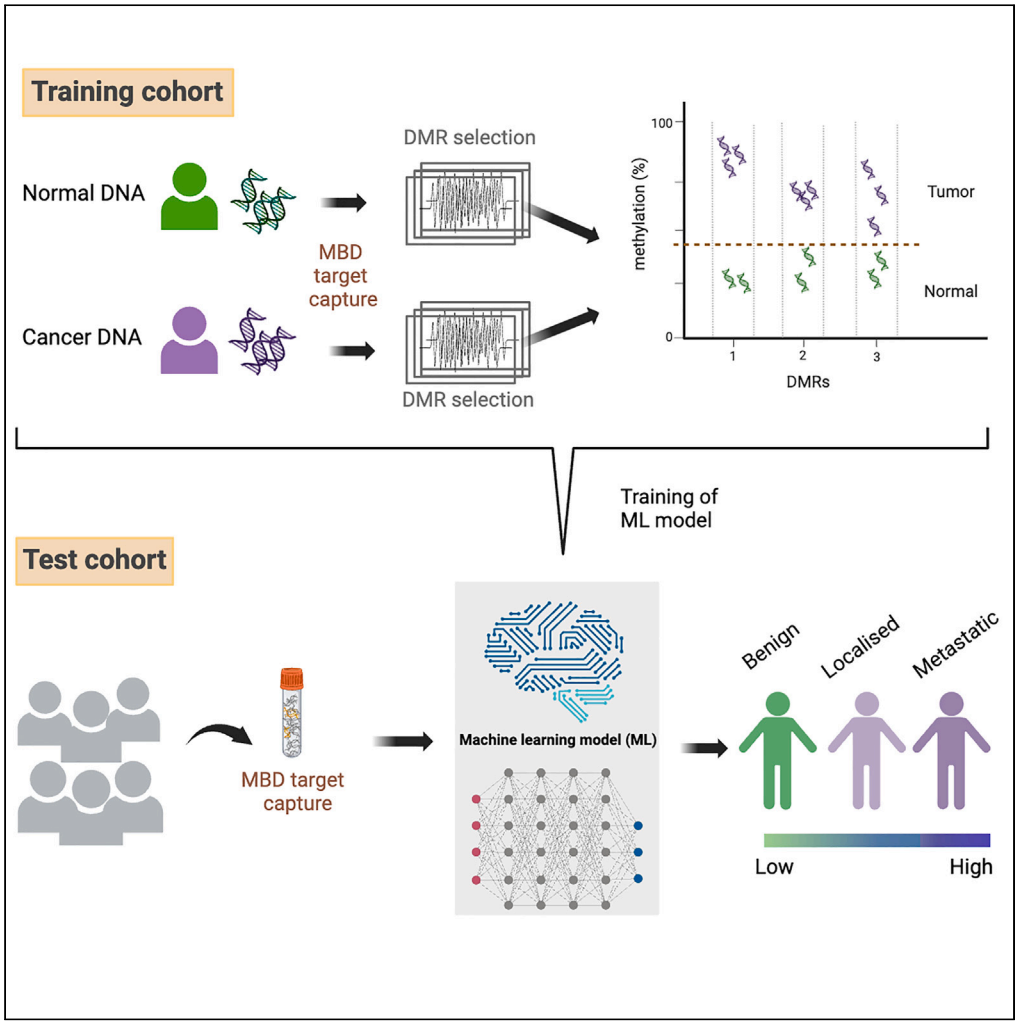
iScience

CellPress
OPEN ACCESS

Article

Prostate cancer detection through unbiased capture of methylated cell-free DNA



Ermira Lleshi, Toby Milne-Clark, Henson Lee Yu, ..., Vincent J. Gnanapragasam, Charlie E. Massie, Harveer S. Dev

hsd26@cam.ac.uk

Highlights

Enrichment of methylated cell-free DNA identifies prostate cancer biomarkers

A machine learning model detects prostate cancer using the identified biomarkers

The biomarkers are enriched for genes in cancer-related signaling pathways

Lleshi et al., iScience 27, 110330
July 19, 2024
https://doi.org/10.1016/j.isci.2024.110330

iScience

CellPress
OPEN ACCESS

Article

Prostate cancer detection through unbiased capture of methylated cell-free DNA

Ermira Lleshi,[1,4,8] Toby Milne-Clark,[1,8] Henson Lee Yu,[1] Henno W. Martin,[1] Robert Hanson,[1] Radoslaw Lach,[1] Sabrina H. Rossi,[1] Anja Lisa Riediger,[2,3] Magdalena Görtz,[2] Holger Sültmann,[3] Andrew Flewitt,[4] Andy G. Lynch,[5,6] Vincent J. Gnanapragasam,[7] Charlie E. Massie,[1,9,10] and Harveer S. Dev[1,9,11,*]

SUMMARY

Prostate cancer screening using prostate-specific antigen (PSA) has been shown to reduce mortality but with substantial overdiagnosis, leading to unnecessary biopsies. The identification of a highly specific biomarker using liquid biopsies, represents an unmet need in the diagnostic pathway for prostate cancer. In this study, we employed a method that enriches for methylated cell-free DNA fragments coupled with a machine learning algorithm which enabled the detection of metastatic and localized cancers with AUCs of 0.96 and 0.74, respectively. The model also detected 51.8% (14/27) of localized and 88.7% (79/89) of patients with metastatic cancer in an external dataset. Furthermore, we show that the differentially methylated regions reflect epigenetic and transcriptomic changes at the tissue level. Notably, these regions are significantly enriched for biologically relevant pathways associated with the regulation of cellular proliferation and TGF-beta signaling. This demonstrates the potential of circulating tumor DNA methylation for prostate cancer detection and prognostication.

INTRODUCTION

Prostate cancer is the third most commonly diagnosed cancer in men and accounts for 7% of all cancer-related deaths worldwide.[1] Molecular stratification for prognostication may improve outcomes for patients at the highest risk of clinical progression. Presently, we rely on using Prostate Specific Antigen (PSA) as a biomarker to initially evaluate men for prostate cancer. However, the limitations of this approach, including high rates of false positives and false negatives, have led to uncertainty regarding the recommendation for universal PSA testing as a screening approach.[2,3]

In recent years, the blood-based test – Stockholm-3 – was developed which combined PSA testing with family history as well as clinical, genetic, and protein biomarkers and was found to decrease the rate of overdiagnosis, thereby decreasing unnecessary biopsies by 34% compared to using PSA alone.[4] Multiple studies, including our own, have demonstrated the genomic heterogeneity of prostate cancers.[5,6] Wyatt and colleagues established that genetic alterations and copy number variations found in metastatic prostate cancer tissues are concordant with those detected in liquid biopsies.[7] However, studies that focused on early-stage prostate tumors demonstrate a lack of consensus mutations, with the most prevalent genetic alterations found in the SPOP gene only occurring in 8–13% of cases.[8,9] Consequently, non-genetic approaches may offer a more robust strategy for cancer diagnostics.

Epigenetic changes have been described preceding genetic alterations and may share similar patterns even among genetically distinct tumors, particularly in early-stage cancers. For example, we previously showed that HES5 promoter hypermethylation is found among 38 out of 39 (97%) early-stage prostate cancers.[10] In fact, several studies have shown the utility of epigenetic markers for cancer detection, risk prediction, and treatment monitoring.[11,12] These techniques include single-base resolution approaches such as bisulfite-sequencing or enzymatic methyl conversion-sequencing, which may be highly informative but are costly.[13,14] Alternatively, indirect techniques that are more economical provide methylation information at lower resolution. Advancements in both approaches as well as sequencing technologies,

[1]Early Cancer Institute, Department of Oncology, University of Cambridge, Cambridge CB2 0XZ, UK
[2]University Hospital Heidelberg, 69120 Heidelberg, Germany
[3]Division of Cancer Genome Research, German Cancer Research Center (DKFZ), National Center for Tumor Diseases (NCT), 69120 Heidelberg, Germany
[4]Department of Engineering, University of Cambridge, Cambridge, UK
[5]School of Mathematics and Statistics, University of St Andrews, St Andrews KY16 9SS, UK
[6]School of Medicine, University of St Andrews, St Andrews KY16 9TF, UK
[7]Department of Surgery, University of Cambridge, Addenbrooke's Hospital Site, Cambridge, UK
[8]These authors contributed equally
[9]These authors contributed equally
[10]Deceased
[11]Lead contact
*Correspondence: hsd26@cam.ac.uk
https://doi.org/10.1016/j.isci.2024.110330

present a unique opportunity to explore the potential of using these methods with plasma samples, circumventing the costs, risks, and challenges associated with prostate biopsies.

Bryzgunova et al. have demonstrated an ability to discriminate normal, benign prostatic hyperplasia, and prostate cancer within a small cohort of samples by evaluating the DNA methylation levels of just three genes (*GSTP1, RNF219, KIAA1539*).[15] Because the method targets only a few loci, it reduces the sensitivity of this cell-free DNA (cfDNA) based detection assay, given that, on average, there are only around 1,000 genome equivalents present per milliliter of plasma. Chen et al. employed an immunoprecipitation-based strategy which enriched for methylated fragments in the plasma demonstrating the capability of distinguishing localized from metastatic prostate cancers.[16] However, there was a lack of evidence to discriminate any of these cancers from normal cases, which is essential for the practical real-world use of these assays within a clinical diagnostic pathway.

Since multiple cancers may have shared epigenetic features, a recent study applied this approach to the detection of multiple cancers simultaneously vis-à-vis a pan-cancer methylation detection platform. They reported a promising performance of using cfDNA methylation in detecting various cancer types (sensitivity of 51% and specificity of 99.5%), although the performance varied greatly across different cancer types.[17] For prostate cancer, a sensitivity of only 11.2% was achieved.[17]

In this study, methyl CpG-binding domain protein (MBD) was used for the enrichment of methylation-rich fragments. We employed machine learning techniques to develop a classifier capable of distinguishing cancer and non-cancer cases based on these enriched fragments. The panel of differentially methylated regions (DMRs) selected through this strategy exhibited accurate performance in detecting metastatic prostate cancers. Notably, these DMRs also detect early-stage disease, albeit with lower accuracy. In addition, we also investigated the biological implications of these DMRs to infer underlying mechanisms and gain insights into their functional significance. We shed light on potential molecular pathways and regulatory processes which may underpin these signals and contribute to prostate cancer development and progression. Hence, this allows for the allowed discrimination of prostate cancer from non-cancer cases using circulating tumor DNA methylation signals.

RESULTS

Evaluation of enrichment methods for methylated cell-free DNA targets

We began by evaluating two common strategies for enriching methylated fragments in cfDNA samples: cell-free methylated DNA immunoprecipitation-sequencing (cfMeDIP-seq)[18] and cell-free methyl CpG-binding Domain protein-sequencing (cfMBD-seq).[19] Both techniques preferentially enrich for methylated DNA by using either an antibody that captures methylated DNA (cfMeDIP) or an antibody that is specific for methyl group-binding proteins (cfMBD); however, studies show that the latter may yield more high-quality reads with fewer duplicates.[19] To validate this finding, we performed both techniques and compared their ability to capture cancer specific reads. We first identified prostate cancer methylation signatures by calling differentially methylated regions between tumor and normal samples from TCGA as described in our previous work.[10] Since MeDIP-seq and MBD-seq infer methylation levels from the sequencing coverage, we divided the human genome into 100-base windows and identified 6,285 such non-overlapping 100-base bins that intersected with the combined DMRs from TCGA. These final set of 6,285 regions served as our targets to assess the performance of cfMeDIP-seq and cfMBD-seq in capturing informative reads.

We used PC3 as a cell model for advanced prostate cancer, and simulated cfDNA shedding through the collection of tissue culture supernatant (Figure S1). We then performed cfMBD-seq and cfMeDIP-seq on the extracted cfDNA to obtain a preliminary evaluation of the number of reads that could be obtained from each method and quantified the overlap with our pre-defined tissue-derived DMRs (Figure 1A). We observed cfMBD-seq was able to capture reads across a wider range of CpG densities compared to cfMeDIP-seq, particularly those with very low CpG's density (Figure 1B). This was a consistent finding when comparing the coverage across CpG islands, shelves, and shores (Figure S2A).

To further analyze the methylation information obtained from cfMBD-seq and cfMeDIP-seq, we utilized the 100-base bins previously described and calculated the percentage of methylation based on the number of reads aligned to each bin, following established methods.[20] The level of methylation is directly proportional to the read depth, as more methylated fragments are more likely to be bound by the antibody (MeDIP) or methyl-binding protein (MBD). We observed a strong correlation between the coverage and the methylation level obtained from both methods (Figure 1C; Figure S2B). We then compared their sequence coverages and the inferred methylation levels with a direct methylation level obtained from a single-base resolution enzymatic methyl conversion method (EM-seq). Both cfMBD-seq and cfMeDIP-seq demonstrated a modest correlation with cell-free EM-seq (cfEM-seq) (r^2 of 0.56 and 0.52, respectively) (Figure 1D; Figure S2C). From the correlation plots between the enrichment methods and cfEM-seq in Figure 1C, we also see a stronger correlation for DNA fragments with higher methylation levels. Based on these findings, we set a 30% methylation level threshold for both cfMBD-seq and cfMeDIP-seq, while the threshold for cfEM-seq was set at 50%. We then counted the number of fragments that met these criteria and were also present in the 6,285 fragments previously identified in the tissue samples. We found that cfMBD-seq captured 5,507 (88%) of these regions while cfMeDIP-seq captured 5,033 (80%) regions (Figure S2D). Consequently, we selected cfMBD-seq as the preferred method for further experiments.

Training a machine learning model to detect prostate cancer signatures

In order to simultaneously analyze the methylation levels of thousands of fragments, we employed gradient boosted and random forest machine learning algorithms to learn patterns of methylation from the plasma samples of a cohort of 35 normal (i.e., prostate cancer-free) and 62 patients with metastatic prostate cancer. All demographic and clinicopathological information of the patients are summarized in Table S1.

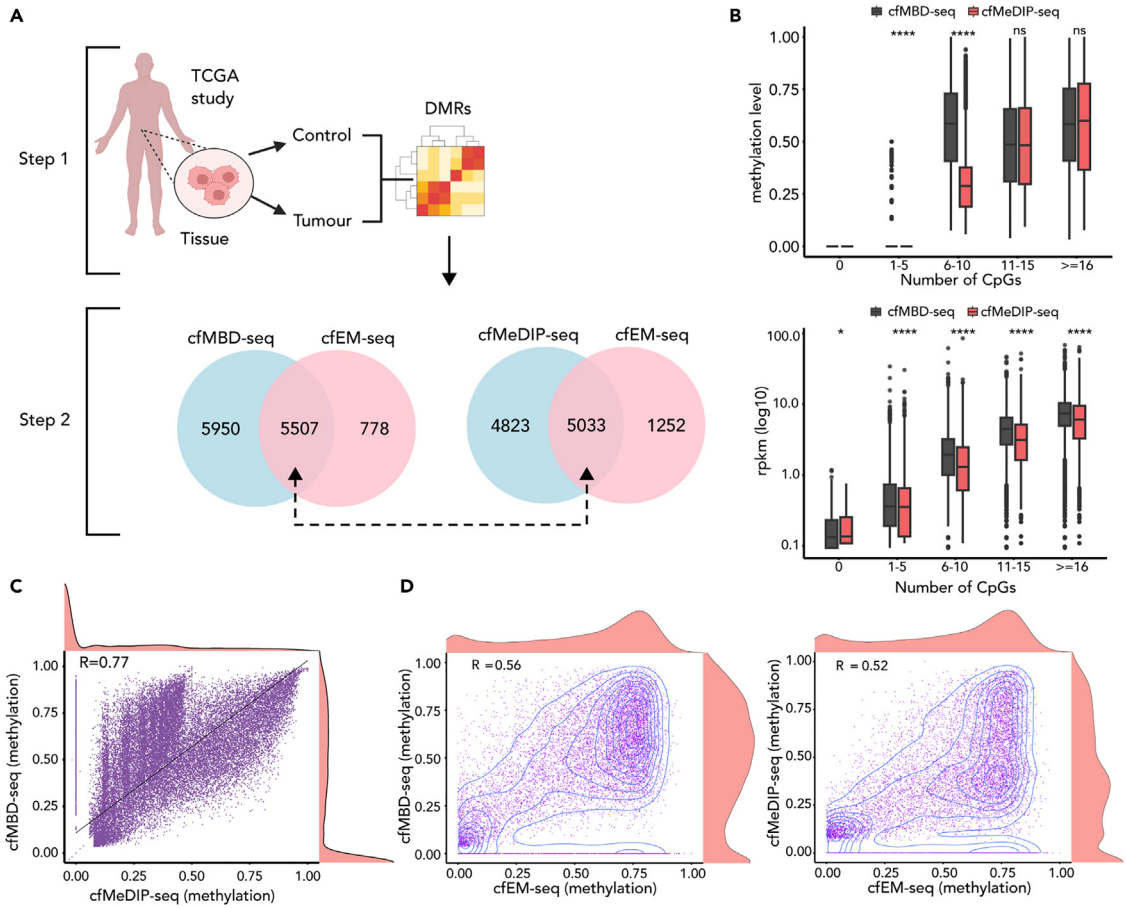


Figure 1. Comparison of methylation enrichment approaches in cfDNA

(A) Workflow for selecting the cancer-specific differentially methylated regions defined by dividing the human genome into 100 bp windows and choosing those that intersects with the DMRs discovered using TCGA methylation data. The Venn diagrams show the overlap of the defined regions with the regions enriched using MBD-seq and MeDIP-seq.

(B) Comparison of the methylation levels (top) and coverage (bottom) obtained from MBD-seq and MeDIP-seq across varying numbers of CpGs showing that MBD-seq consistently reports higher values and greater sensitivity than MeDIP-seq (Kruskal-Wallis test: *$p < 0.05$, ****$p < 0.0001$, **ns** not significant).

Correlation of the C absolute methylation levels that were calculated from cfMBD-seq and cfMeDIP-seq; and D methylation levels of cfMBD-seq (left) and cfMeDIP-seq (right) and cfEM-seq. For both C and D, each dot is a 100-bp genomic window. Spearman's rank correlation coefficients are shown.

These samples were collected from patients on a prostate cancer diagnostic pathway, following clinical referral to a specialist due to elevated PSA, leading to MRI and prostate biopsy for tissue diagnosis. We randomly assigned 70% of the samples as the training set (24 normal controls and 43 cancer cases), while the remaining samples constituted the test set. Methylation levels exhibited substantial variation within the training set among the patients with prostate cancer. Previous studies have indicated a proportional relationship between methylation levels and the fraction of tumor-specific content in cfDNA.[21] Therefore, we applied a previously established method that uses copy number aberrations (IchorCNA[22]) to estimate tumor fraction based on shallow whole genome bisulfite sequencing (sWGBS) for a subset of our cohort. We were able to confirm the correlation between methylation levels and circulating tumor fraction (Figures S3A and S3B).

We extended the application of our copy number aberration (CNA) analysis on a targeted sequencing approach by showing that we are able to obtain the similar results whether the method was applied to sWGBS or MBD data (r = 0.98) (Figure S3C). This facilitated the use of cfMBD-seq data to infer the tumor fraction for the remaining samples. We observed a range of detectable copy number aberrations in control cases of up to 15% (Figure S3D), which may be higher than the reported noise levels from other studies.[23,24] But this is concordant with the copy number variation map of the human genome, which reveals that CNVs take up 5–10% of the human genome and may not necessarily have functional or health implications.[25] In addition, the CNVs are mostly concentrated on non-coding regions which are preferentially selected in MBD.[25] Hence, we selected metastatic cases with tumor fractions exceeding 15%, as indicated in the workflow in Figure 2. The process of identifying DMRs was repeated 30 times, resulting in a total of 32,679 DMRs that were consistently identified in all 30 iterations. Among these DMRs, 24,654 (75.4%) exhibited hypermethylation in tumors compared to normal samples, while 8,280 (25.3%) were

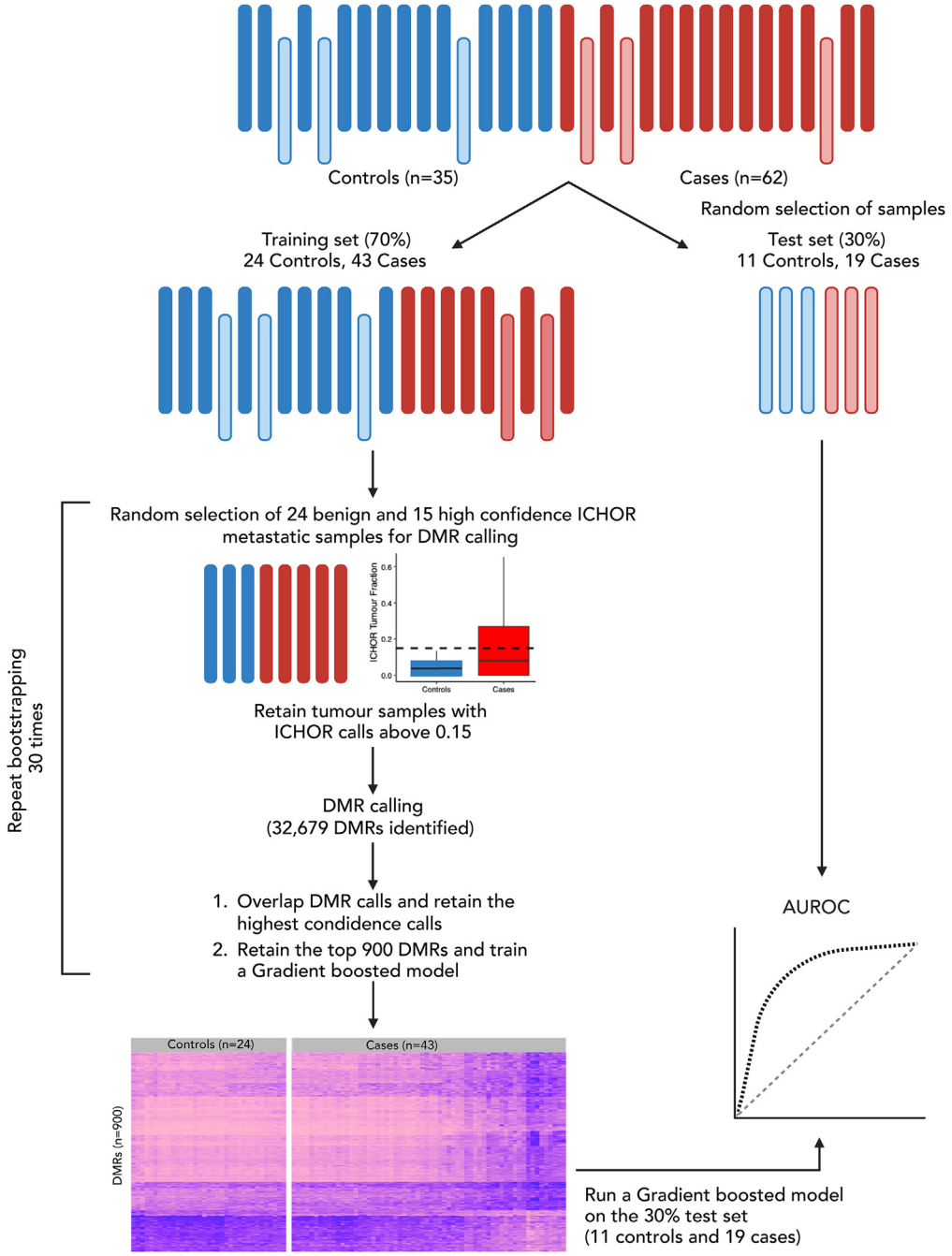


Figure 2. Schematics of building a gradient boosted machine learning model

Patient samples (n = 97; 35 benign controls and 62 metastatic cases) were randomly assigned to training (n = 67; 24 benign and 43 cases) and test (n = 30; 11 benign and 19 cases) sets. Differentially Methylated Regions (DMRs) were identified using a subset consisting of all 24 benign cases and a random selection of 15 metastatic cases that had high tumor fraction, as determined by ichorCNA analysis. This subset selection process was repeated 30 times, with each iteration randomly selecting a set of 15 cases that met the tumor fraction threshold (indicated by the horizontal dotted line on the inset ichor tumor fraction graph). The 900 DMRs exhibiting the largest absolute difference in methylation, and consistently identified in all 30 iterations, were used to train the gradient boosted model using all 67 samples in the training set. The performance of the model was then evaluated on the independent test set.

hypomethylated. From the pool of 32,679 DMRs, we selected the top 900 with the highest absolute difference in methylation levels for further analysis, based on their optimal AUC for distinguishing metastatic from control cases (Figures S4A and S4B). These 900 DMRs were used as input to train a machine learning model across the full 70% training set, irrespective of tumor fraction.

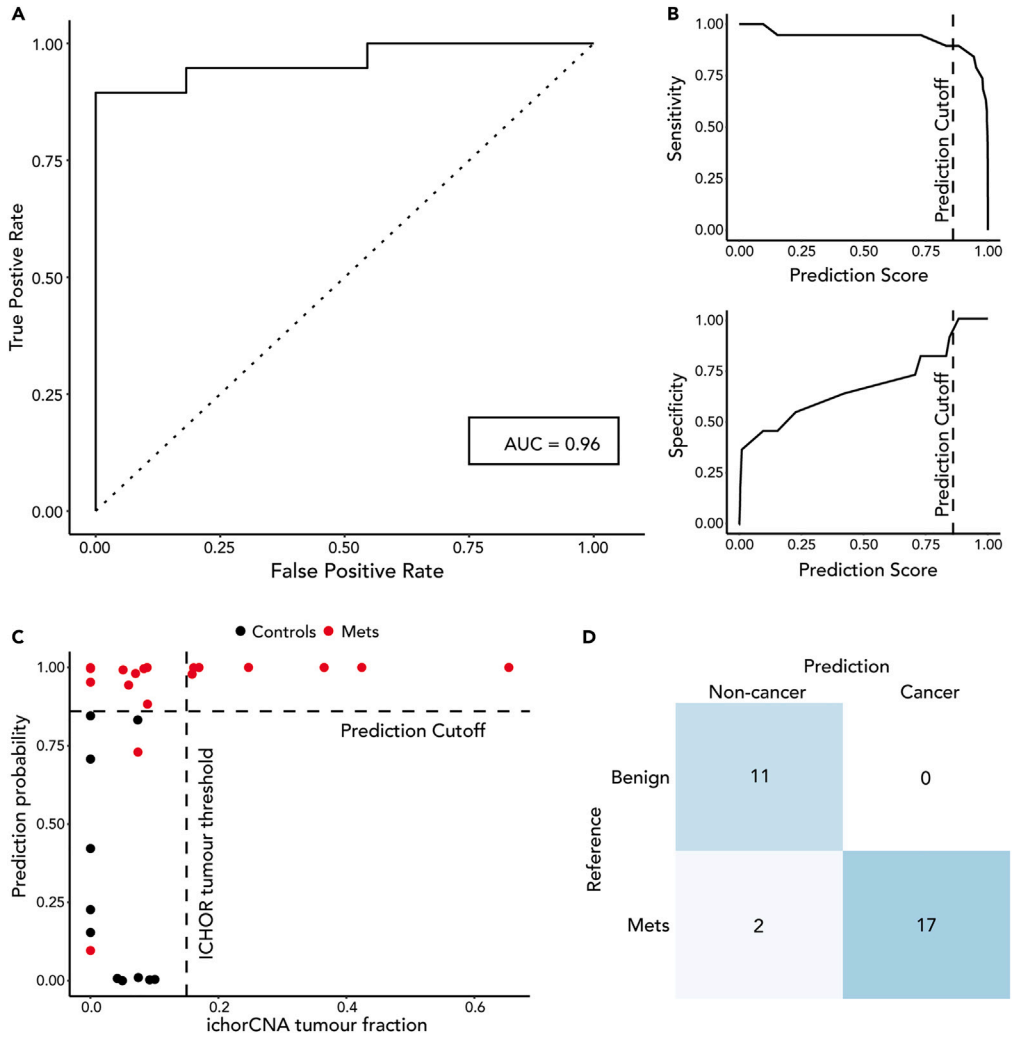


Figure 3. Performance of the machine learning model in the detection of metastatic prostate cancer

(A) ROC showing the performance of the machine learning model in detecting metastatic prostate cancer samples in the test set (n = 30).

(B) Sensitivity (top) and specificity (bottom) of the model at different prediction score cut-off values. The selected cut-off value (0.86) is indicated by the dotted lines, corresponding to a specificity of 100% and a sensitivity of 95%.

(C) Comparison of the performance of the machine learning model (with a prediction probability cut-off of 0.86; horizontal dotted line) and ichorCNA (with a threshold set at 0.15; vertical dotted line) in classifying benign and metastatic cases.

(D) Summary of the number of correctly and incorrectly classified samples using the machine learning model.

Performance of the machine learning model in differentiating normal vs. tumor samples

The average methylation data, heatmap analysis, and principal component analysis (PCA) clustering of the methylation values for the top 900 DMRs alone did not effectively distinguish between cases and controls (Figures S5A–S5C). However, both the gradient boosted model and Random Forest model demonstrated the ability to differentiate the test set, achieving AUC scores of 0.96 and 0.95 respectively (Figure 3A; Figures S6A and S6B). By setting the machine learning probability cut-off at 0.86, the model achieved a sensitivity of 95% and specificity of 100% (Figure 3B). This indicates that the classifier using methylation values outperforms analysis solely based on copy-number (Figure 3C) with AUCs of 0.96 vs. 0.70 (Figure S6A). The gradient boosted model correctly identified 17 out of 19 cases (Figure 3D) while genetic analysis identified only 12 out of 19 cases (Figure S6C).

Performance of the machine learning model in detecting early cases of prostate cancer

We have demonstrated an ability to discriminate metastatic and cancer-free patients from a prostate cancer diagnostic pathway. We proceeded to evaluate the performance of our cfMBD-seq approach in detecting early cases of prostate cancer. We hypothesized that molecular features harbored in metastatic prostate cancer cases would also be present (albeit at lower amounts) in non-metastatic disease. Therefore,

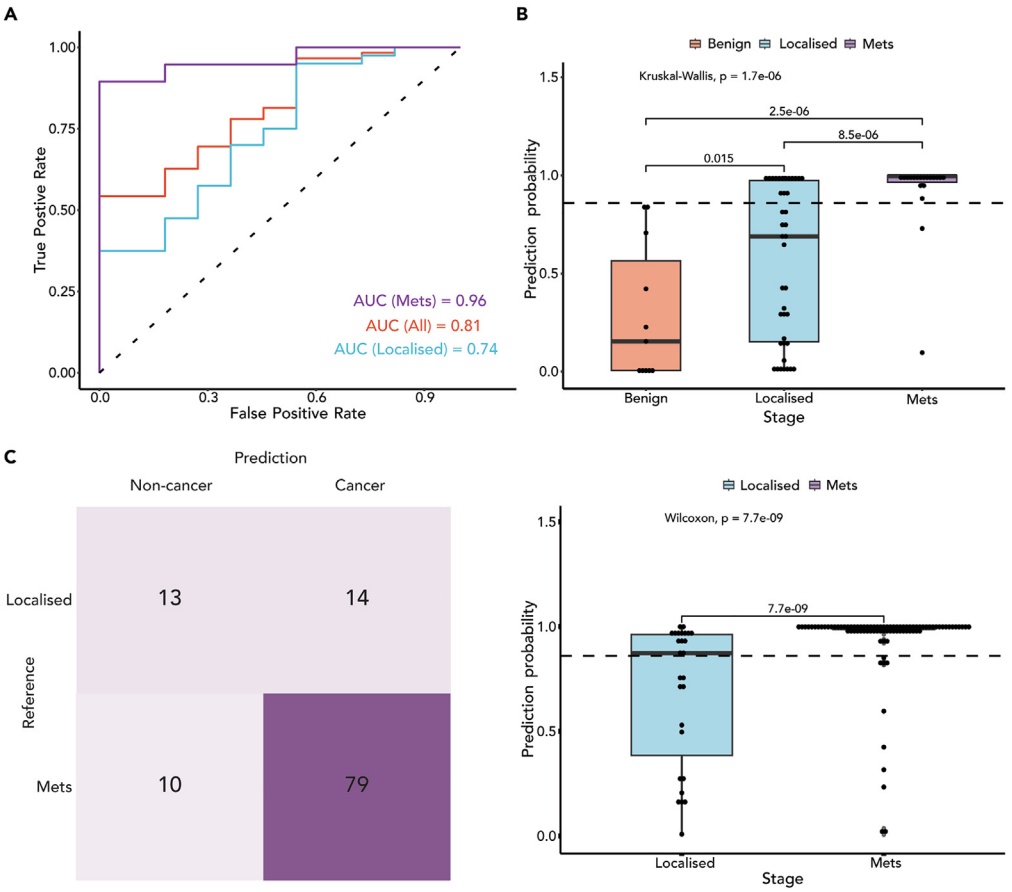


Figure 4. Performance of the machine learning model in the detection of localized prostate cancer

(A) Comparison of the ROC curves generated from the machine learning model in detecting metastatic, localized, and all cancers together.

(B) Boxplot displaying the distribution of prediction scores for controls (n = 11), localized (n = 40), and metastatic cases (n = 19).

(C) Performance of the machine learning model (left) and the boxplots showing the distribution of scores (right) in detecting prostate cancer cases in an external cfMeDIP-seq dataset from Chen et al., 2022[16] consisting of localized (n = 27) and metastatic (n = 89) prostate cancer samples.

we utilised the same set of DMRs, and thresholds as defined previously. We analyzed plasma samples from 40 patients with localized prostate cancer and achieved an AUC of 0.74, for discriminating such cases from cancer-free controls (Figure 4A). Furthermore, we observed an increasing probability score using the gradient boosted machine learning model when comparing normal, localized, and metastatic cases (Figure 4B), suggesting the potential to adjust the threshold values for early cancer cases. When we segregated the cohort of 40 non-metastatic cancer cases into clinically defined good and poor prognosis, based on the UK National Institute of Care and Excellence (NICE): Cambridge Prognostic Group (CPG) scores ("1" being very good prognosis and "5" being poor prognosis),[26] we observed similar performance in sensitivity (Figures S7A and S7B). It is possible that further improvement in performance could be achieved by training the model specifically on localized cases, although the available number of cases was insufficient for this analysis. In order to validate our model, we locked in all the parameters used in the machine learning model and tested it on a previously published dataset that used cell-free MeDIP-seq in differentiating localized (n = 30) from metastatic prostate cancers (n = 103). Despite the difference in the method used to enrich for methylated fragments (MeDIP vs. MBD), and after removing samples with limited coverage on the 900 chosen DMRs, we are able to generate similar sensitivity of detecting about 51% (14 out of 27) of the localized and 89% (79 out of 89) of the metastatic PrCa samples (Figure 4C). We were unable to test the specificity of the method due to the lack of benign samples in the external cohort and combining datasets from multiple sources may introduce unwanted biases and inter-experimental variabilities that may affect the validation exercise.

Functional annotations of the methylation alterations used to discriminate cancer from non-cancer cases

While these 900 DMRs (Table S2) were obtained from an uninformed methylation capture method, the regions were nevertheless successful in distinguishing cancer from non-cancer cases. As such, we sought to evaluate the biological significance of these epigenetic differences.

We observed that these DMRs are distributed fairly evenly across the genome, with no apparent gross organization of hyper or hypo-methylated associations (Figure 5A). We observed at chromosome 8, the p-arm being mostly hypomethylated while the q-arm is mostly

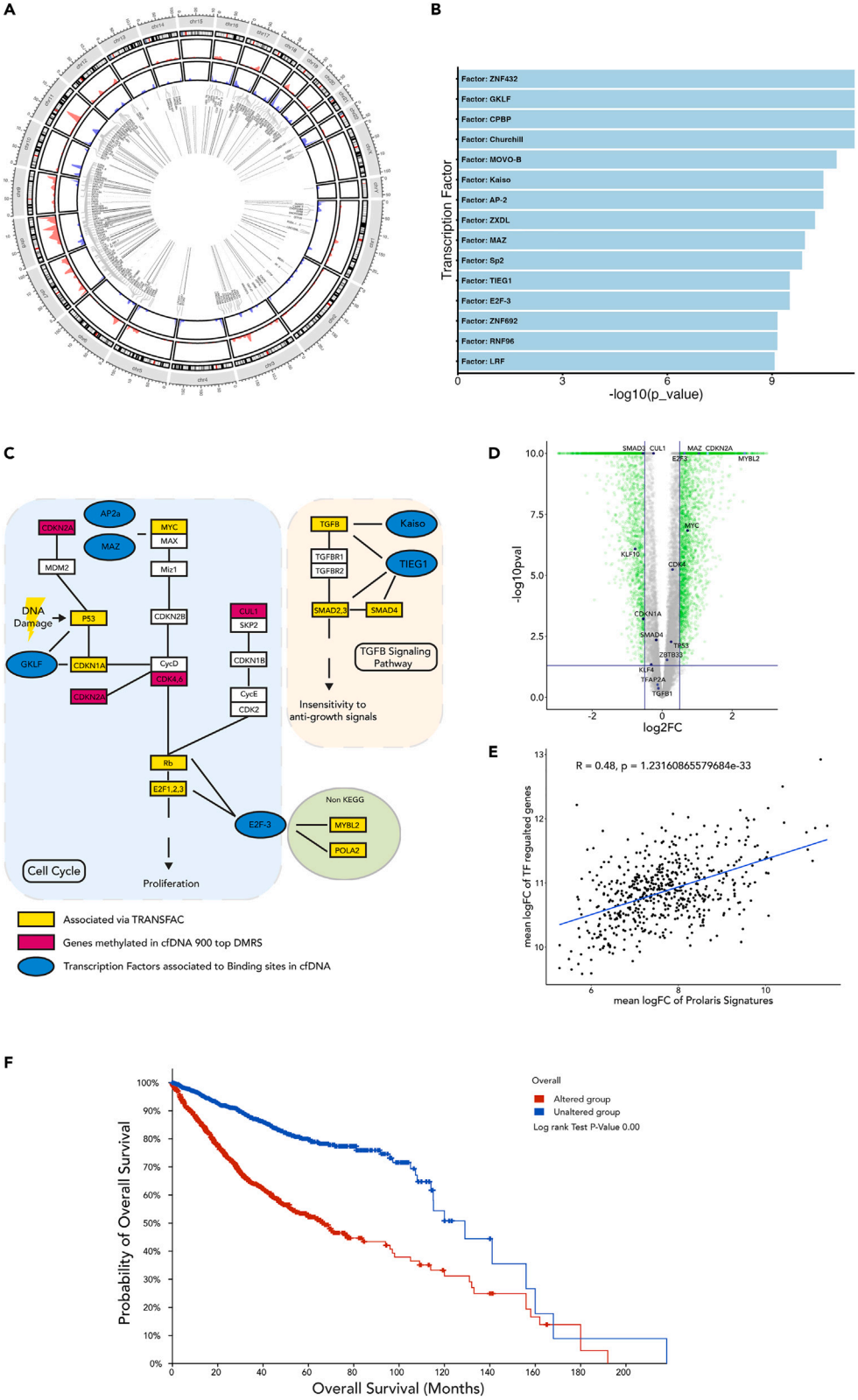

A
B
Factor: ZNF432
Factor: GKLF
Factor: CPBP
Factor: Churchill
Factor: MOVO-B
Factor: Kaiso
Factor: AP-2
Factor: ZXDL
Factor: MAZ
Factor: Sp2
Factor: TIEG1
Factor: E2F-3
Factor: ZNF692
Factor: RNF96
Factor: LRF
Transcription Factor
-log10(p_value)
C
AP2a
CDKN2A
MAZ
MYC
MAX
MDM2
Miz1
DNA Damage
P53
CDKN2B
CUL1
SKP2
GKLF
CDKN1A
CycD
CDK4,6
CDKN1B
CDKN2A
CycE
CDK2
Rb
E2F1,2,3
E2F-3
Non KEGG
MYBL2
POLA2
Cell Cycle
Proliferation
TGFB
Kaiso
TGFBR1
TGFBR2
TIEG1
SMAD2,3
SMAD4
TGFB Signaling Pathway
Insensitivity to anti-growth signals
Associated via TRANSFAC
Genes methylated in cfDNA 900 top DMRS
Transcription Factors associated to Binding sites in cfDNA
D
-log10pval
log2FC
E
R = 0.48, p = 1.23160865579684e-33
mean logFC of TF regualted genes
mean logFC of Prolaris Signatures
F
Overall
Altered group
Unaltered group
Log rank Test P-Value 0.00
Probability of Overall Survival
Overall Survival (Months)

Figure 5. Functional annotation of the epigenetic neighborhood revealed by the machine learning model for cfDNA methylation analysis

(A) Circos plot illustrating the distribution of hypermethylated (red, outer) and hypomethylated (blue, inner) DMRs across the genome with respect to the tumor samples. Each line within the innermost circle represents the associated gene of the DMRs.

(B) Top transcription factors that regulate the genes associated with the 900 DMRs used as input for the machine learning model. Data generated using the g:Profiler web tool using the associations of transcription factors and their downstream genes from the TRANSFAC database,[29] and showing their respective adjusted enrichment *p*-values.

(C) Section of the KEGG pathway "Pathways in Cancer" highlighting the transcription factors (identified in panel **B**, in blue) and their downstream targets (identified separately from the TRANSFAC public database, shown in yellow) demonstrating enrichment for cancer-relevant signaling pathways. Genes that are associated by TRANSFAC outside the KEGG pathway indicated are marked in green.

(D) Volcano plot displaying the fold-change and significance of gene expression differences between normal and tumor tissue samples (from TCGA dataset). The genes identified in the previous panel are labeled (log2FC and *p*-value cutoffs are indicated in blue lines).

(E) Correlation analysis between the gene expression data of the upregulated genes in our plasma-based 'enriched gene set' and the expression of the Prolaris gene set from TCGA tumor tissues, which are known to be associated with the risk of cancer progression.

(F) Kaplan-Meier survival plot showing the difference in overall survival between individuals with genetic or epigenetic alterations in the genes associated with the DMRs, and those without any alterations.

hypermethylated. Chromosome 8p deletions[27] and 8q gains[28] are common to prostate cancer, hence we examined whether methylation levels were related to a loss or gain in copy number (which could translate into hypomethylated and hypermethylated regions, respectively in MBD-seq) rather than methylation level changes per se. We did not observe a clear relationship between the distribution of CpG methylations and ploidy (Figure S8).

We next performed functional enrichment analysis of the genetic neighborhood surrounding our 900 DMRs using g:Profiler,[29] which revealed the enrichment of transcription factor binding sites within this dataset (Figure 5B). Considering the relevance of these transcription factors to prostate cancer, we further investigated their involvement in gene regulatory pathways. Our findings indicated that these transcription factors and their associated genes are predominantly associated with two major pathways: cell proliferation and TGF-beta signaling (Figure 5C). This observation is consistent with existing literature, as sustained cell proliferation is a well-known hallmark of cancer,[30] and dysregulation of the TGF-beta signaling pathway has been implicated in the prostate cancer progression.[31] Further analysis highlights that many of these genes represent G1/S cell cycle checkpoint signature genes (Figure S9A).

Since the DMRs we identified and evaluated were obtained from plasma samples, we sought to investigate the concordance of these regions with large independent tissue-based data from TCGA. We found that the transcription factor binding sites enriched in these TCGA-derived DMRs from compared to that of the entire 450k array panel (for which TCGA methylation data was obtained from) are similar to those enriched in the 900 DMRs we identified in the plasma (Figure S9B). Then, using the RNA expression of the TCGA dataset, we show that many of the genes associated with the transcription factors enriched for in the plasma (referred to as "enriched gene set") were also differentially expressed in the tissue. (Figure 5D; Figure S9C, and Table S3). This plasma-based "enriched gene set" are also compared with the expression of the 'Prolaris' genes in the tumor tissues in the TCGA database, which consists of 31 genes associated with cellular proliferation, and which may be relevant to clinical 'aggressiveness' of prostate cancer.[32] We reasoned that proliferation-linked genes found in cancers, correlate with tumor burden and would be more likely to correlate with the abundance of tumour-associated methylation alterations. Indeed, we found a correlation between the expression of our "enriched gene set" and the "Prolaris genes" (Figure 5E; Figure S9D) despite no direct overlap between the specific genes in both sets.

Finally, we evaluated the impact of our plasma-based gene set on the survival of patients with prostate cancer in the TCGA cohort. The presence of genetic and/or epigenetic alterations in our unique gene set, was associated with a poorer prognosis (Figure 5F). To eliminate the confounding effects of well-established cancer drivers *TP53* and *MYC*, we repeated the analysis by excluding these two genes, and retained this strong relationship between the altered gene set and survival (Figure S9E).

DISCUSSION

We conducted an evaluation of two common methylation enrichment strategies and found that MBD-seq is a reliable method for capturing informative reads that are relevant to prostate cancer. Despite the preference of these techniques for CpG regions with higher methylation density, we were still able to generate a valuable set of DMRs that can be used for prostate cancer detection across disease stages, within a clinically relevant diagnostic setting.

Our study builds upon prior work which identified metastatic prostate cancer tissue mutations from plasma samples containing sufficient cfDNA.[7] In a similar work aimed at detecting mutations in ctDNA on other types of cancers, Wan and colleagues first defined a patient-specific panel of mutations from tissue biopsies and then employed a computational enrichment method. Their method was able to detect up to 10^{-5} - 10^{-6} mutant molecule fraction for samples with $>10^5$ informative reads, which is defined as the product of the coverage and the number of mutations in the panel.[33] Our work employs a similar strategy of increasing informative reads by simultaneously analyzing multiple DMRs, but we obviate the need for a patient-specific background because of the high degree of shared methylation patterns across patients.[10] Herein, we have demonstrated superior performance than other methylation-based alternatives for prostate cancer identification (Table S4).

The multi-modal analysis of cfDNA offers a wealth potential information that can be captured for clinical utility.[34,35] In our study, we demonstrated the capacity to estimate tumor fraction from copy number aberration analysis in cfDNA obtained through our capture-based

approach. Using tumor fraction information, we identified tumor samples with sufficient content to train a machine learning model that can identify unique DNA methylation patterns in cfDNA, distinguishing metastatic prostate cancer cases from cancer-free subjects. This approach proved effective in distinguishing metastatic cases with varying tumor fraction burdens and revealing a higher sensitivity detection afforded by methylation alteration analysis. Significantly, we also observed the extrapolation of our method to the detection of localized cancer cases. By relying on the same features defined in advanced cancer, we were able to detect a substantial proportion of localized cases.

The functional analysis we conducted revealed the cancer relevance of the epigenetic neighborhoods identified in our study. These neighborhoods were found to be associated with pathways closely linked to prostate cancer progression and survival, including cell proliferation and TGF-beta signaling. The identification of TGF-beta signaling pathway is intriguing, given its crucial role in various oncogenic events including increased proliferation, decreased apoptosis, epithelial-to-mesenchymal transition, and evasion of immune surveillance.[36] Targeting this pathway may hold promise as a therapeutic strategy for castration-resistant prostate cancer.[36]

In addition, several of the transcription factors identified in our analysis have been strongly linked to prostate cancer. For example, *GKLF* (also known as *KLF4*), is a known regulator of prostate stem cell homeostasis, and its overexpression has been correlated with prostate tumorigenesis.[37] Similarly, the transcription factor Kaiso has been implicated in the progression of prostate cancer,[38] while *MAZ* has been shown to promote prostate cancer metastasis to the bone.[39] Despite the fact that the highlighted DMRs were obtained from cfDNA using a non-locus-specific enrichment method, we were able to validate and confirm these findings through comparison with tissue methylation and transcriptomic data. This represents compelling evidence for the biological significance of these specific regions that we have highlighted and supports their application in prostate cancer diagnostics.

Our findings indicate that there are common features shared by metastatic cancers that are also present in the earliest stage of the disease. Understanding the origins of these epigenetic changes and determining the relative contribution of tumor epithelia versus microenvironmental features will be crucial for advancing our knowledge in this area. It is noteworthy, that our machine learning model showed similar performance in distinguishing between patients with good (CPG1) and poor (CPG5) prognosis (Figure S7B). This suggests that the captured DMRs may be more reflective of tumor burden, which are comparable among localized cancers and significantly greater in metastatic disease (Figure S10). It is also important to note the heterogeneity of each prognostic subgroup, and the nearly 70% cancer-specific survival of CPG5s, reflecting the variability of outcome.[26] Furthermore, these DMRs were originally identified in high tumor burden-metastatic cancers. Therefore, the generated probability score may reflect an epigenetic bottleneck specific to the metastatic process, suggesting that the score will only increase as the potential to metastasise increases.

Limitations of the study

The performance of the model on localized cancers was limited; however, the identification of metastasis linked DMRs may hold prognostic relevance for localized cases. Future validation on larger cohorts with well-defined outcomes is also required to test this outcome since the samples in our cohort lack the needed clinical follow-up data to make the necessary correlations. Further validation and follow-up studies are also required to determine appropriate applications of the tool. In addition, a targeted capture method on the informative DMRs may yield greater coverage depths that allow for more sensitive detection of localized cancers and improve the differentiation of low- and high-risk cancers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Study participants
- METHOD DETAILS
  - Sample collection and preparation
  - Cell-free DNA extraction from cell culture media
  - Enzymatic conversion, library preparation, and target capture (cfEM-seq)
  - cfMEDIP-seq and cfMBD-seq
  - Sequencing data processing
  - Sample quality control for cfMeDIP-seq and cfMBD-seq
  - Assessing tumour fraction using ichorCNA
  - Selection of the differentially methylated regions
  - Functional analysis of DMRs
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110330.

ACKNOWLEDGMENTS

We would like to dedicate this article to the memory of Charlie Massie, co-lead author of the study, and colleague, mentor, and friend to his co-authors. We were fortunate to work with such a dedicated, passionate, and talented scientist, and his memory will continue to inspire us to make a positive impact in all that we do.

The authors would also like to acknowledge the support and assistance of: Alex Keates and Kelly Leonard of the CU-TRACT office, Dr Anne Warren, Dr Simon Pacey, Dr Shubha Anand and the Cancer Molecular Diagnostics Laboratory team, Professor Grant Stewart, Professor Ian Mills, and our lab members, for their valuable contributions, support, and feedback in the implementation of this project and preparation of this article. We are grateful to the Cambridge Cancer Center, and its core facilities. Thanks to our director, Professor Rebecca Fitzgerald, and all our colleagues at the Early Cancer Institute. We extend our sincerest thanks to the patients and their families, who generously donated their samples for our research, and without whom, none of this would have been possible. We are grateful for the support and collaboration of the CRUK Prostate International Cancer Genome Consortium. This research was made possible through funding from various sources; Cancer Research UK, CRUK Career Development Fellowship, University of Cambridge W.D. Armstrong Trust Fund, John Black Prostate Cancer Foundation Young Investigator Award. This research was also supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

AUTHOR CONTRIBUTIONS

EL, AF, and CEM developed the concept of the project. EL, RL, and CEM planned the experimental design. EL performed the wet lab experiments. TMC wrote the code and performed the machine learning. TMC and HLY performed the function annotations. RNA sequencing data pre-processing and data analysis was performed by TMC and supported by AGL. HWM performed the tumor fraction analysis. TMC, EL, HLY, HWM, and HD analyzed and interpreted the data, and wrote the article with input from all the others; VJG provided and coordinated for the collection of the samples and their clinical and prognostic annotation. EL, TMC, HWM, RH, SHR, RL, CEM, HLY, and HD provided technical and scientific inputs throughout the project. TMC, HLY, ALR, MG, and HD analyzed the external dataset. CEM and HD supervised the project.

DECLARATION OF INTERESTS

The authors do not claim any competing interest in this work.

Received: August 4, 2023
Revised: May 2, 2024
Accepted: June 18, 2024
Published: June 20, 2024

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
2. Ilic, D., Djulbegovic, M., Jung, J.H., Hwang, E.C., Zhou, Q., Cleves, A., Agoritsas, T., and Dahm, P. (2018). Prostate cancer screening with prostate-specific antigen (PSA) test: A systematic review and meta-analysis. BMJ *362*, k3519. https://doi.org/10.1136/bmj.k3519.
3. Grossman, D.C., Curry, S.J., Owens, D.K., Bibbins-Domingo, K., Caughey, A.B., Davidson, K.W., Doubeni, C.A., Ebell, M., Epling, J.W., Kemper, A.R., et al. (2018). Screening for prostate cancer US: Preventive services task force recommendation statement. JAMA, J. Am. Med. Assoc. *319*, 1901–1913. https://doi.org/10.1001/jama.2018.3710.
4. Ström, P., Nordström, T., Aly, M., Egevad, L., Grönberg, H., and Eklund, M. (2018). The Stockholm-3 Model for Prostate Cancer Detection: Algorithm Update, Biomarker Contribution, and Reflex Test Potential. Eur. Urol. *74*, 204–210. https://doi.org/10.1016/j.eururo.2017.12.028.
5. Cooper, C.S., Eeles, R., Wedge, D.C., Van Loo, P., Gundem, G., Alexandrov, L.B., Kremeyer, B., Butler, A., Lynch, A.G., Camacho, N., et al. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat. Genet. *47*, 367–372. https://doi.org/10.1038/ng.3221.
6. Wedge, D.C., Gundem, G., Mitchell, T., Woodcock, D.J., Martincorena, I., Ghori, M., Zamora, J., Butler, A., Whitaker, H., Kote-Jarai, Z., et al. (2018). Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nat. Genet. *50*, 682–692. https://doi.org/10.1038/s41588-018-0086-z.
7. Wyatt, A.W., Annala, M., Aggarwal, R., Beja, K., Feng, F., Youngren, J., Foye, A., Lloyd, P., Nykter, M., Beer, T.M., et al. (2017). Concordance of Circulating Tumor DNA and Matched Metastatic Tissue Biopsy in Prostate Cancer. J. Natl. Cancer Inst. *109*, djx118. https://doi.org/10.1093/jnci/djx118.
8. Fraser, M., Sabelnykova, V.Y., Yamaguchi, T.N., Heisler, L.E., Livingstone, J., Huang, V., Shiah, Y.J., Yousif, F., Lin, X., Masella, A.P., et al. (2017). Genomic hallmarks of localized, non-indolent prostate cancer. Nature *541*, 359–364. https://doi.org/10.1038/nature20788.
9. Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–689. https://doi.org/10.1038/ng.2279.
10. Massie, C.E., Spiteri, I., Ross-Adams, H., Luxton, H., Kay, J., Whitaker, H.C., Dunning, M.J., Lamb, A.D., Ramos-Montoya, A., Brewer, D.S., et al. (2015). HES5 silencing is an early and recurrent change in prostate tumourigenesis. Endocr. Relat. Cancer *22*, 131–144. https://doi.org/10.1530/ERC-14-0454.

11. Jerónimo, C., Bastian, P.J., Bjartell, A., Carbone, G.M., Catto, J.W.F., Clark, S.J., Henrique, R., Nelson, W.G., and Shariat, S.F. (2011). Epigenetics in prostate cancer: Biologic and clinical relevance. https://doi.org/10.1016/j.eururo.2011.06.035.
12. Jones, K., Zhang, Y., Kong, Y., Farah, E., Wang, R., Li, C., Wang, X., Zhang, Z., Wang, J., Mao, F., et al. (2021). Epigenetics in prostate cancer treatment. J. Transl. Genet. Genom. *5*, 341–356. https://doi.org/10.20517/jtgg.2021.19.
13. Kurdyukov, S., and Bullock, M. (2016). DNA methylation analysis: Choosing the right method (MDPI AG). https://doi.org/10.3390/biology5010003.
14. Lan, X., Adams, C., Landers, M., Dudas, M., Krissinger, D., Marnellos, G., Bonneville, R., Xu, M., Wang, J., Huang, T.H.M., et al. (2011). High resolution detection and analysis of CpG dinucleotides methylation using MBD-seq technology. PLoS One *6*, e22226. https://doi.org/10.1371/journal.pone.0022226.
15. Bryzgunova, O., Bondar, A., Ruzankin, P., Laktionov, P., Tarasenko, A., Kurilshikov, A., Epifanov, R., Zaripov, M., Kabilov, M., and Laktionov, P. (2021). Locus-specific methylation of gstp1, rnf219, and kiaa1539 genes with single molecule resolution in cell-free dna from healthy donors and prostate tumor patients: Application in diagnostics. Cancers *13*, 6234. https://doi.org/10.3390/cancers13246234.
16. Chen, S., Petricca, J., Ye, W., Guan, J., Zeng, Y., Cheng, N., Gong, L., Shen, S.Y., Hua, J.T., Crumbaker, M., et al. (2022). The cell-free DNA methylome captures distinctions between localized and metastatic prostate tumors. Nat. Commun. *13*, 6467. https://doi.org/10.1038/s41467-022-34012-2.
17. Klein, E.A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., Chung, G., Clement, J., Gao, J., Hunkapiller, N., et al. (2021). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann. Oncol. *32*, 1167–1177. https://doi.org/10.1016/j.annonc.2021.05.806.
18. Shen, S.Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M.H.A., Chadwick, D., Zuzarte, P.C., Borgida, A., Wang, T.T., Li, T., et al. (2018). Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature *563*, 579–583. https://doi.org/10.1038/s41586-018-0703-0.
19. Huang, J., Soupir, A.C., and Wang, L. (2022). Cell-free DNA methylome profiling by MBD-seq with ultra-low input. Epigenetics *17*, 239–252. https://doi.org/10.1080/15592294.2021.1896984.
20. Lienhard, M., Grasse, S., Rolff, J., Frese, S., Schirmer, U., Becker, M., Börno, S., Timmermann, B., Chavez, L., Sültmann, H., et al. (2017). QSEA-modelling of genome-wide DNA methylation from sequencing enrichment experiments. Nucleic Acids Res. *45*, e44. https://doi.org/10.1093/nar/gkw1193.
21. Beltran, H., Romanel, A., Conteduca, V., Casiraghi, N., Sigouros, M., Franceschini, G.M., Orlando, F., Fedrizzi, T., Ku, S.Y., Dann, E., et al. (2020). Circulating tumor DNA profile recognizes transformation to castration-resistant neuroendocrine prostate cancer. J. Clin. Invest. *130*, 1653–1668. https://doi.org/10.1172/JCI131041.
22. Adalsteinsson, V.A., Ha, G., Freeman, S.S., Choudhury, A.D., Stover, D.G., Parsons, H.A., Gydush, G., Reed, S.C., Rotem, D., Rhoades, J., et al. (2017). Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat. Commun. *8*, 1324. https://doi.org/10.1038/s41467-017-00965-y.
23. Wallander, K., Eisfeldt, J., Lindblad, M., Nilsson, D., Billiau, K., Foroughi, H., Nordenskjöld, M., Liedén, A., and Tham, E. (2021). Cell-free tumour DNA analysis detects copy number alterations in gastro-oesophageal cancer patients. PLoS One *16*, e0245488. https://doi.org/10.1371/journal.pone.0245488.
24. Berchuck, J.E., Baca, S.C., McClure, H.M., Korthauer, K., Tsai, H.K., Nuzzo, P.V., Kelleher, K.M., He, M., Steinharter, J.A., Zacharia, S., et al. (2022). Detecting Neuroendocrine Prostate Cancer Through Tissue-Informed Cell-Free DNA Methylation Analysis. Clin. Cancer Res. *28*, 928–938. https://doi.org/10.1158/1078-0432.CCR-21-3762.
25. Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome (Nature Publishing Group). https://doi.org/10.1038/nrg3871.
26. Gnanapragasam, V.J., Bratt, O., Muir, K., Lee, L.S., Huang, H.H., Stattin, P., and Lophatananon, A. (2018). The Cambridge Prognostic Groups for improved prediction of disease mortality at diagnosis in primary non-metastatic prostate cancer: A validation study. BMC Med. *16*, 31. https://doi.org/10.1186/s12916-018-1019-5.
27. Macoska, J.A., Trybus, T.M., Benson, P.D., Sakr, W.A., Grignon, D.J., Wojno, K.D., Pietruk, T., and Powell, I.J. (1995). Evidence for Three Tumor Suppressor Gene Loci on Chromosome 8p in Human Prostate Cancer. Cancer Res. *55*, 5390–5395.
28. Steiner, T., Junker, K., Burkhardt, F., Braunsdorf, A., Janitzky, V., and Schubert, J. (2002). Gain in Chromosome 8q Correlates with Early Progression in Hormonal Treated Prostate Cancer. Eur. Urol. *41*, 167–171.
29. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). G:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. *47*, W191–W198. https://doi.org/10.1093/nar/gkz369.
30. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. https://doi.org/10.1016/j.cell.2011.02.013.
31. Thompson-Elliott, B., Johnson, R., and Khan, S.A. (2021). Alterations in TGFβ signaling during prostate cancer progression. Am. J. Clin. Exp. Urol. *9*, 318.
32. Cuzick, J., Fisher, G., Mstat, R., Ba, W., Park, J., Bs, Y., Flake, D.D., Wagner, S., Gutin, A., Lanchbury, J.S., et al. (2011). Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. Lancet Oncol. *12*, 245–255. https://doi.org/10.1016/S1470.
33. Wan, J.C.M., Heider, K., Gale, D., Murphy, S., Fisher, E., Mouliere, F., Ruiz-Valdepenas, A., Santonja, A., Morris, J., Chandrananda, D., et al. (2020). ctDNA Monitoring Using Patient-specific Sequencing and Integration of Variant Reads. Sci. Transl. Med. *12*, eaaz8084.
34. Snyder, M.W., Kircher, M., Hill, A.J., Daza, R.M., and Shendure, J. (2016). Cell-free DNA Comprises an in Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell *164*, 57–68. https://doi.org/10.1016/j.cell.2015.11.050.
35. Erger, F., Nörling, D., Borchert, D., Leenen, E., Habbig, S., Wiesener, M.S., Bartram, M.P., Wenzel, A., Becker, C., Toliat, M.R., et al. (2020). CfNOMe - A single assay for comprehensive epigenetic analyses of cell-free DNA. Genome Med. *12*, 54. https://doi.org/10.1186/s13073-020-00750-5.
36. Cao, Z., and Kyprianou, N. (2015). Mechanisms navigating the TGF-B pathway in prostate cancer. Asian J. Urol. *2*, 11–18. https://doi.org/10.1016/j.ajur.2015.04.011.
37. Xiong, X., Schober, M., Tassone, E., Khodadadi-Jamayran, A., Sastre-Perona, A., Zhou, H., Tsirigos, A., Shen, S., Chang, M., Melamed, J., et al. (2018). KLF4, A Gene Regulating Prostate Stem Cell Homeostasis, Is a Barrier to Malignant Progression and Predictor of Good Prognosis in Prostate Cancer. Cell Rep. *25*, 3006–3020.e7. https://doi.org/10.1016/j.celrep.2018.11.065.
38. Wang, H., Liu, W., Black, S., Turner, O., Daniel, J.M., Dean-Colomb, W., He, Q.P., Davis, M., Yates, C., and Kaiso. (2015). a Transcriptional Repressor, Promotes Cell Migration and Invasion of Prostate Cancer Cells through Regulation of miR-31 Expression. *7*, 5677.
39. Yang, Q., Lang, C., Wu, Z., Dai, Y., He, S., Guo, W., Huang, S., Du, H., Ren, D., and Peng, X. (2019). MAZ promotes prostate cancer bone metastasis through transcriptionally activating the KRas-dependent RalGEFs pathway. J. Exp. Clin. Cancer Res. *38*, 391. https://doi.org/10.1186/s13046-019-1374-x.
40. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.
41. Li, H. (2013). Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.
42. Krueger, F., and Andrews, S.R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572. https://doi.org/10.1093/bioinformatics/btr167.
43. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
44. Lienhard, M., Grimm, C., Morkel, M., Herwig, R., and Chavez, L. (2014). MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. Bioinformatics *30*, 284–286. https://doi.org/10.1093/bioinformatics/btt650.
45. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Preprint at arXiv. https://doi.org/10.48550/arXiv.1603.04467.
46. Cavalcante, R.G., and Sartor, M.A. (2017). Annotatr: Genomic regions in context.

Bioinformatics *33*, 2381–2383. https://doi.org/10.1093/bioinformatics/btx183.

47. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.
48. Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., and Peterson, H. (2023). G:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. *51*, W207–W212. https://doi.org/10.1093/nar/gkad347.
49. Kanehisa, M. (2000). Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. *28*, 27–30.
50. Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. (2003). TRANSFAC®: Transcriptional regulation, from patterns to profiles. https://doi.org/10.1093/nar/gkg108.
51. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.
52. Shen, S.Y., Burgener, J.M., Bratman, S.V., and De Carvalho, D.D. (2019). Preparation of cfMeDIP-seq libraries for methylome profiling of plasma cell-free DNA. Nat. Protoc. *14*, 2749–2780. https://doi.org/10.1038/nprot.201.012.
53. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Project, G., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human plasma samples	Cambridge University Hospitals	DIAMOND biomarker study
Critical commercial assays		
Qiasymphony DSP Circulating DNA mini kit	Qiagen	Cat no. 937556
QIAamp Circulating Nucleic Acid Kit	Qiagen	Cat no. 55114
EM-seq Conversion module	New England Biolabs	Cat no. E7125L
ACCEL-NGS Methyl-Seq DNA Library Kit	Swift	Cat no. 30024
SeqCap Epi Enrichment Kit	Roche NimbleGen	Cat no. 07145519001
KAPA HyperPrep Kit	Roche	KK8502
QIAquick PCR Purification Kit	Qiagen	Cat no. 28104
MagMeDIP kit	Diagenode	Cat no. C02010021
MethylMiner Kit	Thermo Fisher	Cat no. ME10025
Deposited data		
Processed data	This paper	https://github.com/MassieLab/cfMBD-seq-for-Prostate-cancer-detection/tree/main/data
MeDIP data of localised vs metastatic prostate cancer from the CPC cohorts	Chen et al.[16]	EGA dataset: http://ega-archive.org/datasets/EGAD00001007972
MeDIP data of localised vs metastatic prostate cancer from the VPC cohort	Chen et al.[16]	EGA dataset: http://ega-archive.org/datasets/EGAD00001008711
MeDIP data of localised vs metastatic prostate cancer from the Barrier cohort	Chen et al.[16]	EGA dataset: http://ega-archive.org/datasets/EGAD00001008712
Infinium Human Methylation 450 K array - prostate cancer	TCGA	https://portal.gdc.cancer.gov/projects/TCGA-PRAD
Experimental models: Cell lines		
PC3	ATCC	CRL - 1435; RRID: CVCL_0035
Software and algorithms		
FastQC v0.11.4	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC v1.11	Ewels, Philip et al.[40]	https://multiqc.info/
TrimGalore v0.4.4	Babraham Bioinformatics	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
R v4.2.2	The R Foundation	https://www.r-project.org/
BWA-mem	Li, H. et al.[41]	https://github.com/lh3/bwa
Bismark v0.22.1	Krueger, F. et al.[42]	https://github.com/FelixKrueger/Bismark
Samtools	Li, H. et al.[43]	https://github.com/samtools
MEDIPS v1.44.0	Lienhard, M. et al.[44]	https://bioconductor.org/packages/release/bioc/html/MEDIPS.html
ichorCNA	Adalsteinsson, V.A. et al.[22]	https://github.com/broadinstitute/ichorCNA
QSEA	Lienhard, Matthias et al.[20]	https://www.bioconductor.org/packages/release/bioc/html/qsea.html
Python v3.7	Python software foundation	https://www.python.org/
Tensorflow	Abadi, Martín et al.[45]	https://github.com/tensorflow/tensorflow

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
AnnotatR v1.12.1	Cavalcante, R et al.[46]	https://www.bioconductor.org/packages/release/bioc/html/annotatr.html
TxDb.Hsapiens.UCSC.hg38.knownGene v3.4.6	Bioconductor Core Team and Bioconductor Package Maintainer	https://www.bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg38.knownGene.html
Circlize	Gu, Z. et al.[47]	https://github.com/jokergoo/circlize
gprofiler	Kolberg, L. et al.[48]	https://biit.cs.ut.ee/gprofiler/gost
KEGG	Kanehisa, M. et al.[49]	https://www.genome.jp/kegg/
Transfac	Matys, V. et al.[50]	https://genexplain.com/transfac/
cBioportal	Cerami, Ethan et al.[51]	https://www.cbioportal.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests fir resources and reagents should be directed to and will be fulfilled by the lead contact, Dr Harveer Dev (hsd26@cam.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw sequencing data are available upon request from the lead contact. The processed data tables containing the beta values of all the samples at the 900 selected DMRs are available in https://github.com/MassieLab/cfMBD-seq-for-Prostate-cancer-detection/tree/main/data. The TCGA data and the validation dataset used are obtained from the TCGA and EGA websites respectively. The link to the dataset and the accession codes are listed in the key resources table.
- All codes used are deposited in https://github.com/MassieLab/cfMBD-seq-for-Prostate-cancer-detection.
- All additional information required to reanalyse the data in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Study participants

This study was conducted in accordance with all applicable ethical regulations. Access to samples and permission to use was covered by the DIAMOND biomarker study (Ethics 03/18, CI: VJG). All of the experiments conducted were compliant with relevant laws or guidelines of the University of Cambridge and were approved by the National Research Ethics Service (NRES) Committee East of England, UK. All participants were male, aged 47-91 years old, and predominantly British. The age, T-stage, and PSA at diagnosis are summarised in Table S1, including p-values from performing one-way ANOVA with a Tukey's post hoc test. The samples were split 70/30 into training and test sets respectively such that the age and PSA levels of the two groups are comparable with each other.

All peripheral blood samples were obtained from patients who underwent the prostate cancer diagnostic pathway and were recommended for biopsies. They were then classified as either normal/benign, non-metastatic good prognosis (CPG1), non-metastatic poor prognosis (CPG5), or metastatic at presentation, using the NICE CPG stratification system.

METHOD DETAILS

Sample collection and preparation

Plasma was extracted from the whole blood samples according to the biobank's internal protocol and were aliquoted in 2 – 4 mL fractions and were immediately stored at -80°C for long-term storage. The cell-free DNA was isolated from plasma using Qiasymphony DSP Circulating DNA mini kit (Qiagen) at the Cambridge Cancer Molecular Diagnostics Laboratory (CMDL) and quantified via Qubit 4.0 (Thermo Fisher Scientific) before use.

The external validation dataset was obtained from the CPC (n=30 localised samples), VPC (n=67 metastatic samples), and Barrier cohorts (n=14 metastatic samples) in the Chen et al. (2022)[16] paper.

Cell-free DNA extraction from cell culture media

Human prostate cancer cell line PC3 was obtained from the American Type Culture Collection (ATCC). Cells were cultured for 72h at 37°C in 5% CO_2 and were harvested at >90% confluency before cfDNA isolation. Cell culture media was first aspirated from the adherent PC3 cells, and then followed by double centrifugation, first at 1900 xg and subsequently at 16,000 xg for 10 min and 30 min, respectively. Following centrifugation, cfDNA was isolated using QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer's instructions. The isolated cfDNA was quantified using a Qubit 4.0 (Thermo Fisher Scientific) fluorometer.

Enzymatic conversion, library preparation, and target capture (cfEM-seq)

Libraries were prepared from 50 ng of isolated cfDNA from the supernatant of a growing PC3 cell culture by enzymatically converting unmethylated cytosines into uracil using the EM-seq Conversion module (New England BioLabs) and appending a truncated universal adapter onto the resulting methyl-converted single-stranded DNA products using the ACCEL-NGS Methyl-Seq DNA Library Kit (Swift Biosciences). Unique indices were incorporated to the adapter by subjecting them to 16 cycles of PCR using KAPA Hifi HotStart Uracil+ReadyMix (Roche) with primers from the Methyl-Seq - Dual Indexing Kit (Swift Biosciences), and then purifying the final DNA libraries with a 1.0X magnetic bead clean up step. The converted DNA libraries were then quantified using Qubit 4.0 Fluorometer (Thermo Fisher Scientific) and eight samples were pooled together at equimolar amounts for targeted probe hybridization capture. The probes were synthesised by NimbleGen based on the cancer-specific target regions that were described in the results section. They were then used to enrich for the chosen fragments in the converted libraries using the SeqCap Epi Enrichment Kit (NimbleGen) protocol. The product of which is then subjected for paired end sequencing at 150 bp on a NovaSeq 6000 system.

cfMEDIP-seq and cfMBD-seq

In order to enrich for methylated DNA in the plasma or cell-free supernatant samples, we either performed the cell-free Methylated DNA Immunoprecipitation (cfMeDIP) or cell-free methyl CpG binding domain protein (cfMBD) coupled with high throughput sequencing.

This cfMeDIP-seq protocol was adapted from a previously reported method by Shen et al.[52] First, 100 ng of cfDNA was end-repaired and A-tailed using KAPA HyperPrep Kit (Roche) according to the manufacturer's protocol. Then, these fragments were ligated with 480 nM of NEBNext unmethylated hairpin adaptors (NEBNext Multiplex Oligos for Illumina, New England Biolabs) at 20^{O}C for 20 min and were purified with AMPure XP beads (Beckman Coulter). After bead purification, the looped adaptors were opened through the addition of USER enzyme (New England Biolabs), which selectively cleaves the uracil nucleotide found in the middle of the hairpin loop. The product was then purified with QIAquick PCR Purification Kit (Qiagen) and was spiked with 0.3 ng of methylated and unmethylated *A. thaliana* DNA in the appropriate buffers from the MagMeDIP kit (Diagenode) to act as controls for the recovery of methylated DNA and background unmethylated DNA being enriched, respectively. The library mixture was then incubated at 95^{O}C for 10 min to generate single-stranded cfDNA libraries and then quickly transferred on ice for 10 min. The samples were then divided into two PCR tubes containing 7.9 uL and 87 uL aliquots for the 10% input control (IC) and immunoprecipitation (IP), respectively. The 5mC antibody provided in the MagMeDIP kit was diluted 15-fold prior to addition to the immunoprecipitation reactions and incubated for 17 h at 4^{O}C with rotation. The samples were then purified using Diagenode iPure Kit v2 kit as instructed by the kit without performing any fragment size selection afterwards. The eluted libraries were amplified for 9 cycles for IC and 12 cycles for IP libraries using Kapa HiFi Hotstart Mastermix (Roche) and 0.3 uM of NEBNext multiplex oligos (New England Biolabs). The success of the enrichment by the antibody was determined by qPCR amplification of the spiked-in unmethylated and methylated *A. thaliana* DNA (Diagenode). After the enrichment step, the eluted libraries were amplified using reagents from Kapa HyperPrep Kit (Roche), purified with magnetic bead clean up, and were sequenced at 150 bp on a NovaSeq 6000 system.

For the MBD-seq, 200 ng of cfDNA was used for the cell line experiments, while 5 ng was used for plasma cfDNA samples. Instructions for <1 ug of input cfDNA from the MethylMiner Kit (Thermofisher Scientific) protocol were followed in both cases. Briefly, 10 uL of pre-washed beads in 1X buffer were coupled with 3.5 ug of biotinylated MBD2 protein and incubated at room temperature for 1 h on an SLA-ROM-5 rotating mixture according to the vendor's supplied protocol. The washed bead-biotin/MBD2 complex was then added to each sample. Then, 200 ul of the 1X Bind/Wash buffer was added, and the solution was incubated on an SLA-ROM-5 rotator (Scitech Labapp) for 1h at room temperature. After mixing the cfDNA-MBD2/biotin-bead complex, the mixture was placed on a magnetic rack to separate the beads from the supernatant. The beads were then washed 3 times with 200 uL of 1X Bind/Wash buffer followed by 3 min incubations at room temperature on a rotator. The bound methylated DNA molecules were eluted from the beads with 2000 mM twice, each with 200 uL. The combined 400 ul of methylated DNA eluate was then purified by ethanol precipitation. Then, the sequencing library was prepared from the purified DNA using KAPA HyperPrep Kit and sequenced in NovaSeq 6000 as described above.

Sequencing data processing

After sequencing, the raw reads were examined for their quality using FastQC version 0.11.4 and MultiQC v1.11[40] Raw reads were trimmed using Trim Galore version 0.4.4 using default settings for pair-end mode. The trimmed reads were then aligned to hg38 using BWA-mem[41] using pair-end mode defaults. For the methyl-converted template, the trimmed reads were aligned to hg38 using the Bismark tools version 0.22.1. The final bam files were generated from SAM alignment files using SAMtools[53] version 1.9.

Sample quality control for cfMeDIP-seq and cfMBD-seq

For cfMeDIP-seq and cfMBD-seq, the sequencing results were subjected to quality control check by saturation analysis using the R Bioconductor package MEDIPS version 1.44.0.[44] All samples generated a reasonable number of unique reads (>17 million). Upon visual assessment of the saturation curves and background coverage analysis, all the samples indicated full saturation in terms of sequencing depth; hence, all the samples were included in the downstream analysis.

Assessing tumour fraction using ichorCNA

We ran the ichorCNA workflow on either shallow whole genome bisulfite sequencing data or on MBD-seq data, which involved first running readCounter from hmmcopy with window size set to 1MB and then generating a custom background using healthy samples as the normal panel. We then ran ichorCNA using the recommended settings to generate estimated tumour fractions.[22]

Selection of the differentially methylated regions

The genome is first binned into 100-bp non-overlapping windows. Then, each 100-bp region is classified as a differentially methylated region between metastatic prostate cancers (with higher than 15% tumour fraction based on ichorCNA) and control samples using the R package QSEA.[20] The details of choosing the samples are described in Figure 2. The resulting 100-bp DMRs were then merged with neighbouring regions if they were within 300 bp of each other to create larger DMRs and simplify analysis. Then, the DMRs were then used as input for the Gradient Boosted Decision Tree and Random Forest machine learning models, which were ran in python 3.7 using TensorFlow.[45] We thereafter used the Gradient Boosted Decision Tree model with the following hyperparamter values: num_trees=100, growing_strategy="-BEST_FIRST_GLOBAL", shrinkage = 0.1, Maximum depth = 6, min_examples = 5. We also evaluated the relationship between methylation level at the DMRs and copy number changes at each region. Using the sWGS ichorCNA copy number estimates generated while assessing the tumour fraction, we assigned a copy number to the DMR based on the 1Mb window it overlapped with.

Functional analysis of DMRs

The functional annotations of the DMRs were performed using the R package annotatR (v1.12.1),[46] TxDb.Hsapiens.UCSC.hg38.knownGene (v3.4.6). Circos plot in Figure 5 was generated using the R Bioconductor package Circlize.[47] Enhancers were annotated using the GeneHancer database. Gene enrichment analysis was performed using gprofiler2 (cut-off Benjamini-Hochberg adjusted p-value = 0.05)[29] by taking the gene annotations of the DMRs using annotatR as input and setting the background to default for MBD-derived DMRs and 450k array genes for TCGA-derived DMRs. Gene signalling and regulatory network pathways are derived from KEGG[49] and TRANSFAC[50] databases. Kaplan-Meier curves were generated from cBioportal.[51]

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analysis performed in R and are indicated in the figure legends. For comparisons of methylation levels, the data are shown as boxplots, and Kruskal-Wallis test was performed to determine whether the difference between groups were significant. In addition, asterisks are added to show statistical significance as: *p<0.05, **p<0.01, ***p<0.001, and ****p<0.0001, **ns** not significant. For comparisons of methylation levels at a CpG-level between cfMedIP-seq and cfMBD-seq, Spearman's rank correlation coefficients were reported.

The sample size (n) represents the number of patient samples used in the cohort. Specifically, for the training cohort: benign (n=24), metastatic (n=44); for the test cohort: benign (n=11), localised (n=40), metastatic (n=19); for the validation cohort: localised (n=27), metastatic (n=89). The clinical demographics were compared by performing one-way ANOVA with a Tukey's post hoc test. For gene annotations, we reported p-values for the transcription factors generated from gProfiler.[48] For the linear correlation between log fold changes between transcription factors, spearman correlation coefficients and p-values were reported. For the Kaplan-Meier curve, the log-rank test p-value from cBioPortal[51] is reported.