Identification of Methylation Biomarkers for Early Detection of Prostate Cancer

Yiyi Pu
Hangzhou Medical College
Chao Li
Hangzhou Medical College
Haining Yuan
Hangzhou Medical College
Xiaoju Wang (wangxj@zjams.com.cn)
Hangzhou Medical College

Research Article

Keywords: DNA methylation, early diagnosis, biomarker, prostate cancer, bioinformatics

DOI: https://doi.org/10.21203/rs.3.rs-272705/v1

Abstract

Background: DNA methylation has been widely used for development of cancer diagnosis biomarker. However, the clinical translational rate is low. Databases, such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO), offer great opportunities for DNA methylation biomarker identification. By taking advantage of the public databases, we aimed to identify cancer specific biomarkers based on DNA methylation level for early detection purpose.

Results: We performed a pan-cancer methylation analysis using datasets from TCGA and validated the results using GEO datasets. To identify early-diagnosis biomarkers, we focused on the localized tumors, and identified the biomarkers that can effectively distinguish the localized tumors from normal tissues. After comparing biomarkers for all cancer types, we identified a large group of cancer specific biomarkers. Within all 26 prostate cancer specific biomarkers selected, we confirmed three biomarker sets by multiplex analysis. First, 7 biomarkers (cg26140475, cg24891312, cg24522654, cg21359747, cg03254336, cg12697139, cg19034132) could detect localized prostate tumors from normal tissues (AUC > 0.9). Second, 9 biomarkers (cg17220055, cg26140475, cg24891312, cg09853702, cg22400059, cg16736279, cg27639613, cg06011086, cg00664697) could distinguish between low and high Gleason score prostate tumors (AUC = 0.79). Last, a single biomarker (cg26140475) completely separated prostate tumor from other urinary tumors (AUC = 1).

Conclusions: Our study identified and validated a panel of methylation-based biomarkers which could be used for prostate cancer early diagnosis.

Background

DNA methylation is one of the main epigenetic modifications with great potential for biomarker development [1]. Many genes were found as potential biomarkers in risk prediction, diagnosis, prognosis and treatment response in various cancer types based on methylation levels [2]. With a large number of differentially methylated genes identified, however, only 14 of them have been translated into clinical tests [3]. There are multiple reasons for such low translational activity: 1) general bench-to-bedside problems in biomarker development, such as gap between academic discovery and clinical translation, lack of widely accepted standards in biomarker development process etc [4]; 2) characteristics of epigenetic biomarkers as biomarkers, such as lack of accurate location information [3] and heterogenetic cells in samples [5]. In addition, from the peer-reviewed literatures, we speculated that lack of the patient classification in the first step of biomarker development might be another reason. In most comparative analyses for potential biomarkers, the authors simply divided the patient samples into normal tissues and tumors, and thus identified the genes with methylation level differences between two groups [6–9]. Due to the dynamic features of methylation profile in cancer stages [10], however, the general classification methods might not be sufficient.

Metastatic lesions complicate the clinical treatment and thus markedly contribute to cancer associated death. Therefore, it is critical to detect tumor at early stage (before metastasis). In order to identify the early-diagnosis biomarkers, in this study we focused on the localized tumor samples and identified a panel of biomarkers that can effectively distinguish localized tumors from normal tissues across many cancer types. After comparing biomarkers across all cancer types, we identified a large number of biomarkers specific for prostate cancer diagnosis. Three sets of biomarkers were identified by further selection and multiplex analysis. For early detection, a panel of 7 biomarkers (cg26140475, cg24891312, cg24522654, cg21359747, cg03254336, cg12697139, cg19034132) could distinguish between localized prostate tumors and normal tissues. More interestingly, a single biomarker 'cg26140475' allowed us to identify prostate tumor from all urinary tumors. In summary, our study identified and validated a group of methylation-based biomarkers which could be used for early diagnosis of prostate cancer.

Methods

Data collection

All cancer types with more than 5 normal samples and metastatic information were selected from TCGA (https://portal.gdc.cancer.gov/). DNA Methylation data (Illumina Human Methylation 450) for 16 main cancer subtypes were downloaded for this study: bladder urothelial carcinoma (BLCA), breast invasive ductal carcinoma (D_BRCA), breast invasive lobular carcinoma (L_BRCA), colon adenocarcinoma (COAD), esophageal adenocarcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), renal clear cell carcinoma (KIRC), renal papillary cell carcinoma (KIRP), hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), rectal adenocarcinoma (READ), follicular thyroid carcinoma (F_THCA) and papillary thyroid carcinoma (P_THCA).

The probe annotation file was downloaded from GEO (GPL13534, HumanMethylation450_15017482, Illumina Inc.).

Clinical data were downloaded from cBioPortal for Cancer Genomics (http://www.cbioportal.org/) [11] by its web API.

Data processing

In AJCC CANCER STAGING MANUAL, tumor node metastasis (TNM) system is used as a general criterion to classify cancers by size and extent of the primary tumor (T), involvement of regional lymph node (N), and presence or absence of distant metastases (M) [12]. Based on TNM system, tumor samples in our study were classified into two main groups based on their metastasis state: localized tumors (N0 & M0: No regional lymph node metastases and No distant metastases), metastatic tumors (Regional lymph node metastases or Distant metastases or both). Sample compositions of all 16 cancer types were listed in Supplementary table S1. Early diagnosis biomarkers in our study were defined as the biomarkers that can effectively distinguish localized tumors from normal tissues.

Methylation levels were measured as β values for all 485,577 probes. β value, calculated as the ratio of methylated probe signal and total probe signal, ranges from 0 (entirely unmethylated) to 1 (entirely methylated). M value, calculated as log2 ratio of methylated probe signal and unmethylated probe signal, can be transferred from β value by equation (1). M values were found to be more statistically valid [13] and thus were used in our methylation analyses.

$$M = \log 2\left(\frac{\beta}{1-\beta}\right) \quad (1)$$

For each cancer type, probes containing >50% missing data in normal or tumor samples were removed. For cancer types with less than 10 normal samples, only probes containing no missing data in normal samples were used. The number of probes left for further steps ranged from 395,529 (READ) to 396,062 (BLCA & KIRC). Then, 'impute.knn' function from R package 'impute' (1.48.0) [14] was used to estimate missing data by 10 nearest neighbor averaging. Finally, probes located on sex chromosomes, and cross-reactive probes (probes that hybridize to alternate sequences) [15] were removed.

Data analysis and Biomarker selection

R package 'limma' (3.30.13) [16] was used to compare among normal samples (N), localized tumor samples (LT) and metastatic tumor samples (MT) for all 16 cancer types respectively based on their M values. Potential biomarkers need to satisfy following criteria:

- Significant methylation difference between N and LT (FDR < 0.05);
- Large mean β difference (> 0.1) between N and LT;
- Significant methylation difference between N and MT (FDR < 0.05);
- No correlation between factors (sex, race and age) and methylation level (FDR > 0.05).

After getting potential biomarkers for all cancer types, two types of biomarkers with biological and clinical significance were further selected:

1. Pan-cancer biomarkers were defined as biomarkers that existed in ≥ 80% cancer types and all cancer types revealed the same variation trend from N to LT.
2. Cancer-specific biomarkers were defined as biomarkers that existed in ≥ 60% cancer types and one cancer type (two cancer subtypes were also allowed here) revealed opposite variation trend from N to LT with other cancer types.

Biomarker validation

Validation datasets were selected from GEO based on two criteria. First, data was produced by the same platform (Illumina Human Methylation 450). Second, each dataset contained more than 5 normal tissues and tumor tissues.

A total of 11 GEO validation datasets from 7 main cancer types were as follows: GSE60185 (46 normal and 208 BRCA), GSE69914 (50 normal and 305 BRCA), GSE42752 (19 normal and 22 COAD-READ), GSE48684 (17 normal and 64 COAD-READ), GSE61441 (46 normal and 46 paired KIRC), GSE56588 (10 normal and 224 LIHC), GSE77269 (20 normal and 20 paired LIHC), GSE66836 (19 normal and 164 LUAD), GSE49149 (29 normal and 167 PAAD), GSE76938 (63 normal and 73 PRAD), GSE112047 (16 normal and 31 PRAD).

Processing of GEO datasets was in general the same as TCGA: imputing missing values, then analyzing M values between normal tissues and tumor tissues by 'limma' package. Biomarkers were validated only if they performed similar methylation variation pattern between TCGA and GEO datasets. Validated pan-cancer sites need to have consistent significant methylation variation (FDR < 0.05) between normal tissues and tumors in ≥ 80% GEO datasets. Validated cancer-specific biomarkers must have consistent significant methylation variation (FDR < 0.05) in ≥ 60% GEO datasets and matched specific cancer type with TCGA.

Model construction and validation

Logistic regression models were constructed based on methylation level (β) to measure the predictive ability of the biomarkers. In order to construct model using the least critical biomarkers, we referred to the Leave One Out Cross Validation (LOOCV) method. For dataset with n samples, each time we left one sample out and built model with the lowest Akaike Information Criterion (AIC) based on the rest n-1 samples. After summarizing all n models, we selected biomarkers included in all models as critical ones. The final model was constructed only by the critical biomarkers.

Four datasets from GEO was downloaded and used as validating datasets. GSE47915 (4 Gleason-6 prostate tumors and 4 benign prostate tissues), GSE76938 (63 normal and 73 PRAD), GSE112047 (16 normal and 31 PRAD) and GSE52955 (Kidney: 6 normal and 17 tumor samples; Bladder: 5 normal and 25 tumor samples; Prostate: 5 Normal and 25 tumor samples).

Pyrosequencing sensitivity test in urine samples

LNCaP cells were cultured in RPMI-1640 medium supplemented with 10% FBS. One control group (30ml normal urine) and two experimental groups (30ml normal urine spiked in 10^3 or 10^4 LNCaP cells) were prepared. DNA was extracted by urine DNA extraction kit (Ningbo AJcore, China) and then treated with QiagenEpiTect Bisulfite Kit (Qiagen, 59104) for bisulfite conversion. After PCR amplification, the methylation level of cg26140475 loci was measured by pyrosequencing.

The authors state that they have obtained approval from Hangzhou Medical College's medical ethics committee and have followed the principles outlined in the Declaration of Helsinki for all human experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Results

Identification of methylation biomarkers

To take advantage of tumor node metastasis (TNM) information from the TCGA datasets, we divided the tumor samples into localized and metastatic tumors. The localized tumors here were defined as tumors with neither regional lymph node metastases nor distant metastases. We then performed a multi-comparison for normal tissues and localized tumors or metastatic tumors based on a linear model. Genomic sites with a significant and detectable methylation difference between localized tumors and normal tissues (FDR < 0.05, mean β difference > 0.1) were considered as the biomarker candidates. In order to minimize the influence by the metastatic tumors in the future clinical application, we also took into account the significant methylation difference between metastatic tumors and normal tissues (FDR < 0.05). After removing the genomic sites with methylation status correlated with irrelevant factors (e.g. race, age, sex), we identified a large group of methylation biomarkers for each cancer type. The number of biomarkers ranged from 20,053 for F_THCA to 256,401 for LUSC. By integrating biomarkers from all cancers, we identified pan-cancer biomarkers and cancer-specific biomarkers respectively. The whole biomarker selection procedure was summarized in Fig. 1.

- Pan-cancer biomarkers

A panel of 383 sites with uniform methylation pattern was identified as candidate pan-cancer biomarker, which showed commonality across multiple cancer types and therefore had potential in general cancer screening. After GEO datasets validation, we confirmed a total of 299 pan-cancer methylation sites (289 hyper-methylated and 10 hypo-methylated) (Fig. 2A). Detailed information of pan-cancer biomarkers was listed in Supplementary table S2.

- Cancer-specific biomarkers

In order to identify the cancer-specific biomarkers, we searched for the unique methylation patterns for a particular cancer type. We found 359 sites with opposite methylation status between the certain cancer type and others. After GEO datasets validation, we identified 173 cancer-specific biomarkers (125 PRAD-specific, 20 PAAD-specific, 10 BRCA-specific, 9 KIRC-specific, 7 LIHC-specific and 2 COADREAD-specific) (Fig. 2B). Detailed information of cancer-specific biomarkers was listed in Supplementary table S3.

Development of prostate cancer specific biomarkers

In total, we identified 125 prostate cancer specific methylation loci (Fig. 2B). To further narrow down to the most significant biomarkers for potential clinical usage, we included Gleason score in our selection procedure. Gleason score grades prostate cancer by the histological appearance of carcinoma cells [17]. Given the prognosis difference between Gleason score 3+4 & 4+3 [18] and the sample size in each group, we divided all prostate tumors into two groups (low Gleason score group: 6 & 3+4; high Gleason score group: 4+3 & 8-10). After a multi-comparison between normal and low/high Gleason score tumors, 26 methylation biomarkers were identified with a significant difference (FDR < 0.05) between normal tissues

and low/high Gleason score tumors, and more importantly, between low Gleason score tumors and high score ones. Among the 26 selected biomarkers, 15 loci were located in intergenic region and the rest 11 were located in 11 genes (*HIVEP3*, *KCMF1*, *RAP1GDS1*, *PRB2*, *OSBPL3*, *PEX14*, *PKD1L1*, *ALDH1A3*, *SDK1*, *FAM13A*, *BRE*) respectively (Table 1).

Predictive ability of the prostate cancer specific biomarkers

- Normal tissues vs. localized prostate tumors

By following LOOCV method, we identified 7 critical biomarkers from 26 biomarkers (cg26140475, cg24891312, cg24522654, cg21359747, cg03254336, cg12697139, cg19034132). A logistic regression model, 12.234*cg26140475 + 4.247*cg24891312 + 14.924*cg24522654 - 5.902*cg21359747 - 13.520*cg03254336 - 3.001*cg12697139 - 4.841*cg19034132 - 0.425, was built to validate the discriminatory power to separate early prostate tumors (Gleason 6) from normal tissues (AUC = 1, dataset: GSE47915, Fig. 3A). Because the GEO dataset (GSE47915) has a small number of patients with the Gleason score, we validated the model using a broad classification criterion (tumor & normal) in different datasets. Our data showed that the model can effectively detect prostate tumors from normal tissues (AUC = 0.92, GSE76938; AUC = 0.99, GSE112047, Fig. 3A).

- Low Gleason score tumors vs. high Gleason score tumors

We also identified 9 out of 26 biomarkers to have the ability to separate between low and high Gleason score prostate tumors. Because no appropriate dataset available, 10 fold cross validation method was used to evaluate the model. Data was divided into 10 parts and each time 9 parts were used to construct the model and the rest 1 part to draw the receiver operating characteristic curve (ROC). Our analysis showed the 9 biomarkers (cg17220055, cg26140475, cg24891312, cg09853702, cg22400059, cg16736279, cg27639613, cg06011086, cg00664697) was able to separate low from high Gleason score prostate tumors (average AUC = 0.79, Fig. 3B).

- Prostate cancer vs. other urinary tumors

Clinical relevant biomarkers are expected to be prostate cancer specific, especially among other urinary related cancers. Therefore, we built a logistic regression model for two groups of the localized tumors, PRAD and non-PRAD (BLCA, KIRC and KIRP). Based on the large differences among the urinary tumors (Fig. 3C), a simple model using one biomarker cg26140475, 21.3216*cg26140475 - 5.2742, can correctly separate the prostate tumors from other urinary tumors in GSE52955 with AUC = 1.

Pyrosequencing's ability in detecting single loci methylation

To detect the single loci methylation level, pyrosequencing has been proved to be an appropriate method [19]. In order to investigate the sensitivity in urinary samples, we designed an experiment by spiking in prostate cancer cells to the normal urine to mimic cancer status. After reviewing the methylation level of all commercially available prostate tumor cells from GEO datasets (GSE114598, GSE34340, GSE49143,

GSE54758, GSE66872, GSE71626, GSE79185, GSE86829) (Fig. 4A), we found LNCaP cells have raised methylation level (β was around 0.6, by Illumina Human Methylation 450) of loci cg26140475. Therefore, we spiked in different number of LNCaP cells to normal urine and measured the methylation level of biomarker “cg26140475” by pyrosequencing. After comparing the control urine with the spiked samples (10^3, 10^4 cells/30ml urine), we found that pyrosequencing can detect the methylation level of biomarker in the urine sample spiked 10,000 cells (Fig. 4B).

Discussion

While numerous publications identified the methylation biomarkers from public databases, most of them didn’t specify the tumor samples based on the clinical information. To our best knowledge, this study is the first one to classify TCGA tumor samples into localized tumors and metastatic tumors based on TNM information. Using a pan-cancer comparative analysis on locus methylation level between normal tissues and localized tumors, we identified a large number of cancer specific biomarkers and finally focused on 26 potential prostate cancer specific biomarkers. Models with multiplex biomarkers can effectively distinguish between normal & localized prostate tumors (7 biomarkers, AUC > 0.9), low & high Gleason score prostate tumors (9 biomarkers, AUC = 0.79) and prostate tumors & other urinary related tumors (1 biomarker, AUC = 1).

The anatomical position of urethra and urinary organs makes it possible for urinary components mixing into urine. When urinary cancer occurs, urine therefore becomes a perfect medium for molecular biomarker detection. Indeed a large number of urinary DNA methylation tests have already been developed for bladder cancer, prostate cancer and renal cancer, mostly depending on quantitative methylation-specific polymerase chain reaction (qMSP) [20]. For prostate cancer, commercially available urine-based methylation test is still missing. The only methylation-based test is ‘ConfirmMDx’ (MDxHealth, Inc, Irvine, CA), which relies on the methylation status of three genes (GSTP1, APC and RASSF1) from biopsy tissues to avoid unnecessary repeat biopsies [21].

To satisfy the general application, multiple urine tests had been developed for prostate cancer. For example, O'reilly et al. combined a DNA methylation panel of previously reported genes (GSTP1, SFRP2, IGFBP3, IGFBP7, APC, and PTGS2) to detect high-risk prostate tumors by urine samples [22]. Similarly, Zhao et al. developed a urinary methylation assay based on methylation level of two genes (HOXD3 and GSTP1) [23] and Bakavicius et al. combined PSA test and urinary methylation tests of three genes (RARB, RASSF1, GSTP1) [24]. The technique of qMSP, however, only measures the general methylation of gene region and requires aberrant methylation in CpG islands [25]. Our study released the limit by identifying locus within both CpG islands and non-CpG islands regions. Particularly, pyrosequencing [19], a technique to quantify single loci methylation level, ensures an appropriate way for the clinical application in the future. For example, Yao et al. demonstrated pyrosequencing could effectively detect the methylation difference of loci cg05163709 between positive and negative biopsies patients’ urines [26]. Our result showed that pyrosequencing was able to detect the loci methylation level of 10,000 cancer cells spiked in 30ml urine. Though the sensitivity wasn’t high enough, urine samples collected in the early morning or

after DRE (digital rectal examination) could largely increase prostate cell concentration. Furthermore, using filtrations to collect urinary cells was also found to be effective in increasing sensitivity [27].

Although this study identified a panel of biomarkers which are promising for clinical application in the future, based on our preliminary result of pyrosequencing sensitivity, further study with clinical urine samples will be required.

Conclusions

This study identified and validated a panel of methylation-based biomarkers which could be used for prostate cancer early diagnosis.

Abbreviations

AIC: Akaike Information Criterion; AUC: area under the curve; BLCA: bladder urothelial carcinoma; COAD: colon adenocarcinoma; D_BRCA: breast invasive ductal carcinoma; ESCA: esophageal adenocarcinoma; F_THCA: follicular thyroid carcinoma; GEO: Gene Expression Omnibus; HNSC: head and neck squamous cell carcinoma; KIRC: renal clear cell carcinoma; KIRP: renal papillary cell carcinoma; LIHC: hepatocellular carcinoma; LOOCV: Leave One Out Cross Validation; LT: localized tumor samples; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; L_BRCA: breast invasive lobular carcinoma; N: normal samples; MT: metastatic tumor samples; PAAD: pancreatic adenocarcinoma; PRAD: prostate adenocarcinoma; PSA: prostate specific antigen; P_THCA: papillary thyroid carcinoma; qMSP: quantitative methylation-specific polymerase chain reaction; READ: rectal adenocarcinoma; ROC: receiver operating characteristic curve; TCGA: The Cancer Genome Atlas; TNM: tumor node metastasis.

Declarations

Ethics approval and consent to participate: The authors state that they have obtained approval from Hangzhou Medical College's medical ethics committee and have followed the principles outlined in the Declaration of Helsinki for all human experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Consent for publication: Not applicable

Availability of data and materials: All data generated or analysed during this study are included in this published article and its supplementary information files.

Competing interests: The authors declare that they have no competing interests.

Funding: This research was supported by Youth Foundation of Zhejiang Academy of Medical Sciences (grant no. CF2008Q04) and National Natural Science Foundation of China (grant no. 81773626). The authors have no other relevant affiliations or financial involvement with any organization or entity with a

financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Authors' contributions: YY Pu and XJ Wang were responsible for the overall conception and design. YY Pu and C Li were responsible for the experimental verification. YY Pu, HN Yuan and C Li were responsible for analysis and interpretation of the data, and drafting of the manuscript. YY Pu and XJ Wang were responsible for revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: The authors thank all colleagues in School of Bioengineering, especially Yanmei Zhang, Jianfei Fang for their helpful advices.

References

1. Laird PW: **The power and the promise of DNA methylation markers**. *Nat Rev Cancer* 2003, **3**(4):253-266.
2. Tsai HC, Baylin SB: **Cancer epigenetics: linking basic biology to clinical medicine**. *Cell Res* 2011, **21**(3):502-517.
3. Koch A, Joosten SC, Feng Z, de Ruijter TC, Draht MX, Melotte V, Smits KM, Veeck J, Herman JG, Van Neste L *et al*: **Analysis of DNA methylation in cancer: location revisited**. *Nat Rev Clin Oncol* 2018, **15**(7):459-466.
4. van Gool AJ, Bietrix F, Caldenhoven E, Zatloukal K, Scherer A, Litton JE, Meijer G, Blomberg N, Smith A, Mons B *et al*: **Bridging the translational innovation gap through good biomarker practice**. *Nat Rev Drug Discov* 2017, **16**(9):587-588.
5. Garcia-Gimenez JL, Mena-Molla S, Beltran-Garcia J, Sanchis-Gomar F: **Challenges in the analysis of epigenetic biomarkers in clinical samples**. *Clin Chem Lab Med* 2017, **55**(10):1474-1477.
6. de Almeida BP, Apolonio JD, Binnie A, Castelo-Branco P: **Roadmap of DNA methylation in breast cancer identifies novel prognostic biomarkers**. *BMC Cancer* 2019, **19**(1):219.
7. Tu Y, Fan G, Xi H, Zeng T, Sun H, Cai X, Kong W: **Identification of candidate aberrantly methylated and differentially expressed genes in thyroid cancer**. *J Cell Biochem* 2018, **119**(11):8797-8806.
8. Fan G, Tu Y, Chen C, Sun H, Wan C, Cai X: **DNA methylation biomarkers for hepatocellular carcinoma**. *Cancer Cell Int* 2018, **18**(1):140.
9. Li H, Liu JW, Liu S, Yuan Y, Sun LP: **Bioinformatics-Based Identification of Methylated-Differentially Expressed Genes and Related Pathways in Gastric Cancer**. *Dig Dis Sci* 2017, **62**(11):3029-3039.
10. McMahon KW, Karunasena E, Ahuja N: **The Roles of DNA Methylation in the Stages of Cancer**. *Cancer J* 2017, **23**(5):257-261.
11. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E *et al*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. *Cancer Discov* 2012, **2**(5):401-404.

12. Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A: **AJCC cancer staging manual (7th)**. New York: Springer; 2010.
13. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis**. *BMC Bioinformatics* 2010, **11**:587.
14. Hastie T, Tibshirani R, Narasimhan B, Chu G: **Impute: Imputation for microarray data**. *Oral History Review* 2011(1):128-130.
15. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R: **Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray**. *Epigenetics* 2013, **8**(2):203-209.
16. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**(7):e47.
17. Humphrey PA: **Gleason grading and prognostic factors in carcinoma of the prostate**. *Mod Pathol* 2004, **17**(3):292-306.
18. Epstein JI, Zelefsky MJ, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, Vickers AJ, Parwani AV, Reuter VE, Fine SW *et al*: **A Contemporary Prostate Cancer Grading System: A Validated Alternative to the Gleason Score**. *Eur Urol* 2016, **69**(3):428-435.
19. Busato F, Dejeux E, El Abdalaoui H, Gut IG, Tost J: **Quantitative DNA Methylation Analysis at Single-Nucleotide Resolution by Pyrosequencing(R)**. *Methods Mol Biol* 2018, **1708**:427-445.
20. Larsen LK, Lind GE, Guldberg P, Dahl C: **DNA-Methylation-Based Detection of Urological Cancer in Urine: Overview of Biomarkers and Considerations on Biomarker Design, Source of DNA, and Detection Technologies**. *Int J Mol Sci* 2019, **20**(11).
21. Wojno KJ, Costa FJ, Cornell RJ, Small JD, Pasin E, Van Criekinge W, Bigley JW, Van Neste L: **Reduced Rate of Repeated Prostate Biopsies Observed in ConfirmMDx Clinical Utility Field Study**. *American health & drug benefits* 2014, **7**(3):129-134.
22. O'Reilly E, Tuzova AV, Walsh AL, Russell NM, O'Brien O, Kelly S, Dhomhnallain ON, DeBarra L, Dale CM, Brugman R *et al*: **epiCaPture: A Urine DNA Methylation Test for Early Detection of Aggressive Prostate Cancer**. *JCO precision oncology* 2019, **2019**:1-18.
23. Zhao F, Olkhov-Mitsel E, Kamdar S, Jeyapala R, Garcia J, Hurst R, Hanna MY, Mills R, Tuzova AV, O'Reilly E *et al*: **A urine-based DNA methylation assay, ProCUrE, to identify clinically significant prostate cancer**. *Clin Epigenetics* 2018, **10**(1):147.
24. Bakavicius A, Daniunaite K, Zukauskaite K, Barisiene M, Jarmalaite S, Jankevicius F: **Urinary DNA methylation biomarkers for prediction of prostate cancer upgrading and upstaging**. *Clin Epigenetics* 2019, **11**(1):115.
25. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW: **MethyLight: A High-Throughput Assay to Measure DNA Methylation**. *Nucleic acids research* 2000, **28**(8):e32.

26. Yao L, Ren S, Zhang M, Du F, Zhu Y, Yu H, Zhang C, Li X, Yang C, Liu H *et al*: **Identification of specific DNA methylation sites on the Y-chromosome as biomarker in prostate cancer**. *Oncotarget* 2015, **6**(38):40611-40621.

27. Larsen LK, Jakobsen JS, Abdul-Al A, Guldberg P: **Noninvasive Detection of High Grade Prostate Cancer by DNA Methylation Analysis of Urine Cells Captured by Microfiltration**. *J Urol* 2018, **200**(4):749-757.

Tables

Table 1. Location information of 26 PRAD-specific biomarkers for modeling.

Probe ID	Corresponding gene	Gene region	Chromosome	MAPINFO	Relation to CpG island
cg17220055	*HIVEP3*	5'UTR	1	42248998	-
cg03598112	-	-	17	28691021	-
cg19291696	-	-	13	27769437	-
cg26140475	-	-	8	126525558	-
cg07002540	*KCMF1*	3'UTR	2	85281128	-
cg27655168	-	-	17	39059398	-
cg14454094	*RAP1GDS1*	Body	4	99352003	-
cg24891312	*PRB2*	TSS1500	12	11549555	-
cg09853702	-	-	12	47665230	-
cg15041658	*OSBPL3*	Body	7	24866066	-
cg22400059	*PEX14*	Body	1	10571508	-
cg13414270	-	-	2	45465395	-
cg26541218	*PKD1L1*	Body	7	47826387	-
cg24522654	-	-	12	19535154	-
cg21359747	*ALDH1A3*	Body	15	101420636	Island
cg16736279	*SDK1*	Body	7	3995563	-
cg08548284	-	-	1	14382172	-
cg18006637	*FAM13A*	Body	4	89713820	-
cg18270378	-	-	7	3284394	-
cg03254336	-	-	10	114635839	-
cg27639613	*BRE*	Body	2	28408263	-
cg12697139	-	-	1	209571889	-
cg06011086	-	-	8	61823275	S_Shore
cg19034132	-	-	10	75692227	-
cg00664697	-	-	16	80061290	-
cg03649855	-	-	6	45576388	-