.6%) for low-grade tumors.[27] A 2-marker-based methylation assay using mass spectrometry, termed utMeMA for the detection of urinary tumor DNA, had a sensitivity of 91.7% and a specificity of 77.3% in the external validation cohort of 175 participants and showed a sensitivity of 69.2% for low-grade tumors.[28] In addition, a GHSR/MAL panel with 9 methylation markers achieved good performance in urine pellets with a sensitivity of 78.6% and a specificity of 91.7%, but without further validation.[29] In addition to these studies, our research group previously developed two new models, UCdetector (sensitivity 78.6%, specificity 87.5% in the external validation cohort of 52 participants) and GUseek (without an independent validation cohort), for UCs or genitourinary cancer by using urinary cell-free DNA and cellular DNA based on CNVs and DNA methylation markers, which need further validation.[12,30] Here, the key considerations of (i) and (ii) promised, as demonstrated, that our method UCseek could perform well compared to the other methods. As expected, our UCseek exhibited an excellent prediction sensitivity of 92.7% in the external validation cohort of 206 participants, which outperformed the previous UC detection methods, i.e., UroCAD, UCdetector and GHSR/MAL panel. Moreover, UCseek outperformed utMeMA (90.7% vs. 83.1%) and UCdetector (90.7% vs. 87.5%) in terms of prediction specificity in the independent validation cohort (n = 206). To further validate the performance of UCseek, a large-scale, multicenter, and prospective clinical trial (ChiCTR2200063932) for low- and high-grade UCs in China is ongoing. Collectively, UCseek showed remarkable sensitivity and specificity for the prediction of UCs, which has great application potential in the clinic.

In contrast, it has been shown that low-grade UCs accumulate fewer CNVs than high-grade UCs[13,31]; thus, patients with limited CNVs cannot be identified by methods such as UroCAD. Moreover, the moderate decrease in detection specificity for DNA methylation markers may be due to some cancer-specific comethylated markers also being tightly coupled with methylated CpGs in urothelial cells. Consistently, it has been shown that the loss of methylation linkage disequilibrium in cancer cells was validated, but the majority of MHBs in cancers still contain tightly coupled CpGs compared with matched normal tissues.[16]

Moreover, compared to our previous diagnostic model GUseek, UCseek achieved better results in UC detection, especially at low sequencing depths (0.3X-0.5X), which indicates that it can become a low-cost cancer detection method. A unified threshold for samples with various sequencing depths can be assigned for cancer diagnosis. More importantly, the performance of UCseek was further validated in an independent validation cohort (n = 206), including a surveillance cohort (n = 42). These advantages were not proven by GUseek before, so UCseek is more likely to be applied in clinical practice.

We acknowledge several limitations of our study. First, this study only primarily analyzed a single landmark timepoint and did not systematically collect serial longitudinal urine draws for all patients in the surveillance cohort. Second, only a limited number of benign diseases and low-grade patients in an independent validation set were included in the current study. Furthermore, a multicenter, blinded clinical trial including more benign diseases is needed to validate the clinical application of UCseek to identify UC from patients with urinary system diseases. Additionally, further studies are needed to validate and expand its clinical utility in detecting minimal residual disease and predicting recurrence.

Taken together, UCseek could be a highly specific, robust, and noninvasive urothelial carcinoma diagnostic method with improved accuracy compared with the DNA methylation model or the CNV model alone. It may be used as a noninvasive approach for diagnosis and recurrence surveillance in UCs prior to the use of cystoscopy, which would largely reduce the burden on patients. We optimized an improved approach named UCseek for the noninvasive diagnosis and monitoring of UCs in both low- and high-grade tumors and in early- and advanced-stage tumors, even at ultralow sequencing depths, which may reduce the burden of cystoscopy and blind second surgery. Overall, our findings help to make a major step toward the clinicalization of methods for diagnosing UC and provide more precise and personalized treatment recommendations for patients who require long-term monitoring for recurrence.

Contributors

WC, XB, LZ and XL conceived the project. PW, YS and DZ performed the experiments. PW, YL and JZ verified the underlying data, had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. PW, YS, and DZ conducted the bioinformatics analyses. JZ, JS, MZ, YL, ZY, YT, QZ, JQ, and MZ provided technical or material support. WC, PW, YS and JZ wrote the manuscript with help from all of the authors. All authors read and approved the final manuscript.

Data sharing statement

The original sequencing data of the discovery set, independent validation set, and urothelial carcinoma tissue data in this study have been deposited in the Genome Sequence Archive in the National Genomics Data Center, Beijing Institute of Genomics (China National Center for Bioinformation) of CAS under accession Nos. PRJCA001203, HRA002284 and HRA001562 at https://ngdc.cncb.ac.cn/gsa/. For the TCGA-BLCA cohort, copy number segment data were obtained from UCSC Xena at https://xenabrowser.net/datapages/.

Declaration of interests

The authors declare no potential conflicts of interest.

Acknowledgments

We thank the UC cancer patients and their families for their participation and for providing their UC specimens for the advancement of cancer research. We also thank the Peking University First Hospital, the Affiliated Hospital of Qingdao University and Cancer Hospital Chinese Academy of Medical Sciences for preparing the specimens for this study. This work was supported by the National Key R&D Program of China (2018YFC2000100, 2019YFA0110900 to W.C.), the CAS Strategic Priority Research Program (XDA16010102 to W.C.), the National Natural Science Foundation of China (82173061 to W.C., 82103426 to Y.L.), CAMS Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-C&T-B-052, 2021-I2M-1-033), Beijing Hope Run Special Fund of Cancer Foundation of China (LC2017A15).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2023.104437.

References

1 Babjuk M, Burger M, Comperat EM, et al. European association of urology guidelines on non-muscle-invasive bladder cancer (TaT1 and carcinoma in situ) - 2019 update. *Eur Urol*. 2019;76:639–657.

2 Lotan Y, Svatek RS, Sagalowsky AI. Should we screen for bladder cancer in a high-risk population? *Cancer*. 2006;107:982–990.

3 Dimashkieh H, Wolff DJ, Smith TM, Houser PM, Nietert PJ, Yang J. Evaluation of urovysion and cytology for bladder cancer detection: a study of 1835 paired urine samples with clinical and histologic correlation. *Cancer Cytopathol*. 2013;121:591–597.

4 Sweis RF, Galsky MD. Emerging role of immunotherapy in urothelial carcinoma-Immunobiology/biomarkers. *Urol Oncol*. 2016;34:556–565.

5 Messer J, Shariat SF, Brien JC, et al. Urinary cytology has a poor performance for predicting invasive or high-grade upper-tract urothelial carcinoma. *BJU Int*. 2011;108:701–705.

6 Tanaka N, Kikuchi E, Kanao K, et al. The predictive value of positive urine cytology for outcomes following radical nephroureterectomy in patients with primary upper tract urothelial carcinoma: a multi-institutional study. *Urol Oncol*. 2014;32:48.e19–48.e26.

7 Chou R, Gore JL, Buckley D, et al. Urinary biomarkers for diagnosis of bladder cancer: a systematic review and meta-analysis. *Ann Intern Med*. 2015;163:922–931.

8 Lin T, Liu Z, Liu L, et al. Prospective evaluation of fluorescence in situ hybridization for diagnosing urothelial carcinoma. *Oncol Lett*. 2017;13:3928–3934.

9 Mancini M, Zazzara M, Zattoni F. Stem cells, biomarkers and genetic profiling: approaching future challenges in urology. *Urologia*. 2016;83:4–13.

10 Springer SU, Chen CH, Rodriguez Pena MDC, et al. Non-invasive detection of urothelial cancer through the analysis of driver gene mutations and aneuploidy. *Elife*. 2018;7:e32143.

11 Mancini M, Righetto M, Zumerle S, Montopoli M, Zattoni F. The bladder EpiCheck test as a non-invasive tool based on the identification of DNA methylation in bladder cancer cells in the urine: a review of published evidence. *Int J Mol Sci*. 2020;21:6542.

12 Ge G, Peng D, Guan B, et al. Urothelial carcinoma detection based on copy number profiles of urinary cell-free DNA by shallow whole-genome sequencing. *Clin Chem*. 2020;66:188–198.

13 Guan B, Liang Y, Lu H, et al. Copy number signatures and clinical outcomes in upper tract urothelial carcinoma. *Front Cell Dev Biol*. 2021;9:713499.

14 Lu H, Liang Y, Guan B, et al. Aristolochic acid mutational signature defines the low-risk subtype in upper tract urothelial carcinoma. *Theranostics*. 2020;10:4323–4333.

15 Xu Z, Ge G, Guan B, et al. Noninvasive detection and localization of genitourinary cancers using urinary sediment DNA methylomes and copy number profiles. *Eur Urol*. 2020;77:288–290.

16 Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet*. 2017;49:635–642.

17 Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011;472:90–94.

18 Baslan T, Kendall J, Rodgers L, et al. Genome-wide copy number analysis of single cells. *Nat Protoc*. 2012;7:1024–1041.

19 Ulz P, Belic J, Graf R, et al. Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. *Nat Commun*. 2016;7:12008.

20 Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J*. 2005;47:458–472.

21 Cao LJ, Tay FH. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans Neural Netw*. 2003;14:1506–1518.

22 Lee G, Lee MJ. Regression discontinuity for binary response and local maximum likelihood estimator to extrapolate treatment. *Eval Rev*. 2022. https://doi.org/10.1177/0193841X221105968.

23 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39:561–577.

24 Database resources of the BIG data center in. *Nucleic Acids Res*. 2018;46:D14–D20.

25 Li W, Li Q, Kang S, et al. CancerDetector: ultrasensitive and noninvasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res*. 2018;46:e89.

26 Vanarsa K, Enan S, Patel P, et al. Urine protein biomarkers of bladder cancer arising from 16-plex antibody-based screens. *Oncotarget*. 2021;12:783–790.

27 Zeng S, Ying Y, Xing N, et al. Noninvasive detection of urothelial carcinoma by cost-effective low-coverage whole-genome sequencing from urine-exfoliated cell DNA. *Clin Cancer Res*. 2020; 26:5646–5654.

28 Chen X, Zhang J, Ruan W, et al. Urine DNA methylation assay enables early detection and recurrence monitoring for bladder cancer. *J Clin Invest*. 2020;130:6278–6289.

29 Hentschel AE, Nieuwenhuijzen JA, Bosschieter J, et al. Comparative analysis of urine fractions for optimal bladder cancer detection using DNA methylation markers. *Cancers (Basel)*. 2020;12:859.

30 Wolff EM, Chihara Y, Pan F, et al. Unique DNA methylation patterns distinguish noninvasive and invasive urothelial cancers and establish an epigenetic field defect in premalignant tissue. *Cancer Res*. 2010;70:8169–8178.

31 Hurst CD, Alder O, Platt FM, et al. Genomic subtypes of non-invasive bladder cancer with distinct metabolic profile and female gender bias in KDM6A mutation frequency. *Cancer Cell*. 2017;32:701–715.