PROSTATE CANCER EPIGENETIC MECHANISM STUDY AND BIOMARKER DISCOVERY USING BIOINFORMATICS APPROACHES

A Dissertation

by

MUTIAN ZHANG

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Yun Huang
Committee Members,	Roderick Dashwood
	Chad Creighton
	Ken Chen
Head of Program,	Carol Vargas-Bautista

August 2021

Major Subject: Medical Sciences

ABSTRACT

Most screening-detected prostate cancer (PCa) is indolent and not lethal. Biomarkers that can predict aggressive diseases independent of clinical features are needed to improve risk stratification of localized PCa patients and reduce overtreatment. Epigenetic, especially methylation biomarkers have better stability in biofluids or samples with a below-average quality. We aimed to identify DNA methylation differences in leukocytes between clinically defined aggressive and non-aggressive PCa to identify potential biomarkers for PCa diagnosis. To accomplish this aim, we performed DNA methylation profiling in leukocyte DNA samples obtained from 287 PCa patients with Gleason Score (GS) 6 and ≥8 using Illumina 450k methylation arrays, and 8 PCa patients using whole genome bisulfite sequencing. We observed the DNA methylation level in the core promoters and the first exon region were significantly higher in GS≥8 patients than GS=6 PCa. We then performed a 5-fold cross validated random forest model on 1,459 differentially methylated CpG Probes (DMPs) between the GS=6 and GS≥8 groups to identify PCa aggressiveness biomarkers. The power of the predictive model was further reinforced by ranking the DMPs with Decreased Gini and re-train the model with the top 97 DMPs (Testing AUC=0.920, predict accuracy=0.847). Similar approaches were performed to detect methylation differences between normal and PCa patient leukocyte DNA. Moreover, we analyzed 8 whole genome bisulfite sequencing (WGBS) patient leukocyte DNA specimens from the patient pool with Model based Analysis of Bisulfite Sequencing

data (MOABS), an integrated tool for bisulfite sequencing analysis. DNA microarray and WGBS results were highly correlated (r=0.946) and mutual biomarkers were identified. To make MOABS analysis widely accessible, we also utilized bioinformatics methods to implement MOABS to the galaxy platform and validated the power of MOABS-Galaxy with quick test and public bisulfite sequencing datasets. In summary, we identified a CpG methylation signature in leukocyte DNA that is associated with PCa aggressiveness and biochemical recurrence and developed the MOABS-Galaxy web service for DNA methylation analysis using bisulfite sequencing data. Our epigenetic mechanism study may provide an alternative option for PCa screening from epigenetic biomarkers, and implementation of MOABS could benefit biologists from non-computational background on bisulfite sequencing data analysis.

DEDICATION

To my parents, Yuewu and Quanying;

My wife Zhenna and my daughter Mia

I would not be here without your love…

ACKNOWLEDGMENTS

I would like to extend my sincere thanks to all of my committee members: Dr. Yun Huang, Dr. Roderick Dashwood, Dr. Chad Creighton, and Dr. Ken Chen. Thank you very much for your kindness and support, and I would not be able to finish the thesis without any of you.

I also would like to express my deepest appreciation for my current and former lab members, Dr. Deqiang Sun, Dr. Jia Li, Dr. Jin Li, Dr. Jianfang Li, and Yue Yin. I feel lucky to be part of such a great team, your guidance and friendship are so important during my Ph.D. journey. I am also thankful to all of my collaborators during Ph.D.: Dr. Jian Gu, Dr. Praveen Rajendran, Dr. Li Ma, Dr. Yuyan Han, Dr. Peijing Zhang, Dr. Zhenna Xiao, Dr. Gavin Johnson, Dr. Peijing Zhang, Dr. Vicky Chen, Dr. Sabeeta Kapoor, Dr. Minjung Lee, Dr. Xiaohua Su, Dr. Robert Tsai, and Dr. Fan Yao. I spent most of my Ph.D. on collaborating projects with you, and I enjoyed sharing perspectives with all of you very much.

Last but not least, I would also like to thank all the faculties and staff at Texas A&M University Institute of Bioscience and Technology. Special thanks go to my graduate program coordinator, Cynthia Lewis. Thank you all for creating a free, collaborative, and supportive community of the IBT graduate program.

NOMENCLATURE

3DCRT	3-dimensional conformal external beam radiotherapy
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
ACS	American Cancer Society
ADT	Androgen deprivation therapy
BCR	Biochemical recurrence
BPH	Benign prostatic hyperplasia
BS-Seq	Bisulfite Sequencing
CDIF	Credible methylation difference
CMR	Differentially methylated region
DMC	Differentially methylated cytosine
DNA	Deoxyribonucleic Acid
DRE	Digital rectal exam
EBRT	External beam radiotherapy
GS	Gleason Score
HRR	Homologous recombination repair
IMRT	Intensity modulated external beam radiotherapy
LHRH	Luteinizing hormone-releasing hormone
mCPRC	Metastatic castration-resistant prostate cancer
MOABS	Model based Analysis of Bisulfite Sequencing data

NGS	Next-generation sequencing
PCa	Prostate cancer
PCSM	Prostate cancer specific mortality
PSA	Prostate specific antigen
RF	Random forest
RNA	Ribonucleic acid
RP	Radical prostatectomy
RRBS	Reduced Representation Bisulfite Sequencing
RT	Radiotherapy
TSG	Tumor suppressor gene
USPSTF	US Preventive Task Force
UTI	Urinary tract infection
WGBS	Whole genome bisulfite sequencing

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Yun Huang (advisor) and Roderick Dashwood (co-advisor) of the Institute of Bioscience and Technology, Texas A&M University, Professor Chad Creighton of Baylor College of Medicine, and Professor Ken Chen of MD Anderson Cancer Center.

All work for Chapter II of the dissertation was completed by Mutian Zhang, in collaboration with Dr. Gu Jian and Dr. Yuyan Han. All the work for Chapter III of the dissertation was completed by Mutian Zhang, in collaboration with Dr. Jin Li.

Funding Sources

This work was supported by a Cancer Prevention and Research Institute of Texas (CPRIT) grant (RP140556), and a National Cancer Institute Specialized Program of Research Excellence (SPORE) grant (CA140388).

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Prostate Cancer Overview

Based on the Cancer Facts & Figures 2021, prostate cancer (PCa) is the most common cancer in men with 248,530 estimated new cases and 34,130 deaths (*1*). The probability of developing invasive PCa for a man from birth to death in the US is 12.1%, which indicates that almost 1 out of 8 men would develop PCa during their lifetime (*1*). However, it is widely known that “More men die with prostate cancer than because of it” – quite a big portion of prostate cancers are not noticed during their lifetime (*2*). Multiple studies reported a PCa prevalence higher than 50% at autopsy for men who are 70 years or older, while most of them were not diagnosed when they were alive (*3*-*5*).

The majority of undiagnosed PCa grows slowly and considered indolent; at the same time, the 5-year survival rate for localized PCa is almost 100%(*6*). However, as mentioned before, there are more than 30k PCa death cases every year, where the most majority of them were from aggressive PCa cases. This fact indicates the importance to distinguish aggressive PCa from indolent PCa. Most of the PCa patients have no symptoms when diagnosed, or only have non-specific urinary symptoms, while the advanced stage patients may have hematuria, bone pain, or other specific symptoms depending on metastasis location (*7*).

The most common screening criteria for PCa in the last 30 years is the blood level

of prostate specific antigen (PSA) (*8*). PSA is a protein made by both normal prostate and PCa cells, usually measured in nanograms per milliliter (ng/ml). While there is no absolute cutoff for the screening test, 4 ng/ml is widely accepted as a warning sign. A patient with PSA higher than 10 ng/ml will be considered high-risk for PCa. The wide use of prostate-specific antigen (PSA) testing for screening and early detection has contributed to the greatly improved survival of PCa (*9*). Another common screening method is the digital rectal exam (DRE), which is less effective than the PSA test, but could serve as a complementary method since DRE may occasionally discover PCa in men with normal PSA levels (*10*).

Screening and diagnosis

As mentioned in the previous section, the most common method for PCa screening is the blood/serum PSA test. However, doctors realized that current PSA tests are leading to too many overdiagnoses and unnecessary treatment (Figure 1). Although overtreatment may increase the survival rate of potential PCa patients, it will trade-off waste of clinical resources, financial pressure on patients, and more importantly, reduced life quality due to the side effects of treatment. Based on a synthesized study, the overtreatment rate may up to 67% (*11*), which indicated urgent needs for accurate screening methods.

Now the screening process for PCa has been much evolved. US Preventive Task Force (USPSTF) and American Cancer Society (ACS) are the most authorized institutes providing PCa screening suggestions. To be specific, USPSTF recommends a PSA test every 1-2 years, while DRE is not necessary. Moreover, men from 55 to 69 years should make individual decisions (take or not), but men above 70 are not recommended to take

the routine screening.

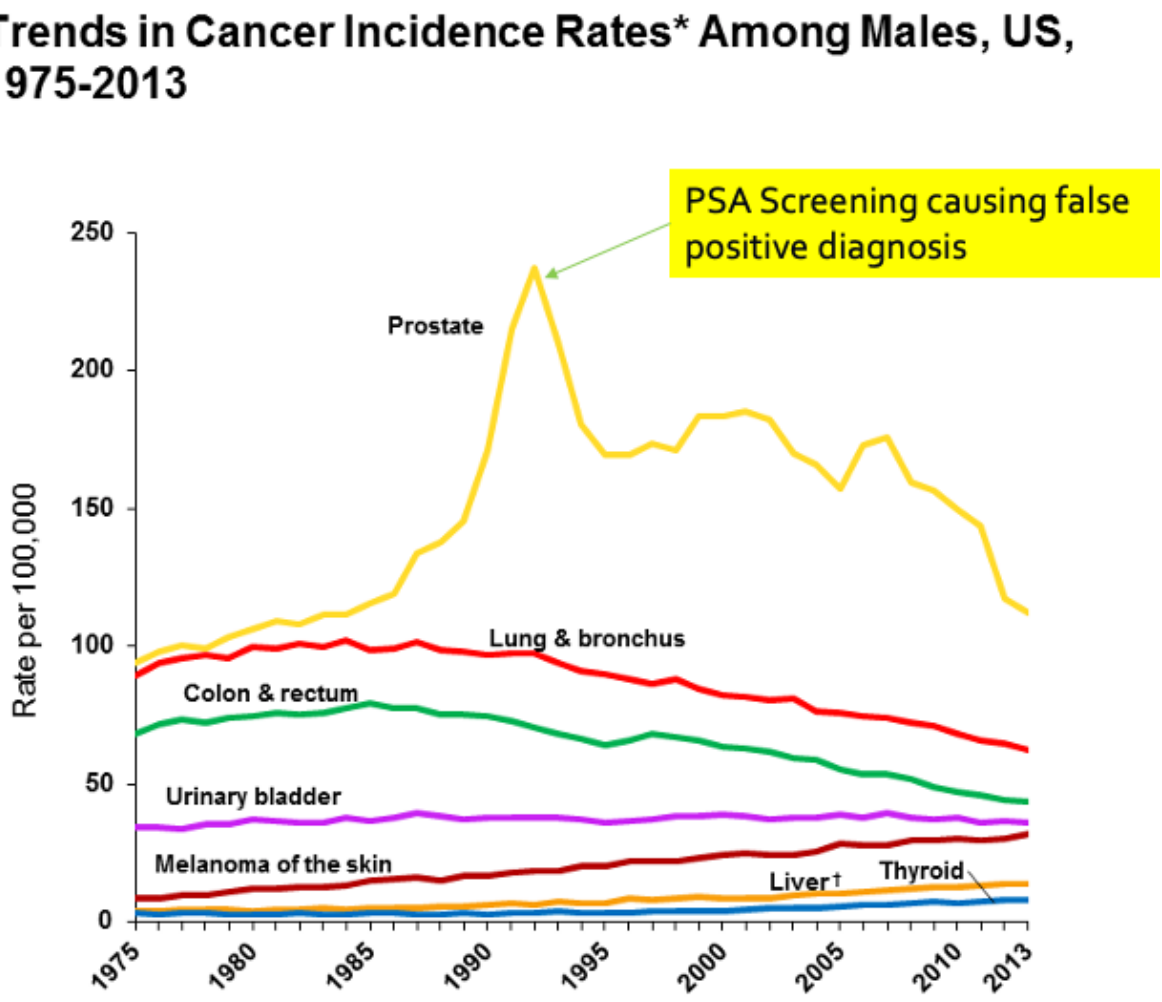


Figure 1. Trends in Cancer Incidence Rates Among Males, US (*12*).

Prostate cancer diagnosis rate shown in yellow. The diagnosis rate at 1990s was elevated by PSA screening low specificity, the rate dropped back to close 100 per 100k men in the 2010s. Figure reprinted from NIH open access SEER Cancer Statistics Review, 1975-2013.

ACS further categorized men into average/high/highest risk. Normal men with average risk should start screening at 50; African Americans or patient's father diagnosed

PCa before 65 are considered high risk and start screening at 45; patients with BRAC1,2 mutation history or with more than one first-degree relative diagnosed PCa at an early age are considered highest risk and start screening early as 40 (*13*).

The criteria to recommend prostate biopsy also varies. The most classic threshold is the high PSA level (> 4.0 ng/ml), which is not highly accurate because many other conditions would raise the PSA level. Inflammatory events such as urinary tract infection (UTI), benign prostatic hyperplasia (BPH), or prostatitis will increase long-term PSA level; Occasional irritation such as sexual activity or bicycling will stimulate PSA level within 48 hours; PSA level will gradually increase with aging (*13*).

Since PSA screening brings many false positives, many other helpful tests are performed for PCa screening. Gallium 68-PSMA PET/CT is a high-sensitive PCa detection method approved by FDA in December 2020 (*14*). Multiparametric MRI for abnormal prostate, along with the Prostate Image-Reporting and data system (PI-RADS) also provides accurate recommendations for the next step (*15*). Other than imaging, molecular and gene marker tests, such as 4K score, PCA3, and prostate health index (PHI) may also be helpful (*16*).

Besides novel screening methods, advanced PSA tests are emerging. PSA density is defined as the quotient of PSA level and prostate volume, and PSA density higher than 0.15 ng/mL/cc is considered as high PCa risk (*17*). Free and bound PSA ratio is another worth mentioning score, where high bound PSA increases risks of cancer. When the ratio is below 10, the patient is highly suspicious of PCa (*18*). Complexed PSA, or alpha 1-anti-chymotrypsin-complexed PSA, showed generally higher specificity but similar sensitivity comparing with total PSA (*19-21*).

When a patient is considered high PCa risk after initial screening, a biopsy will be performed for diagnosis. The traditional biopsy method usually collects 12 cores, while MRI-targeted biopsy takes only 6 selected cores and would reduce unnecessary biopsies by 25% (*22*). A recent study indicated that the combination of transrectal ultrasound-guided prostate biopsy (TRUS) and MRI – MRI-TRUS fusion-guided biopsy had the highest cancer detection rate (*23*). This new technology makes targeted and systemic biopsies available at the same time.

Stages and risk stratification

It is important to stage PCa patients and decide treatment plans specifically. Gleason Score (GS) is the traditional grading score for PCa developed by Dr. Donald Gleason (*24*) (Figure 2). When pathologists examine the cancer cells from a biopsy specimen, they grade cancer cells from 1-5 depending on how well or poorly they are differentiated, where 5 means the most aggressive cell type. The GS is decided by the sum of the top 2 dominant cell types. In most cases, GS ranges from 6 (3+3) to 10 (5+5). Patients with GS=6 or less are considered low risk for aggressive PCa, and patients with GS>=8 usually tend to have poorly differentiated, highly aggressive PCa. However, although GS=7 suggests an intermediate risk, patients can be divided into 3+4 or 4+3 groups with different prognoses. At the same time, GS is highly dependent on biopsy location and pathologists' experience, which tend to cause inaccurate staging, overtreatment, or misdiagnosis.

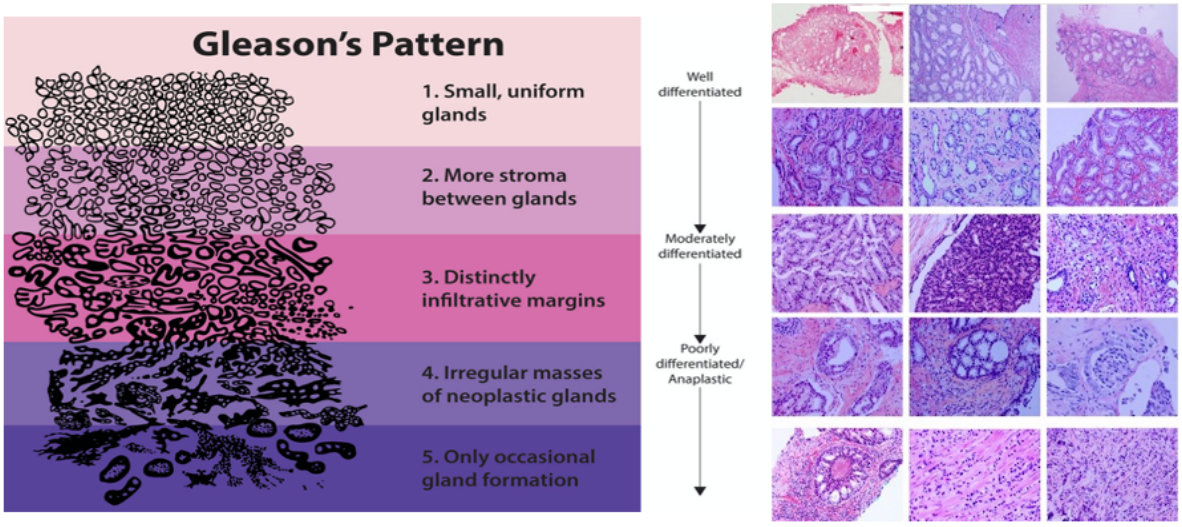


Figure 2. Gleason Scoring system and typical Gleason patterns (*25*).

Examples of Gleason patterns from 1 (well-differentiated) to 5 (poorly-differentiated) are illustrated. Reprinted from Chen et. al., "The evolving Gleason grading system", 2016. Permission to reuse the figure for thesis was obtained from www.copyright.com with license ID 1122005-1.

To better stage PCa, doctors are using the "TNM+PSA+GG" system (*26, 27*) (Table 1). "T" stands for 4 tumor stages, from not palpable (T1) to fixed and invasive tumor (T4). T1 tumor is not detectable by DRE and has 3 subtypes: a T1a patient has less than 5% of tumor tissue, a T1b patient has more than 5%, and a T1c patient has tumor found in needle biopsy tissues. T2 tumor is palpable but still within the prostate: T2a tumor involves less than 50% of one side, T2b tumor involves more than 50% of one side, and T2c tumor involves both sides. T3 tumor is extended outside the prostate, where T3a tumor invades to the bladder neck, and T3b tumor further invades to seminal vesicle. T4 tumor invades adjacent structures other than seminal vesicles, such as the rectum and pelvic wall. "N" describes regional lymph nodes status, where N0 is negative and N1 is positive. "M"

describes distant metastasis status, where M0 is negative and M1 is positive. M1 is further categorized into 3 subgroups, M1a defines non-regional lymph nodes metastasis, M1b stands for bone metastasis, and M1c illustrates metastasis to other sites. PSA levels are normally grouped into <10, 10-20, and >20. “GG” is short for Grade Group from 1-5, and it is decided by GS. GG 1 means GS 3+3, 2 means 3+4, 3 means 4+3, 4 means 4+4 (occasionally 5+3), and 5 means GS equals 9 or 10.

Based on the clinical features above, PCa patients will be staged into 4 major stages: I, II, III, and IV (*28*). Stage I PCa is confined inside of the prostate and cannot be detected by DRE (T1-2, N0, M0, PSA<10, GG1). Stage II PCa is palpable by DRE, while PSA level is less than 20 ng/ml (For example, Stage IIB: T1-2, N0, M0, PSA<20, GG2). Stage III is more complicated. Stage IIIA (T1-2, N0, M0, PSA>20, GG1-4) is an advanced form of Stage II, the tumor is confined in the prostate but the PSA level is more than 20 ng/ml. But for Stage IIIB, the tumor is bigger and outside of the prostate, and there is no requirement for PSA level (T3-4, N0, M0, PSA any, GG1-4). Stage IIIC is specific for Gleason Score equals to 9 or 10, regardless of the tumor size or PSA level (T any, N0, M0, PSA any, GG5). Any metastasis PCa will be considered as Stage IV. Stage IVA PCa has lymph node metastasis (T any, N1, M0, PSA any, GG any), and stage IVB has any types of other distance metastasis (T any, N any, M1, PSA any, GG any) (Table 1).

To determine the optimal treatment strategy for PCa, the NCCN risk stratification schema for localized prostate cancer is widely suggested (*29*). This guideline categorized PCa into 6 different risk level groups: very low, low, favorable intermediate, unfavorable intermediate, high, and very high. The risk levels were evaluated based on tumor size, Grade Group, PSA level, PSA density, and the percentage of positive biopsy cores (Table

2).

Stage		Tumor	Lymph node	Metastasis	Grade Group	PSA
Stage I		T1-2	N0	M0	GG1	PSA <20
Stage II	Stage IIA	T1-2	N0	M0	GG1	PSA 10-20
	Stage IIB	T1-2	N0	M0	GG2	PSA <20
	Stage IIC	T1-2	N0	M0	GG3/4	PSA <20
Stage III	Stage IIIA	T1-2	N0	M0	GG1-4	PSA ≥20
	Stage IIIB	T3-4	N0	M0	GG1-4	PSA any
	Stage IIIC	T any	N0	M0	GG5	PSA any
Stage IV	Stage IVA	T any	N1	M0	GG any	PSA any
	Stage IVB	T any	N any	M1	GG any	PSA any

Table 1. TNM+PSA+GG system for staging of prostate cancer(*30*).

Details of TNM+PSA+GG staging system. Patients will be staged into I, II, III, and IV based on their clinical features. T1: tumor not palpable; T2: tumor palpable but confined within prostate; T3: tumor extended outside the prostate but within prostate seminal; T4: tumor invades to nearby organs (bladder, rectum. etc.) N0: no metastasis to LNs; N1: metastasis to LNs. M0: no distant metastasis; M1: distant metastasis. GG1: Gleason Score=6; GG2: Gleason Score=3+4; GG3: Gleason Score=4+3; GG4: Gleason Score=8; GG5: Gleason Score=9/10.

Risk	Tumor	Grade Group	PSA	Biopsy	Other
Very low risk	T1	GG1	PSA <10	<3 positive	PSA density <0.15
Low risk	T1	GG1	PSA <10	≥3 positive	Not qualified for very low risk
Favorable intermediate risk	T1/2	GG1/2	N/A	<50% positive	
Unfavorable intermediate risk	T2	GG3	N/A	≥50% positive	
High risk	T3	GG4/5	PSA >20	N/A	Not qualified for very high risk
Very high risk	T3/4	GG5	PSA >20	>4 with GS 4/5	

Table 2. NCCN Risk Stratification schema for localized PCa(*29*).

NCCN Risk Stratification schema for localized PCa. Localized PCa patients are stratified into 6 risk groups: very low, low, favorable intermediate, unfavorable intermediate, high, and very high risk. T1: tumor not palpable; T2: tumor palpable but confined within prostate; T3: tumor extended outside the prostate but within prostate seminal; T4: tumor invades to nearby organs (bladder, rectum. etc.) GG1: Gleason Score=6; GG2: Gleason Score=3+4; GG3: Gleason Score=4+3; GG4: Gleason Score=8; GG5: Gleason Score=9/10.

Treatment

There are different treatment strategies for PCa patients based on their stage and risk. For very low risk patients (usually defined as a localized small tumor, PSA <10ng/ml, low GS, and no symptoms), active surveillance will be performed after discussion between physician and the patient – some favorable intermediate risk group patients may also choose active surveillance. This strategy is supported by a 10-year patient tracking study: among 1,643 men diagnosed with localized prostate cancer, 545 underwent active

surveillance, 553 had surgery, and 545 received radiotherapy. However, there are only 8 PCa specific death in the active surveillance group (5 in the surgery group, 4 in the radiotherapy group), which indicates that >99% of the indolent patients would actually survive under active surveillance. At the same time, the progression rate of PCa is significantly higher in the monitoring group (22.9 per 1000 person-years) comparing with the surgery group and radiotherapy group (8.9 and 9.0 per 1000 person-years, respectively) (*31*). This result demonstrates that patients should be cautious about the long-term consequences that may occur. Since many patients may already at an elderly age, it would be a wise choice to undergo active surveillance and enjoy higher quality for the near-term life. Patients under monitoring are still recommended for a PSA test every 3-6 months; if PSA level remains, prostate MRI is recommended every 2-5 years.

For higher risk patients, there are 3 major types of treatment: Radical Prostatectomy (surgery), Radiation Therapy, and Hormone Therapy. In the US, robotic-assisted radical prostatectomy is commonly performed. Comparing with traditional surgery methods (open, or laparoscopic), robotic-assisted surgery will minimize blood loss, and more importantly, it may best assist with the nerve-sparing approach on both sides of the prostate, which is crucial for the return of the erectile function and urinary continence (*32*).

A report shows that patients who underwent robotic-assisted surgeries were significantly more likely to recover from erectile function (*33*). Pelvic muscle exercises (PFMT, Pelvic Floor Muscle Training) help with recovery from urinary continence: patients who exercised may have increased muscle strength, higher Health-Related Quality of Life, better results in pad test and bladder diary (*34, 35*).

Radical Prostatectomy is a solid treatment for early-stage patients. Based on a

long-term survival study from Lancet, 10-year survival for early-stage prostatectomy patients was 94% (*36*). The side effects include bladder neck stricture, long-lasting bladder control problems (urinary continence, from minor dribbling to needing to wear incontinence pads), sexual dysfunction, and surgery risks (bleeding, infection, blood clots, etc.).

There are 2 major types of radiation therapy: external beam RT (EBRT) and brachytherapy. EBRT is widely used for patients from low risk to very high risk. However, for late-stage patients (later than unfavorable intermediate risk group), hormone therapy is usually combined (discussed later). Most cutting-edge technologies include Intensity Modulated Radiation Therapy (IMRT) and Image-Guided RT (IGRT). IMRT is an advanced form of 3D-conformal RT, which selectively targets tumor tissue by intensity and shape, and minimizes the margins of normal tissue. IGRT works similarly, the doctors will take advantage of the images by fiducial markers, ultrasound, MRI, x-ray images of bone structure, CT scan, 3-D body surface mapping, or electromagnetic transponders to localize tumor tissue before RT (*37, 38*).

Brachytherapy is an alternative method for EBRT. “Branchy” is a Greek word that means “from a small distance”. Different from external RT, brachytherapy aims for a more direct and intensive RT to the PCa tissue. Radioactive seeds will be directly inserted into the prostate. While brachytherapy has a higher radioactive concentration, is it also divided into high-dose-rate brachytherapy (HDR-BT) and low-dose-rate brachytherapy (LDR-BT). The LDR-BT seeds can be permanently placed with iodine-125 (125I), palladium-103 (103Pd), and cesium-131 (131Cs) isotopes. HDR-BT usually uses iridium 192 (192Ir) or cobalt 60 (60Co) and stays in the bladder for 1-3 days. Brachytherapy can be performed alone for low risk and favorable intermediate risk patients. For unfavorable intermediate

risk or later stage patients, EBRT is often combined with brachytherapy for the risk of extra-prostate extension (*39, 40*).

Multiple studies reported a slightly lower survival rate of radiation therapy patients comparing with surgery patients (*41, 42*). The side effects of radiation therapy include urinary continence (but most men would recover after RT since structural damage is limited comparing with surgery), sexual dysfunction (30-40% erectile dysfunction rate, peaked at 2 years after treatment, and remained after 3 years), and chronic radiation proctitis such as diarrhea and tenesmus (*43, 44*). Acute urinary obstruction may happen specifically to brachytherapy patients since the insertion may cause inflammation. A study indicated that PCa patients who received radiation therapy may benefit from sildenafil (such as Viagra): up to 68% of patients at 12 months after therapy were responsive to sildenafil (*45*).

Hormone therapy, or Androgen Deprivation Therapy (ADT), is the most effective systematic therapy for patients with hormone-sensitive prostate cancer (*46*). Androgens, especially testosterone, are necessary for prostate cancer cell growth. Testicles produce 95% of testosterone, the remaining 5% are produced by adrenal glands. By reducing blood testosterone level to the castration level (<50 ng/ml), >90% of PCa tissue would shrink, which also indicates that ADT may limit the growth of PCA but not a complete cure. Based on a 2016 report, there are approximately 34% PCa patients received ADT in the US, while up to 68% of ADT use in Eastern Europe (82% in Hungary) (*47*). ADT is usually recommended for metastasis PCa and combined with RT for unfavorable intermediate risk or later stages to make the treatment more effective. Recurrent patients after surgery or RT are often treated with ADT as well.

There are 3 types of ADT to control testicular androgen levels: orchiectomy (surgical castration), Luteinizing hormone-releasing hormone (LHRH) agonists, and LHRH antagonists. Orchiectomy is out-of-date since the permanent removal of the testicles is not widely accepted. LHRH, or Gonadotropin-releasing hormone (GnRH), is a hormone from the hypothalamus which can stimulate luteinizing hormone (LH) in the pituitary gland, thus increasing testosterone production. By providing LHRH agonists, testosterone production would increase by 200% for several weeks, which is called the testosterone flare effect. After that, the pituitary gland will be desensitized to an LHRH agonist, which finally results in testosterone production termination. However, the first few weeks of higher testosterone production can greatly stimulate tumor growth. Patients with bone and spine metastasis may suffer severe pain or even at paralysis risk. The LHRH agonists available in the United States include Leuprolide (Lupron, Eligard), Goserelin (Zoladex), Triptorelin (Trelstar), and Histrelin (Vantas). LHRH antagonists work on the other way, they can deduce testosterone level without the flare effect. The LHRH agonists available in the United States include: Degarelix (Firmagon) and Relugolix (Orgovyx). While most of the drugs need an injection, the newly FDA-approved Relugolix are oral pills. Side effects are mostly from low level of testosterone, such as hot flashes, sexual dysfunction, osteoporosis, gynecomastia, and mental issues. Clinical trials of newer antiandrogens and oral LHRH antagonists are ongoing and standard ADT treatment may alter with time. Besides 95% androgens produced by testicles, there are other treatments to lower the 5% androgen production from adrenal glands. Abiraterone (Zytiga) is an inhibitor to an enzyme called CYP17, which prevents adrenal gland cells from generating androgens. Abiraterone is especially effective for PCa patients who are resistant to LHRH agonist or antagonist.

Another supplementary ADT is to block the androgen receptor (AR). Bicalutamide (Casodex), Enzalutamide (Xtandi), darolutamide (Nubeqa), and apalutamide (Erleada) are current approved drugs. Side effects of these drugs are mainly on feminization, including breast tenderness, decreased body and muscle mass, sexual dysfunction, and other effects such as hypertension, hypokalemia, and seizure. Estrogens used to be the main alternative to orchiectomy, but serious side effects such as blood clots and breast enlargement made it replaced by other ADT drugs. An interesting discussion about ADT is called intermittent ADT (*48*). Patients who receive ADT may stop the treatment for 6-12 months after the PSA level drops close to an undetectable level. When the PSA level rises again (>10 ng/ml), ADT will resume. The overall survival will not drop by doing so, but the life quality would increase. Thus, NCCN published Flash Updates in 2020 suggesting "Intermittent ADT can be considered for men with M0 or M1 disease to reduce toxicity" (*49*). Interestingly, 30% of the patients who stopped ADT still had testosterone levels <50 ng/ml, which indicated ADT has irreversible damage to the patient androgen hormone system (*50*). Other PCa treatment methods include High intensity focused ultrasound (HIFU), focal laser therapy, cryotherapy, and so on.

As previously discussed, PCa patients are stratified into different risk groups, with treatment strategies customized correspondingly. Localized PCa patients, including very low, low, favorable intermediate, and unfavorable intermediate risk groups. For very low and low risk patients, active surveillance is commonly recommended; some low-risk patients may take radical prostatectomy or radiation therapy. For favorable intermediate risk patients, active surveillance is still an option. However, most of them will undergo RT or surgery, and ADT is usually unnecessary. For unfavorable intermediate risk patients,

treatments are similar with the favorable intermediate group, but NCCN also recommends 4-6 months of ADT starting 2 months before RT to better control tumor growth, only if the patient has no or low cardiac morbidity risk. High or very high-risk patients with T3-T4 tumor stage, Gleason Score >=8, or PSA >= 20ng/ml need more complicated treatments. EBRT combined with ADT for 18-24 months is preferred, while brachytherapy boost may serve as a boost for the RT process. Prostatectomy with pelvic lymphoma dissection is still an option if PCa tissue is not fixed to adjacent organs. If the patient undergoes surgery, RT or ADT afterward is controversial, while short-term ADT may be beneficial (*51*). To determine whether RT or ADT is needed, biomarker assay tests such as Decipher may help, however it is not a standard approach yet.

For metastatic PCa, the treatment strategy is mainly dependent on ADT sensitivity. Castration-sensitive patients benefit from ADT and 2nd generation antiandrogen combination, especially for high risk PCa. Intermittent ADT would increase the patient overall survival rate but not PCa-specific survival, as discussed previously. Metastatic castration-resistant PCa is defined as cancer progression (rising PSA level or new metastasis location) despite testosterone level keeps at castration level. Androgen receptor inhibitors such as Enzalutamide, apalutamide, and abiraterone are now FDA-approved to treat these patients, however there is only little chance to expect patient responses if one of the drugs has already failed on them (for example, only ~10% response with abiraterone after enzalutamide failure). Genetic tests for deficient homologous recombination repair (HRR) genes (ATM, BRAC1/2, PALB2) and deficient DNA mismatch repair genes (MLH1, MSH2, MSH6, PMS2) are recommended as well. For patients with deficient HRR genes, PARP inhibitors such as Olaparib and rucaparib are

helpful; deficient DNA mismatch repair genes patients have immunotherapy as an option, pembrolizumab is commonly recommended (*52*).

After treatment, patients will have their blood PSA level tested every 6 months for the first 5 years, and yearly afterward. About 25-30% of PCa patients experienced biochemical recurrence (BCR). BCR is defined as PSA >= 0.2ng/ml after prostatectomy or PSA rising >=2 ng/ml after RT. For BCR after prostatectomy, salvage RT is the standard of care, often combined with a short-term ADT, especially when PSA > 0.6 ng/ml. However, the effect of salvage RT or ADT is controversial. A study in 2011 from Harvard Radiation Oncology Program indicates salvage RT is beneficial for overall patients (*53*). However, a study in 2008 done by John Hopkins School of Medicine reported that salvage RT positive effect was limited to men with a prostate-specific antigen doubling time of fewer than 6 months and remained after adjustment for pathological stage and other established prognostic factors; salvage RT after 2 years of BCR provided no improvement for patient survival, thus made the overall survival rate not significantly increased. Moreover, additional ADT did not contribute to PCa-specific survival (*54*). A 2009 study from Fox Chase Cancer Center also demonstrated that salvage RT only improved BCR-free survival, but not impacted systemic progression and overall survival (*55*). For BCR after RT, radical prostatectomy, cryotherapy, brachytherapy, or HIFU may serve as salvage treatment. It is reported that salvage radical prostatectomy resulted in significantly higher survival comparing with cryotherapy, especially for young and healthy patients with BCR after RT (*56*).

If none of the salvage treatment is applicable to a BCR patient, observation or delayed ADT is recommended. It is the standard cure for patients whose PSA levels

doubled over 12 months or elderly people. However, ADT improves cancer-related survival but not overall survival (*57*). If PCa is much more aggressive (PSA doubling time less than 10 months or resistant to ADT), 2nd generation antiandrogens such as enzluzamine or darolutamide are recommended besides. During the ADT, denosumab every 6 months is beneficial for bone health.

Treatment		Preferred patients	Description	Comments
Active Surveillance		Very low, low risk; good for some intermediate risk	Treatment decision will be made by both clinician recommendations and patient values and preferences.	Mortality remains for very low or low risk patients; however, more long-term responses and optimal procedure needed.
Radical Prostatectomy		Low to intermediate risk patients; can combine with other treatment for high risk or late stage PCa patients	Robotic-assisted radical prostatectomy and nerve-sparing approach are less harmful for patient health.	Irreversible surgery with side effects, especially sexual dysfunction and urinary continence; good overall survival for early-stage patients.
Radiation Therapy	EBRT	Mostly for low and intermediate risk patients; can also serve as a part of combined therapies for high risk and metastatic patients	Most cutting-edge technologies include Intensity Modulated Radiation Therapy (IMRT) and Image-Guided RT (IGRT) -	Advanced RT technologies target tumor tissue accurately; side effects may harm patient life quality; lower long-term survival rate
	Brachy-therapy	Preferred for low risk and favorable intermediate risk patients; can also combine with other methods for unfavorable intermediate, high risk, or metastatic patients	Radioactive seeds will be directly inserted into the prostate; can be high-dose or low-dose.	
Hormone Therapy	Orchiectomy	Mostly intermediate-high risk or late-stage patients; effective for hormone-sensitive patients	Surgical removal of testicles is out-of-date.	Irreversible surgery would harm patient life permanently.
	LHRH agonists		Pituitary gland will be desensitized to an LHRH agonist, which finally results in testosterone production termination.	Testosterone flare in the first few weeks may worsen metastatic PCa.
	LHRH antagonists		LHRH antagonists deduce testosterone level to castration level.	Side effects are mostly form low level of testosterone: hot flashes, sexual dysfunction, osteoporosis, gynecomastia, and mental issues

Table 3. Summary of PCa treatment methods.

Prostate Cancer Markers (PCMs)

As discussed before, primary care physicians need a simple standard for PCa screening, and accurate early detection which can identify PCa is crucial for patient overall survival and life quality. Although the 5-year diagnosis rate is 15-fold with PSA >1.5ng/ml men, the overall rate was only 7.85%, a showing high false-positive rate for PSA tests (*58*). PSA screening has poor specificity and leads to overtreatment (*59*). In fact, most PSA tests are not done by urologists. A report shows that only 6.1% of PSA tests were ordered by urologists, while most PSA tests were scheduled by internal medicine and family medicine doctors (64.9% and 23.7%, respectively) (*60*). Prostate Cancer Markers (PCMs) are introduced to patients instead. A PCM is a molecule that can serve as a sign of normal or cancerous prostate in tissue, blood, or urine. PCM tests lead to precise, targeted, and personalized therapy for PCa patients. Individuals may select which test(s) they would take based on the PCM Bucket Algorithm (Figure 3).

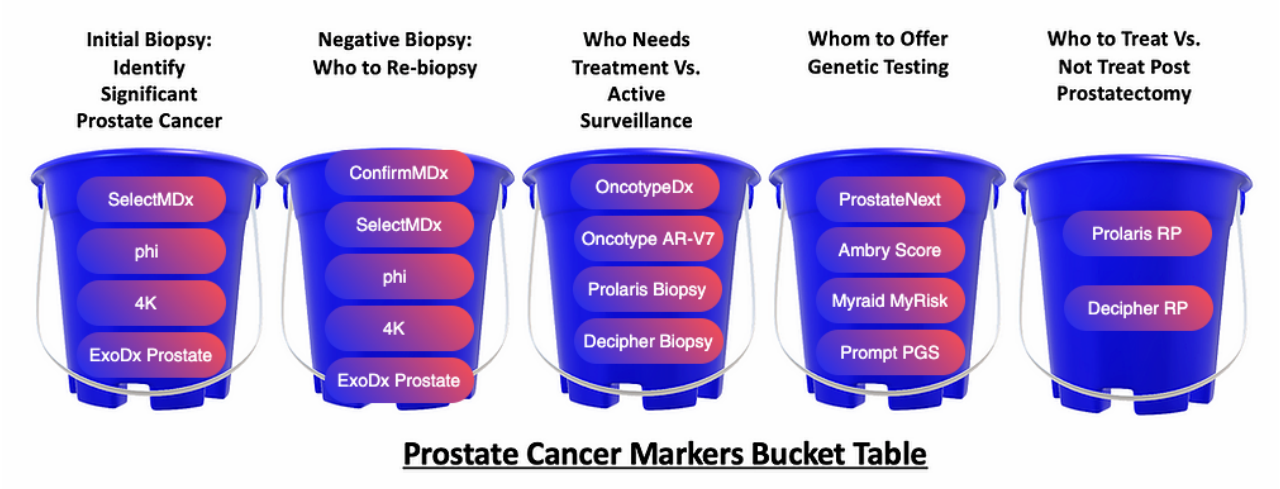


Figure 3. Prostate Cancer Markers bucket table(*61*).

Different PCM tests are available based on patient needs, from early screening and genetic tests to re-biopsy and re-treatment after initial surgeries. PCM can be detected from patient blood, urine, tissue, and saliva specimens. Reprinted with permission from pcamarkers.com.

To access the risk of PCa, there are several tests available: PCA3, Mi-Prostate Cancer, SelectMDx, 4Kscore, Prostate Health Index, Apifiny, ConfirmMDx. PCA3 is short for Progensa Prostate Cancer 3 assay, which measures prostate cancer gene 3 concentration in post-DRE, first-catch urine specimens (*62*). PCA3 score ranges from 0 to 125. Normally, a score below 25 is considered negative and no biopsy need; when the PCA3 score >100, the positive biopsy rate can be up to 50% (Figure 4). Another study demonstrated that PCA3 scores were correlated with tumor grade and volume in prostatectomy samples (*63*). In addition, the National Comprehensive Cancer Network (NCCN) guidelines recommend that PCA3 should be considered in men thought to be at higher risk of having PCa, despite a negative biopsy (*64*).

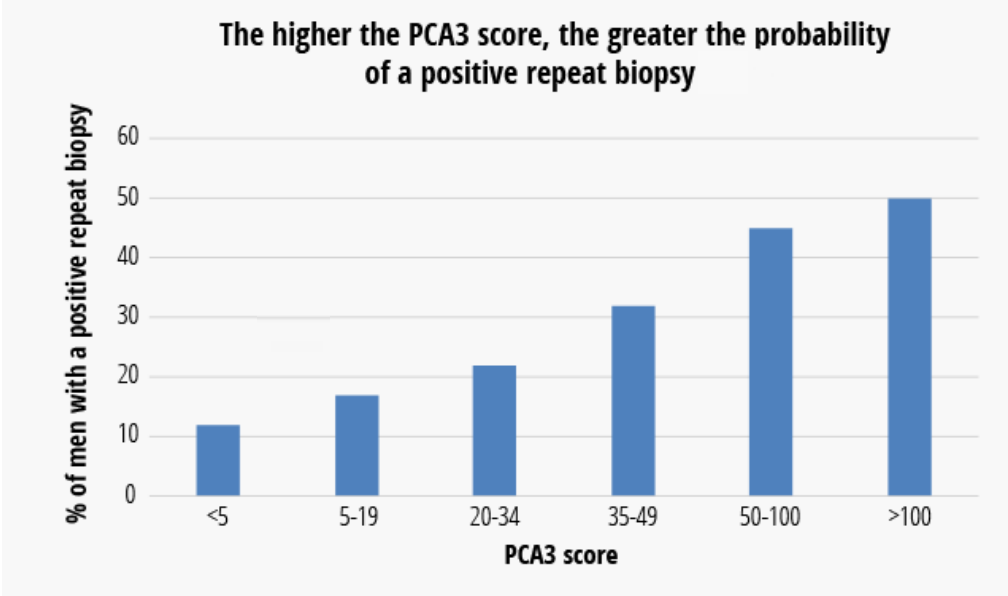


Figure 4. Probability of a positive repeat biopsy based on PCA3 scores(*65*).

PCA3 score is positively correlated with the probability of a positive repeat biopsy. A patient with a >100 PCA3 score can have a 50% positive chance on his next biopsy.

An advanced version of PCA3 is Mi-Prostate Score (MiPS) (*66*). In addition to measuring PCA3 level in urine, it also tests TMPRSS2 and ERG levels, then a multivariable regression generates the combined MiPS score. TMPRSS2-ERG gene fusion is recognized as a PCa early-stage biomarker, and their transcripts are detectable in urine after DRE (*67, 68*). The risk of prostate cancer and late-stage prostate cancer will be delivered to patients. Comparing with PSA3 alone, PSA3+TMRPSS2-ERG has a higher predictive rate for PCa, thus avoid 35-47% of unnecessary biopsies while only missing 1-2.3% of progressive PCa cases. MiPS can be done for a man with any PSA level (*65*).

SelectMDx is a urine biomarker test. After DRE, patient urine specimens will be taken to a laboratory to test DLX1 and HOXC6 mRNA levels, while PSA serves as a reference. DLX3 is related to PCa progression, and HOXC6 regulates PCa cell

proliferation. Combining with other risk factors such as PSA, age, and DRE result, SelectMDx will generate a final risk score for positive biopsy and GS>=7. A study shows that SelectMDx has a 99.6% negative predictive value for GS>=8 prostate cancer. This test can be performed on any PSA level patients (*69*).

4Kscore test is a blood test for 4 specific PCa biomarkers – total PSA, Free PSA, Intact PSA, and Human Kallikrein 2 (HK2). HK2, also known as kallikrein-related peptidase 2 (KLK2), is another member of the KLK family (PSA is also known as KLK3) and considered as a biomarker for PCa. Together with age, DRE results, and prior biopsy results (if available), the 4Kscore algorithm can deliver the risk for aggressive PCa (GS>=7) and the likelihood of distant metastasis within the next 10 years. 4Kscore can be utilized on any PSA level, but not for the patients who have biopsy within 6 months or DRE within 96 hours. A report shows that among 1012 men, the 4Kscore test identified last stage PCa with 95% sensitivity and AUC=0.82; about 36% biopsies could be avoided or delayed (*70*). Another multi-institutional prospective trial study indicates 4Kscore test has 97% sensitivity and 95% negative predictive value in a sample space where more than 50% of the patients are African American (*71*).

Prostate Health Index (phi) is a blood test that combines PSA, free PSA, and pro2PSA. Free PSA and pro2PSA are isoforms of PSA in blood, and they are helpful to improve the specificity for PCa prediction (*72*) (Figure 5). The algorithm for phi is: phi = (pro2 PSA/free PSA) * $\sqrt{PSA}$. Phi ranges from 0 to 55+, patients with different phi scores will fall into 4 categories, the probability of PCa and confidence interval are marked. Based on a 2011 study, phi has 3-folds more specificity than PSA and may reduce 26% of unnecessary biopsies (*73*). Phi is recommended for patients with PSA 2-10 ng/ml and

negative DRE.

phi range	Probability of cancer	95% Confidence interval
0-26.9	9.80%	5.2-15.4%
27.0-35.9	16.80%	11.3-22.2%
36.0-54.9	33.30%	26.8-39.9%
> or =55.0	50.10%	39.8-61.0%

Figure 5. Phi score associate with probability of PCa.

Phi score is positively correlated with the probability of PCa. A patient with a >55 phi score can have a 50.10% chance of PCa in the future.

Apifiny is a blood test targeting 8 different autoantibody biomarkers identified by Wang et. al, (*74*). The study reported 22 candidates, and Apifiny selected 8 of them (*CSNK2A2*, centrosomal protein 164 kDa, *NK3 homeobox 1, aurora kinase interacting protein 1*, 5´-UTR *BMI1*, *ARF6*, chromosome 3´ UTR region *Ropporin/RhoEGF*, and *desmocollin* 3), who are in charge of androgen response regulation, cellular structural integrity, and cell cycle regulation. Apifiny reports a score from 0 to 100, and a score higher than 60 is considered high risk. A study reported that Apifiny has a sensitivity of 0.603 and a negative predictive value of 0.89 (*75*). Apifiny is suitable for PSA>2.5 ng/ml patients, and since it depends on immune reaction, patients who are taking steroids or immunocompromised are not recommended for Apifiny tests.

ConfirmMDx is a tissue test for patients with a negative biopsy result, and wondering if they should re-do the biopsy. ConfirmMDx focuses on DNA methylation

changes in biopsy tissue cells, with help of methylation-specific PCR. To be specific, this test detects the methylation levels of GSTP1, APC, and RASSF1. Methylation silencing on GSTP1 is widely reported in prostate cancer, together with APC and RASSF1 (*76-78*). Based on test results from different areas of the prostate, confirmMDx delivers a report with the likelihood of PCa on repeat biopsy and low/high grade of PCa. Studies indicated that confirmMDx has a 96% of negative predictive value, and may reduce 90% of repeated biopsies (*79, 80*). ConfirmMDx is suitable for negative biopsy patients with any PSA level. This method is not yet approved by FDA.

Another important topic is to decide whether a patient needs treatment or just active surveillance. There are 4 tests available for this category: OncotypeDx, Oncotype AR-V7, Prolaris Biopsy, Decipher Biopsy. OncotypeDx is a test for needle biopsy tissue. A 2014 study identified 17 genes associated with biopsy that predict prostate cancer-specific death, clinical recurrence, adverse pathology, and metastasis (*81*). OncotypeDx tests those 17 genes' expression levels and provides a genomic prostate score (GPS) ranging from 1 to 100. A prospective study demonstrated that OncotypeDx changed 31% of treatment recommendations while physician confidence improved by 85% (*82*). Another retrospective study gave the fact that the rate of active surveillance was increased by 56% (*83*). OncotypeDx test is designed for very low, low, and favorable intermediate risk patients. Ideal patients should have PSA<20 ng/ml, GS=6 or 3+4, previous biopsy sample size 1mm.

Oncotype AR-V7 a blood test to identify metastatic castration-resistant prostate cancer (mCPRC) patients. mCPRC patients are very unlikely to respond to ADT and consider RT instead. Oncotype AR-V7 detects AR-V7 protein level in the nucleus of

circulating tumor cells (CTCs) in patient blood. Androgen receptor splice variant 7 positive patients are largely reported to not benefit from AR-targeted therapies (*84, 85*). AR-V7 negative patients may continue ADT, such as abiraterone or enzalutamide. A study indicated that positive AR-V7 correlated with androgen treatment response, tumor progression, and overall survival (*86*).

The Prolaris cell cycle progression test measures the expression levels of 31 cell cycle progression genes, as well as 15 housekeeping genes, to generate a score ranges from 1 to 10. This score predicts tumor cell proliferation and thus predicts tumor progression. Combining patient clinical features such as age, tumor stage, pre-biopsy PSA level, biopsy positive rate, GS, and risk group, the Prolaris cell cycle assessment provides tumor aggressiveness and 10-year prostate cancer-specific mortality risk (*87*). Validation projects confirmed that the Prolaris test predicted prostate cancer-specific death by providing additional prognostic information, comparing with PSA or GS (*88, 89*). More importantly, studies also demonstrated that Prolaris has a strong ability to predict cancer recurrence after prostatectomy (*90*). For example, a study found that the hazard ratio was 2.55 for doubling expression of Prolaris genes, and the Prolaris score is associated with 10-year survival (p-value=0.13) (*91*). Prolaris is recommended for patients with PSA<100 ng/ml, and who have not undergone any ADT or RT. This test is also useful for post-prostatectomy patients to decide whether they have a high BCR chance and thus need extra treatment.

Decipher biopsy tests 22 selected RNA biomarkers for multiple pathways – most of the RNAs are related to androgen signaling, cell proliferation, and differentiation, cell structure and adhesion, immune response, and cell cycle (*92*). Decipher score ranges

from 0 to 1. This score predicts the likelihood of 5-year metastasis, high-grade PCa, and 10-year prostate cancer-specific mortality. Decipher showed strong ability to distinguish PCa metastasis and GS>=8 patients (hazard ratio 1.72 per 10% Decipher score) (*93*). It is also reported that Decipher system reclassified 46% of patients from original NCCN risk categories (*93*). Decipher biopsy test is recommended for any NCCN risk group, any GS, and any PSA patients after biopsy. However, this is a predictive test to determine treatment or not, so patients already underwent ADT or RT are not suitable. Similar to the Prolaris test, the Decipher test is also useful for post-prostatectomy patients for further treatment decisions.

One last bucket for prostate cancer marker tests is whom to offer genetic tests. PCM tests in this basket include ProstateNext, Ambry Score, Myriad MyRisk, and Prompt PGS. ProstateNext is a blood DNA test focuses on 14 genes associated with hereditary prostate cancer: ATM, BRCA1, BRCA2, CHEK2, EPCAM, HOXB13, MLH1, MSH2, MSH6, NBN, PALB2, PMS2, RAD51D, and TP53. The gene list was from a 2016 study, which identified 16 DNA-repair gene mutations in metastatic PCa men (overall prevalence 11.6%) comparing with localized PCa patients (4.6%) and normal men (2.7%) (*94*). Patient DNA will be sequenced by next-generation sequencing (NGS) methods. Mutations, deletions, and duplications of the genes in the list will be identified. If at least one known mutation is found, the patient is considered positive for the test. If only novel alternations of those genes are found, the patient will be marked as Variant of Unknown Significance (VUS). Patients with high risks may seek help from their health providers with PCa screening plans and prevention options (*94*). The following men should consider ProstateNext: 1) have at least one family member diagnosed with PCa younger than 50; 2)

have at least one family member diagnosed with metastatic PCa; 3) have multiple people on the same side with breast, ovarian, prostate, pancreatic, or other cancers; 4) have at least one family member who was/were found to have a cancer gene mutation.

The Ambry Score is another genetic test for PCa risk evaluation, also known as Polygenic Risk Score (PRS). Ambry Score mainly detects 72 single nucleotide polymorphisms (SNPs) on the human genome, which were identified by previous studies in a large prostate cancer population. Other clinical features include age and ethnicity. The Ambry Score calculation is highly dependent on the accuracy of clinician-provided data (*95*). Sequencing of the SNPs is carried out by a bait-capture methodology using long biotinylated oligonucleotide probes followed by polymerase chain reaction (PCR) and Next-Generation sequencing (*96*). As a result, the remaining lifetime risk will be delivered. For men with no current PCa diagnosis, the average risk cut-off for prostate cancer is 10.2%; for men already diagnosed with PCa, the average PRS score is 1 (higher score indicates increased chance for aggressive PCa) (*97*). Ambry Score is specifically designed for men of Northern European Ancestry with a personal and/or family history of either an early-onset (<50 years old) metastatic PCa or at least one person with prostate, pancreatic, breast, or ovarian cancers. In addition, the patient should have no personal or family history of a mutation in a prostate cancer susceptibility gene, including ATM, BRCA1, BRCA2, CHEK2, EPCAM, HOXB13, MLH1, MSH2, MSH6, NBN, PALB2, PMS2, RAD51D, and TP53. In this case, Ambry Score serves as a supplementary method for ProstateNext.

Myriad Myrisk is another blood or saliva test based on prostate cancer markers. It includes 29 genes panel and identifies elevated risk for 8 hereditary cancers including

PCa. For PCa, target genes are BRAC1&2, MLH1, MSH2, MSH6, PMS2, EPCAM, TP53, and NBN. Multiple clinical reports validated the existence of those gene mutations in PCa patients and the power of gene testing on the prediction of PCa (*98, 99*). Similar to ProstateNext, Myriad Myrisk also delivers positive, negative, and VUS results. Besides, patients without genetic mutations but identified other risk factors (personal clinical risk factors, family history, abnormal additional genetic markers) are marked as elevated risk. Myriad Myrisk test is recommended for a man with 1) personal history of male breast cancer, metastatic PCa, colon or rectal cancer; 2) family history of breast, colon, rectal, or uterine cancer under 50; or ovarian, metastatic prostate, pancreatic cancer at any age. Prompt Prostate Genetic Score (PGS) is a saliva test for the genetic test bucket. It is designed for any man who is eager to know his lifetime risk of developing PCa. Users may collect the sample on their own using a cheek swab. Prompt PGS is based on more than ten years of prostate cancer research. It has been validated in over 10,000 men in some of the most important prostate cancer trials ever performed. Prompt PGS detects prostate cancer risk-associated SNPs and delivers a lifetime prostate cancer risk. The average lifetime risk for US men is 11.1%, and 14.1% for men with a first-degree relative with PCa. It is reported that Prompt PGS would identify 3x more patients at higher risk than family history alone, and it is also 2 times as efficient as PSA screening, detecting 20% more late-stage PCa (*100*).

In summary, there has been a significant process on PCa biomarker research in the past 2 decades. Some of the biomarkers were clinically validated and related tests were commercialized. By referring to the prostate cancer markers bucket algorithm, a patient can make critical decisions with his physician on the following questions: 1) whether the

patient has aggressive PCa after biopsy; 2) whether the patient needs a re-biopsy after negative biopsy; 3) whether the patient needs treatment (prostatectomy, RT, or ADT) or just active surveillance; 4) whether the patient should undergo genetic testing; 5) whether the patient needs salvage treatment after prostatectomy. PCa biomarkers can be detected in patient blood, urine, tissue, or even saliva. The PSA era has ended due to its poor specificity, and a bunch of new biomarkers are emerging on protein, RNA, DNA, or epigenetic levels. Based on the latest NCCN guidelines version 2.2020 for prostate cancer early detection, the probability of higher-grade PCa (GS>=3+4, GG>=2) may be further defined by phi, selectMDx, 4KScore tests. PSA3 score is potentially informative after a negative biopsy. Customized gene test targeting known genetic PCMs by Ambry Score, ProstateNext and so on may deliver a solid prediction for PCa possibility for an individual, especially when he has a PCa family history. However, validation studies across these methods among men, especially races, are variable and mostly focused on the Caucasian population, while African Americans are suffering from the highest PCa death rate. With more clinical data analyzed in the future and payer environment changes, and recommended PCM test map may be highly dynamic.

Cellular and molecular biology of DNA methylation

DNA methylation is a biological process catalyzed by DNA methyltransferases (DNMTs), by which methyl groups are directly added to the DNA molecule (*101*). While both cytosine and adenine can be methylated, methylation of cytosine to form 5-methylcytosine (5mC) is most common in both eukaryotes and prokaryotes. As one of the major events of epigenetics modification, DNA methylation plays a vital role in regulating

the activity of DNA segments without changing the sequences (*102, 103*). DNA methylation typically acts as a gene transcription inhibitor when located in a promoter region. In mammals, DNA methylation is essential for normal development and is associated with several key processes including carcinogenesis, repression of transposable elements, aging, genomic imprinting, and X-chromosome inactivation (*104, 105*) (Figure 6). In plants, DNA methylation can happen in 3 forms: CpG, CHG, or CHH. However, in the human genome, most majority of DNA methylation is on CpG dinucleotides, where non-CpG methylation usually can be detected in embryonic stem cells or during neural development (*106*).

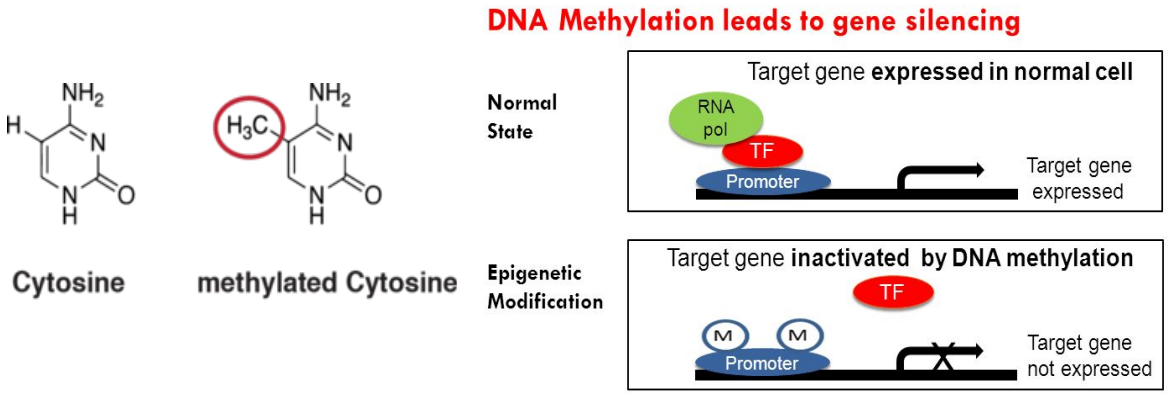


Figure 6. DNA methylation leads to gene silencing.

Left: Structures of cytosine and methylated Cytosine (5mC). Right: mechanism of DNA methylation leads to gene silencing: gene promoter region is blocked and prevents transcriptome factor binding, thus inhibits gene expression. Reprinted from teacher slides of UNC-Chapel Hill Superfund Research Program under Fair Use.

Moreover, CpG is far from randomly distributed in the human genome. Most methylated CpGs are on repetitive elements, such as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeats (LTRs), and other satellite DNA sequences. On the contrary, CpG islands which are identified as 1) length >200bp, usually shorter than 2kb; 2) GC content >50%, and 3) ratio greater than 0.6 of the observed number of CG dinucleotides to the expected number on the basis of the number of Gs and Cs in the segment [Obs/Exp CpG = Number of CpG * N / (Number of C * Number of G)] – have significant low methylation levels (in somatic tissues, ~10% of CpGs in CpG islands are methylated) (*107*). There are about 25,000 CpG islands in the human genome, of which 50% of them are located in gene promoter regions, 25% in gene bodies usually serving as alternative promoters. At the same time, from all ~30k human genes, 60-70% of them have at least 1 CpG island in their promoter region (*108, 109*).

DNA methylation levels are determined by methylation-demethylation balance (*110*). DMNT family controls the methylation process: DNMT1 has a robust affinity towards hemi-methylated DNA, thus is considered as the vital player for delivering DNA methylation patterns during DNA replication. DNMT1 also regulates de novo methylation patterns but only plays a minor role. DNMT3A and DNMT3B are two major DNA methyltransferases in charge of de novo methylation, as well as the co-factor DMNT3L. DNMT2 was considered as a DNA methyltransferase protein, but later identified as an aspartic acid tRNA methyltransferase and has been renamed tRNA aspartic acid methyltransferase 1 (TRDM1) (*111*). Research for DNA demethylation had a significant process during the last decade. TET family members serve as 5mC dioxygenases that

convert 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) (*112-114*).

DNA methylation in prostate cancer

DNA methylation has been found crucial in cancer development and progression in numerous studies (*115*). There are two major methylation alternations in cancer development and progression: locus-specific hypermethylation of tumor-suppressor genes (TSGs) and global hypomethylation (*116-119*). It is believed that the promoter region hypermethylation of TSGs will suppress their activities and global hypomethylation is assumed to induce carcinogenesis by activating oncogenes and increasing genomic instability (*119*). Numerous studies have indicated that PCa tumorigenesis is regulated by aberrant CpG methylation and those sites may serve as biomarkers for aggressive PCa (*120-127*). Recently there is more research focusing on whole genome PBL DNA methylation to discover markers of early detection, cancer risk, and risk factor exposure for various cancer types (*128-132*).

For example, using Illumina's HumanMethylation450 chip (the same platform to be used in this study, querying ~480,000 genome-wide CpG sites) (Figure 7), Florath et al. (*133*) identified a signature of age-related CpG methylation in whole blood DNA of 965 participants of a population-based cohort study. A regression model for age prediction based on 17 CpG sites as predicting variables explained 71% of the variance in age (*133*). ~480,000 selected genome-wide CpG sites methylation level can be measured with Illumina HumanMethylation450 chip platform. Heyn et al. studied CpG methylation profiles

of PBL DNA among 3 major populations (Caucasian, African, and Han-Chinese) and identified differentially methylated CpG sites between races. The host genes of those CpGs contribute to distinct phenotypes, such as response to drugs and environmental agents, and susceptibility to pathogens and diseases (*134*). This study demonstrated that using PBL CpG methylation as biomarkers has a biologically sound theoretical basis.

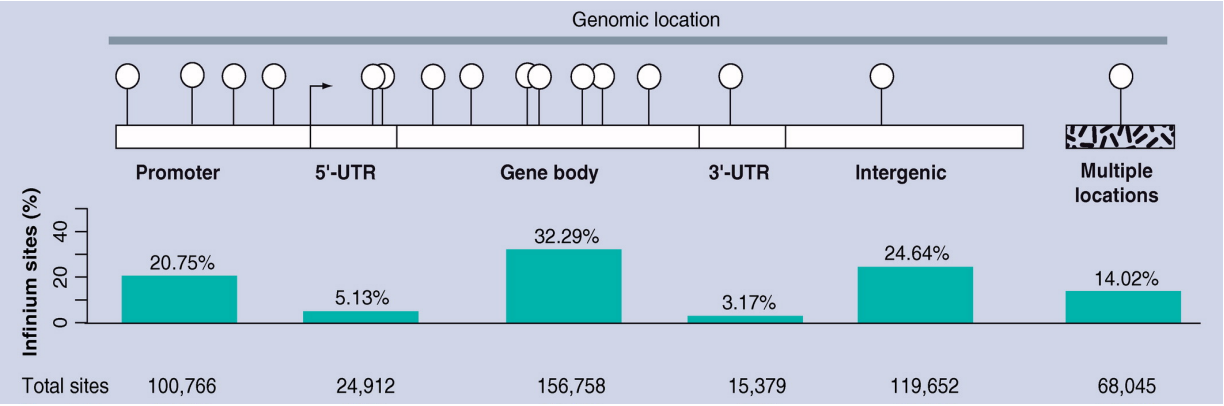


Figure 7. Probe distribution of Illumina Human Methylation 450K Arrays(*135*).

Among all 485,514 Illumina Human Methylation 450K Arrays probes, most of them are on the promoter (100,768 probes, 20.75%), gene body (156,758 probes, 32.29%), and intergenic (119,652 probes, 24.64%) regions. There are 68,045 probes (14.02%) on multiple locations, 24,912 probes (5.13%) on 5'UTR, and 15,379 probes (3.17%) on 3'UTR. Reprinted from Illumina Infinium® HumanMethylation450 BeadChip handbook under Fair Use.

WGBS is the most advanced technology for DNA methylation sequencing, which is a gold standard for DNA methylation measurement. Notably, studies showed that aberrant methylation patterns could be detected by WGBS in PCa (*120, 136-138*) (Figure 8).

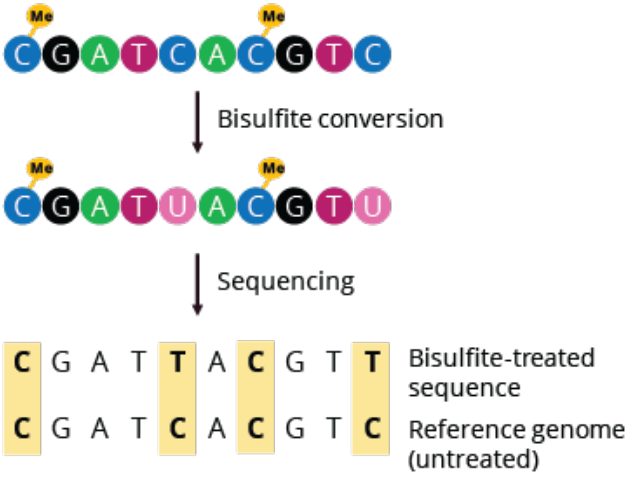


Figure 8. Whole Genome Bisulfite sequencing.

Treatment of DNA sequences with bisulfite converts cytosine residues to uracil (shown in blue), but 5-methylcytosines (shown in blue with marked yellow methyl group) will not be affected. Therefore, methylated cytosines throughout the whole genome can be detected when comparing with the reference sequences. Reprinted from Genewiz under Fair Use.

Both hypermethylation and hypomethylation can be observed in PCa, and abnormal DNA methylation status can lead to malignant carcinoma. Numerous studies suggested that DNA methylation alternations can be considered as early-stage biomarkers, which may help in diagnosis and prognosis.

CpG island hypermethylation and global hypomethylation in PCa

Many CpG islands are abnormally hypermethylated in PCa, thus inhibit some tumor suppressor gene expression (*139*). Major categories include cell cycle control, apoptosis, autophagy, DNA repair, cell adhesion, etc. (*140*). For example, DNA promoter methylation-

induced gene silencing has been studied in PCa. Among more than 50 altered genes, Glutathione S-transferase P1 (GSTP1) hypermethylation is common in PCA, with more than 90% of cases (*141*), and since GSTP1 is detectable in men blood, serum, urine, and plasma, it is further considered as a promising liquid biopsy biomarker (*121*).

In contrast to CpG island hypermethylation, global hypomethylation is observed in human tumors, including PCa. Researchers believe that global DNA hypomethylation is an event that takes place in the tumorigenesis stage of cancer, which is caused by genomic instability, activation of oncogenes, and removal of repression on retrotransposons (*142*). In PCa, associates were found between global hypomethylation and prognosis, tumor stage, and metastasis (*119, 120*).

Numerous studies have shown aberrant CpG site methylation in prostate tissues is involved in prostate tumorigenesis and may serve as biomarkers of aggressive PCa (*124, 143-148*). DNA methylation study has been predominantly performed in target tissues for the obvious reason that TSG activation is normally tissue-specific. However, recently, there has been increasing interest in using genome-wide CpG site methylation profiling (epigenome-wide association study) in peripheral blood DNA to identify markers of risk factor exposure, cancer risk, and early detection (*133, 134, 149-163*).

Biomarker identification in peripheral blood leukocyte (PBL) DNA

Epigenetic, especially methylation biomarkers have better stability in biofluids (plasma, urine, serum, saliva, etc.) or other samples, even if the DNA were extracted from accidentally thawed samples or dried blood spots (*164-167*). This character gives methylation biomarkers an advantage over other biomarkers, especially when the sample

has a below-average quality, Moreover, methylation biomarkers can also indicate target gene function due to their regulation roles and dynamic nature. Comparing to genetic- or protein-based biomarkers, methylation biomarkers may also inform about the effects of lifestyle and environment changes on human health and disease, thus unveil the nature of a disease that can be considered as natural bio-archives (*168*).

In recent days there are more studies focusing on whole genome PBL DNA methylation to discover markers of early detection, cancer risk, and risk factor exposure for various cancer types (*130*). For example, reports found that methylation of long interspersed nucleotide element 1 (LINE-1) (*128*), pericentromeric repeat NBL2, and subtelomeric repeat D4Z4 (*129*) altered in PCa patient blood.

Moreover, biopsy has risks for the patients: the incidence of infection complications can be up to 7% (*169*); about 4% of men are hospitalized following a prostate biopsy (*170*). Additionally, liquid biopsy has been reported as a cancer predictor by identifying single nucleotide polymorphisms (SNPs) from tDNA in patient blood (*171*). Although liquid biopsy DNA-based biomarkers studies are mostly targeting single nucleotide polymorphisms (SNPs), epigenetic changes could have a much larger effect on disease risk due to their profound functional impact on gene expression.

About 75% of men with elevated PSA levels do not have PCa or do not need treatment (*172*). Blood DNA-based biomarkers have the advantages of being minimally invasive, stable, and often powered by robust high-throughput technology. Therefore, there is a critical need to discover methylation signatures of PCa patient PBL through multiple sequencing platforms. Early detection of cancer is the best way to extend the chance of cure, thus increases the survival rate. However, developing an accurate and

non-invasive method for early detection is a challenging task: the radiation-based methods are accurate but invasive and expensive. The methods using protein markers in peripheral blood (such as PSA) are non-invasive but the sensitivity and specificity are not good enough. A promising breakthrough for early diagnosis is to put both PBL DNA and DNA methylation status into consideration. In this case, epigenetic biomarkers in patient PBL DNA can provide strong indications on disease risk evaluation, stratification, progression, and prognosis prediction.

CHAPTER II

EPIGENETIC MARKERS IDENTIFICATION IN PROSTATE CANCER PATIENT BLOOD*

Introduction

Prostate cancer (PCa) is the most common cancer and the second leading cause of cancer death in American men. There will be an estimated 248,530 new cases and 34,130 deaths from PCa in the United States in 2021 (*1*). PCa patients typically exhibit no symptoms until PCa becomes advanced or metastatic. The wide use of prostate-specific antigen (PSA) testing for screening and early detection has contributed to the greatly improved survival of PCa (*8*).

However, many PSA screening-detected PCa are indolent and pose little threat to the survival of patients. Commonly used clinical variables, including PSA level, Gleason score (GS), and tumor stage, are not sufficient to predict which patients will have aggressive diseases and which will have indolent diseases. Thus, the majority of men with localized PCa receive upfront aggressive treatments (radical prostatectomy and radio-therapy), which are often associated with significant side effects, causing overdiagnosis and overtreatment. Biomarkers that can predict aggressive diseases are needed to

---

*Part of this chapter is reprinted with permission from "Genome-wide DNA methylation profiling of leukocytes identifies CpG methylation signatures of aggressive prostate cancer" by Yuyan Han[#], Mutian Zhang[#], Junfeng Xu, Jia Li, Yifan Xu, Timothy C Thompson, Christopher J Logothetis, Deqiang Sun, Jian Gu. Copyright [2021] by all authors. # These authors contributed equally to this work.

improve the risk stratification of PCa patients for better-informed clinical management. Compared with other cancer types, genetic mutations are less common in PCa tumors.

Epigenetic changes including DNA methylation play a prominent role in prostate carcinogenesis and progression (*173*). Global hypomethylation and site-specific hypermethylation in promoter regions of tumor suppressor genes have been frequently observed in most cancers including PCa (*174, 175*).

Recently, there has been growing interest in using DNA methylation in peripheral blood leukocytes as predictors of cancer risks and clinical outcomes (*128, 176-186*). Specific CpG site methylation in leukocyte DNA has been shown to be associated with the risk of PCa (*184-186*) but limited study has systemically investigated the role of leukocyte DNA methylation in predicting aggressive PCa, although DNA methylation alterations are stable, occurred in early stage and can be detected reliably by many methods.

In this study, we performed a CpG methylation profiling in leukocyte DNA from a large number of PCa patients and identified specific leukocyte CpG methylation patterns among GS=6 and GS≥8 patients. We also compared and validated above results with WGBS data.

Materials and Methods

Study population

This study included 287 non-Hispanic white men with histologically confirmed adenocarcinoma of prostate from the University of Texas MD Anderson Cancer Center. Blood specimens were collected from the patients at diagnosis before any treatments. Clinical and follow-up data were abstracted from patient medical records by clinical coding specialists; these data included date of diagnosis, performance status, clinical stage, histological grade and pathological stage, treatment (active surveillance, prostatectomy, radiotherapy, and hormone therapy), and progression (biochemical recurrence and metastasis). The MD Anderson Tumor Registry conducts annual vital status follow-ups for all cancer patients. All patients signed an informed consent form. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Institutional Review Board of MD Anderson Cancer Center. We also included publicly available global DNA methylation data of healthy people (GSE85210) as the control group.

DNA extraction, bisulfite treatment and Illumina 450k beadchip

DNA was extracted with Qiagen mini kit (Qiagen, Germany) according to the manufacturer's protocol. One microgram of genomic DNA was treated with sodium bisulfite using the EZ DNA Methylation-Gold Kit (Zymo Research, Irvine, CA) according to the manufacturer's protocol. In order to minimize the batch effects, similar numbers of samples with GS=6 and GS≥8 were put on the same chip for hybridization. Briefly, whole genome

DNA methylation profiling was performed on 500 ng of bisulfite-treated DNA using the Illumina infimum Human Methylation 450k Beadchip (Illumina, Inc., San Diego, CA, USA) following standardized protocols and manufacturer's instructions. The 450k beadchip contains 485,577 cytosine positions in human genome, among which 365,934 CpG sites are located within known gene regions such as promoter, gene body or untranslated regions (UTRs), 119,830 are in intergenic regions (*187*). Beadchips were scanned on an Illumina HiScan SQ that has a two-color laser fluorescent scanner with a 0.375 um spatial resolution. The intensities of the images were extracted using Genome Studio Methylation Module.

Data analysis

Data analyses were performed with R version 3.4.3, Bioconductor packages, Chip Analysis Methylation Pipeline (ChAMP) (*188*), and bash scripts. Raw intensity data (.idat files) were organized as the initial loading files. The methylation status of each specific CpG site was shown as β-values, calculated as the ratio of the fluorescence intensity signals of the methylated (M) and unmethylated (U) alleles (*189*). β values range between 0 (non-methylated) and 1 (completely methylated). The probe detection p-value threshold was set as 0.01 and any samples showing a high fraction of failed probes (>0.05) were removed. Any probes with less than 3 detected beads in at least 5% of samples were also removed. Non-CpG probes also were removed. Y chromosomes were not ruled out since our dataset contains only male patients. Only one sample from GS=6 group was removed due to high percentage of failed probes. We also carried out normalization of our dataset

in order to remove the differences between type I and type II probe distributions with BMIQ method (*190*).

After normalization, we removed the batch effect caused by sample source. ChAMP called the differentially methylated probes (DMPs) using the corrected matrix of expression values with gene-wise linear models (Figure 9). A total of 10,264 DMPs were identified with FDR<0.05, and 1,459 DMPs with FDR<0.01 were selected as the input for further analysis. To estimate leukocyte subpopulations, we used ChAMP 450k reference databases for whole blood and performed regression method by Houseman et.al (*191*) to deconvolute cell populations for each blood cell type.

WGBS data were analyzed by Model based Analysis of Bisulfite Sequencing data (MOABS) toolkit developed by our lab (*192*). Raw data quality was evaluated by fastqc and multiqc. Qualified reads were mapped to hg38 reference genome with BSMAP module. Methylation calling was done by MCALL module, DMC and DMR identification were done by MCOMP module. Downstream analysis and plotting were performed with customized R, python scripts and Mmint.

.

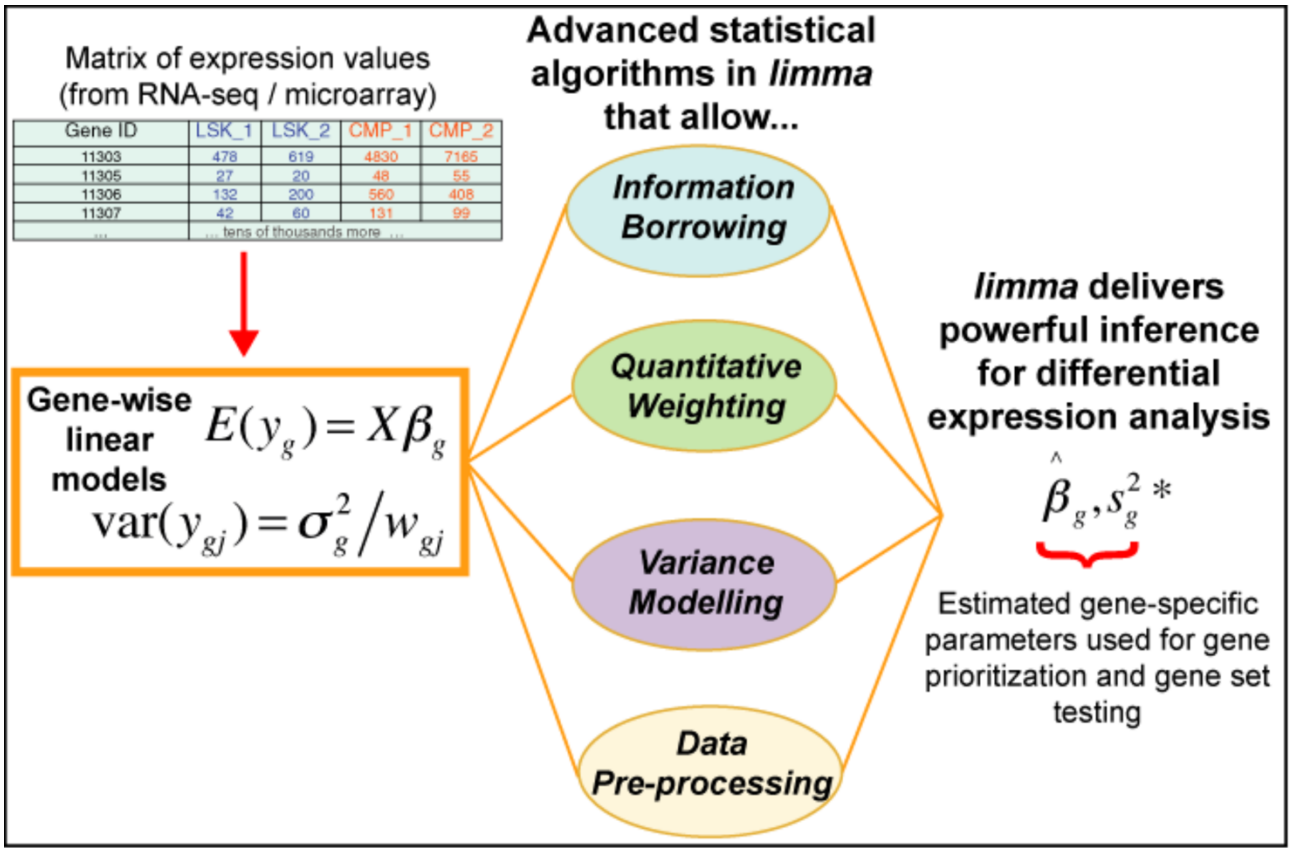


Figure 9. A linear model of ChAMP for 450k DNA microarray data analysis(*193*).

A 287x460k normalized β-value matrix were provided to ChAMP, and a gene-wise linear model was delivered by limma for powerful inference of differential expression analysis. Reprinted from Diboun et. al., “Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma”, 2006. Permission to reprint the figure for thesis was obtained from www.copyright.com with license ID 1122021-1.

Then we performed 5-fold cross-validated random forest model to identify the methylation signature that associates with GS (Figure 10, 11). Training set was determined randomly as 80% of the total dataset for each fold. Random forest trees were not pruned, and the number of trees was set as 400 to increase model power and also to

decrease the FDR. After the first model was trained, probe importance (Decrease Gini) was ranked for further model selection. We decided the best probe number for the random forest model based on the AUC of training and testing set and prediction accuracy.

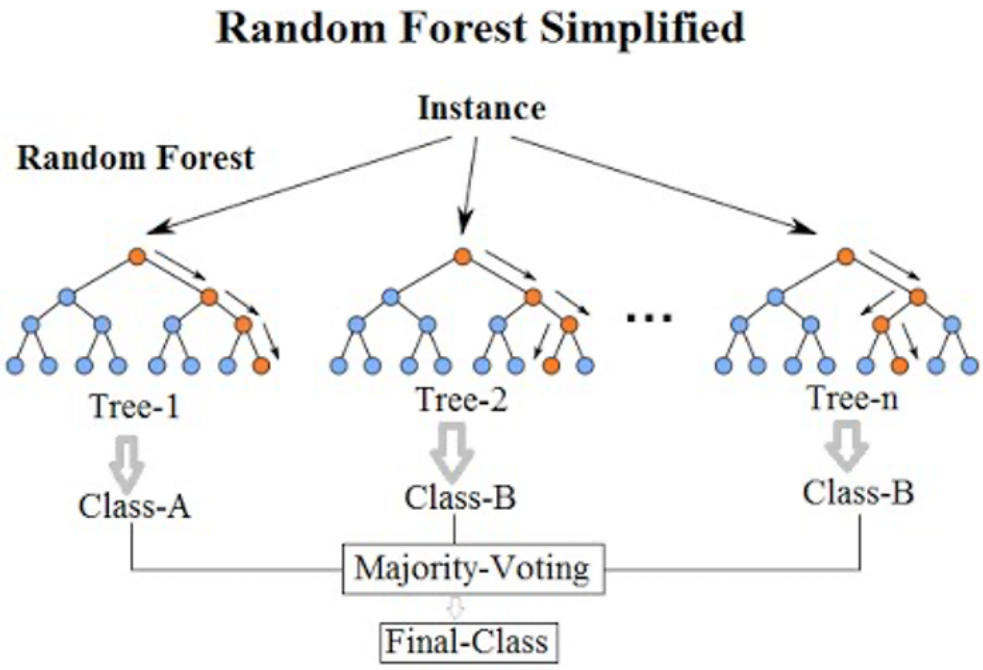


Figure 10. Schematic diagram for a random forest model(*194*).

Each classification decision tree in the forest gives an individual vote based on Gini impurity and information gain from each branch. The final-class prediction will be decided by majority voting. Reprinted from Will Koehrsen under Fair Use.

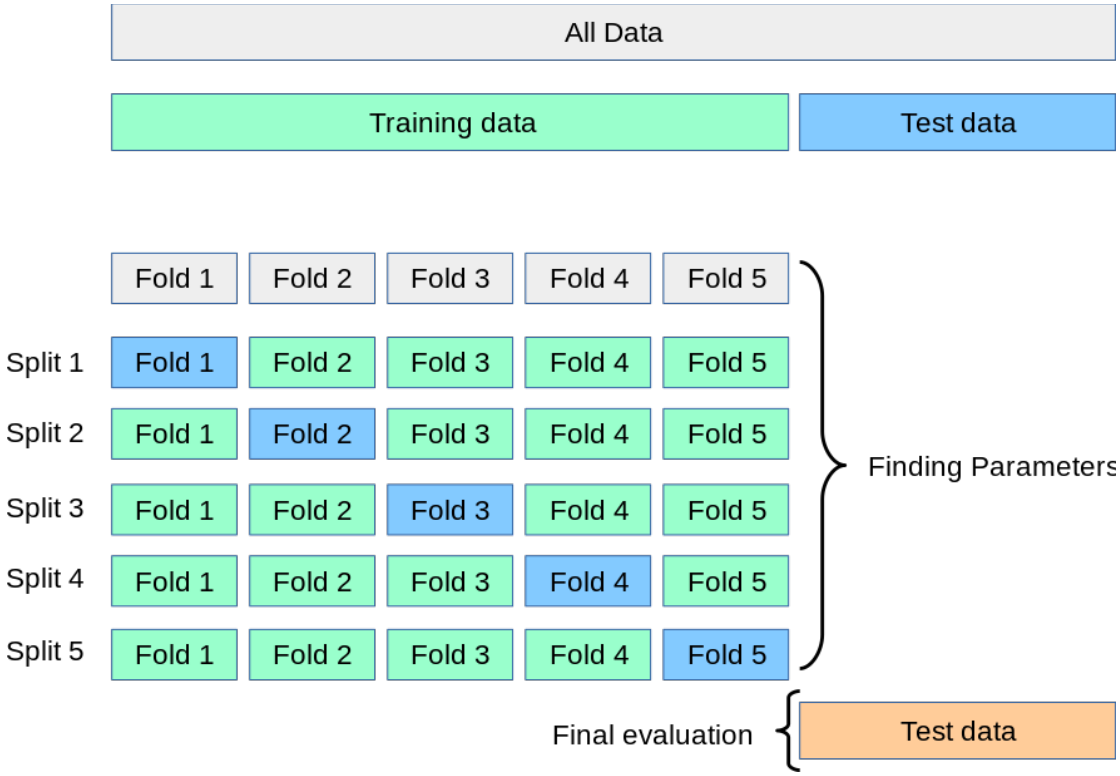


Figure 11. Schematic diagram for a k-fold cross validation model.

When k=5, all samples will be randomly assigned into 5 groups with the same sample size. For iteration 1, the group 1 serves as the testing group and rest of the 4 groups are training sets; group 2 will be the testing groups in iteration 2, and so on. The k-fold cross validation algorithm makes sure that all samples are at least in the testing group once. Reprinted from scikit-learn.org under Fair Use.

Results

Patient characteristics

We performed whole genome CpG methylation profiling in leukocyte DNA from 287 PCa patients with GS=6 and GS≥8. All patients were Caucasians. Most patients (85.3%) were 55 years and older. The mean ages (SD) of GS=6 and GS≥8 patients were 63.49 (SD: 5.46) and 63.68 (7.29), respectively. Only 7.8% were current smokers. About half were GS=6 (N=140) and half GS≥8 (N=147) patients. The patients had predominantly T1 (68.2%) tumors and had PSA<10 ng/ml (72.4%).

Leukocyte DNA methylation patterns in GS≥8 and GS=6 PCa patients

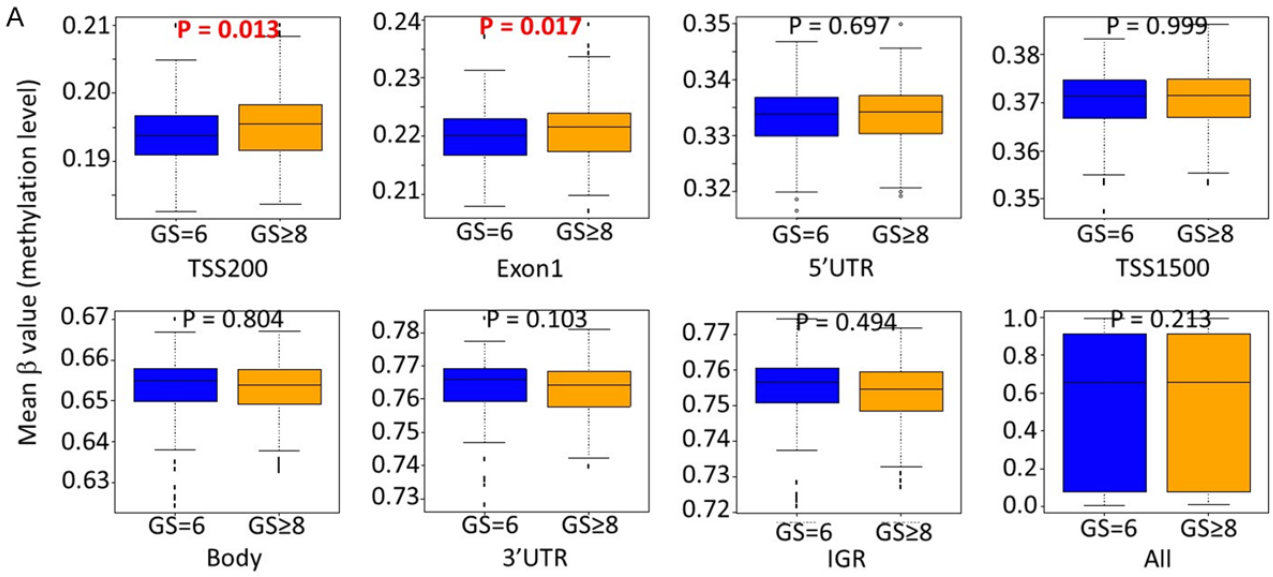
After normalization among all patients, a total of 464,867 cytosine positions in CpG dinucleotides on Human Methylation 450k BeadArray were analyzed. We first compared the global methylation level between GS≥8 and GS=6 patients. Although there were no significant differences in the overall global methylation level (mean β values of all the measured CpG sites) between GS≥8 and GS=6 patients, there were distinct methylation patterns among different functional regions (Figure 12A). The mean β value was the lowest in the core promoter region (TSS-200, within 200 base pairs of the transcription start site [TSS]), followed by Exon 1, 5' untranslated region (UTR), and TSS-1500 (within 1,500 base pairs of the TSS), all of which had mean β values between 0.18 and 0.40; whereas the mean β values of CpG sites located in the gene body, 3' UTR, and intergenic region (IGR) were much higher (0.63 to 0.78). More importantly, the mean β values of CpG

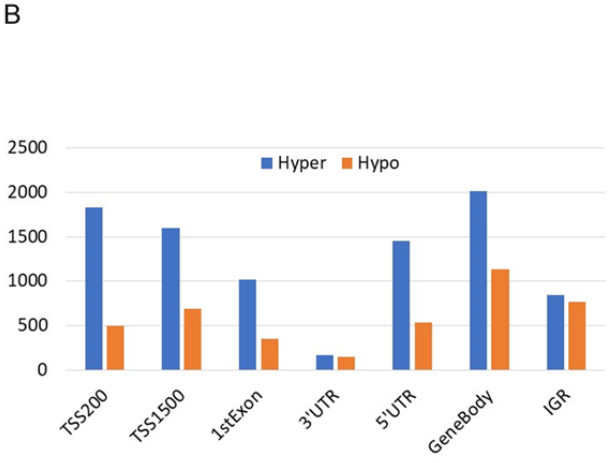
sites in TSS-200 and Exon 1 were significantly higher in GS≥8 patients than in GS=6 patients (P=0.013 and 0.017, respectively), whereas the methylation levels in gene body, 3' UTR, IGR, and overall methylation level (all) were higher in GS=6 than GS≥8 patients, although the difference did not reach statistical significance (Figure 13A).

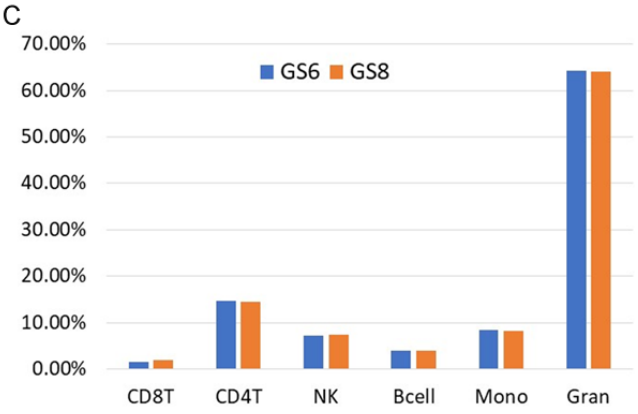
There were 10,264 differentially methylated CpG probes (DMPs) between GS≥8 and GS=6 patients with FDR<0.05, among which 6,876 were hypermethylated and 3,389 were hypomethylated in GS≥8 compared to GS=6 patients. In a breakdown of significant hypermethylated and hypomethylated CpG sites by CpG locations, there were significantly more hypermethylated than hypomethylated CpG sites in transcriptionally active regions, in particular, TSS200, Exon 1, 5' UTR, and TSS1500, whereas the numbers of significantly hypermethylated and hypomethylated CpG sites were similar in 3' UTR and IGR (Figure 12B).

Among 6,876 hypermethylated CpG sites in GS≥8 patients, 3,771 were located in CpG islands. Since hypermethylation in CpG islands is more likely to affect host gene expression, we performed gene set enrich analysis (GSEA) using host genes of these 3,771 DMPs. The top enriched pathways included RNA-binding, enzyme-biding, ribonucleotide binding, and regulation of gene expression.

Leukocyte DNA methylation can be used to quantify different leukocyte subproportions (*191, 195*). We estimated the frequencies of B cell, CD8+ and CD4+ T cell, natural killer cell, granulocytes, and monocytes using methylation profiles (Figure 12C). The frequencies of major leukocyte subpopulations were similar between GS≥8 and GS=6 patients, which indicates the leukocyte methylation differences between GS≥8 and GS=6 is not likely due to different immune cell compositions.







	CD8T	CD4T	NK	Bcell	Mono	Gran
GS=6	1.42%	14.71%	7.25%	3.95%	8.40%	64.25%
GS≥8	1.87%	14.50%	7.30%	3.89%	8.29%	64.15%
P value	0.978	0.965	0.244	0.271	0.850	0.444

Figure 12. Overall leukocyte DNA hypermethylation in transcriptionally active regions in GS⩾8 patients compared to GS=6 patients.

A. Comparisons of mean β value of CpG sites by locations of CpG sites relative to gene structure; B. Comparisons of the total numbers of significantly hypermethylated and hypomethylated CpG sites by locations of CpG sites relative to gene structure. Y-axis shows the number of differentially methylated probes; C. Comparisons of the frequencies of major leukocyte subpopulations between GS≥8 and GS=6 patients. Y-axis show the proportion of each cell type. Abbreviations: TSS200: within 200 bp of the transcription start site (TSS); TSS1500: within 1500 bp of the TSS; UTR: untranslated region; IGR: intergenic regions. Student's T-tests were performed for each comparison.

A leukocyte CpG methylation signature for predicting aggressive PCa

To identify a CpG methylation signature that distinguishes GS≥8 from GS=6 patients, we used the normalized β value of 1,459 CpG sites with FDR<0.01 as input to train the 5-fold cross validated random forest model. The testing set AUC was 0.836 and prediction accuracy was 0.757. After ranking the probes with their contribution to the model (decreasing Gini), we improved the model by training the model with fewer top-ranked DMPs. When we used the top 10 differentially methylated DMPs, the prediction reached 80% and additional DMPs only modestly increased the prediction accuracy, up to 85% (Figure 13A). For the final model with the top 97 DMPs, the testing AUC was 0.920, and predicting accuracy was 0.847 (Figure 13B). The Multidimensional Scaling (MDS) plot indicated a strong ability of our model to cluster patients (Figure 13C). Figure 13D shows the heatmap of using those 97 CpG sites to group GS≥8 from GS=6 patients and there was a clear separation of these two groups. The characteristics of the top 97 differentially methylated CpG sites between GS=6 and GS≥8 patients are shown in Table 1.

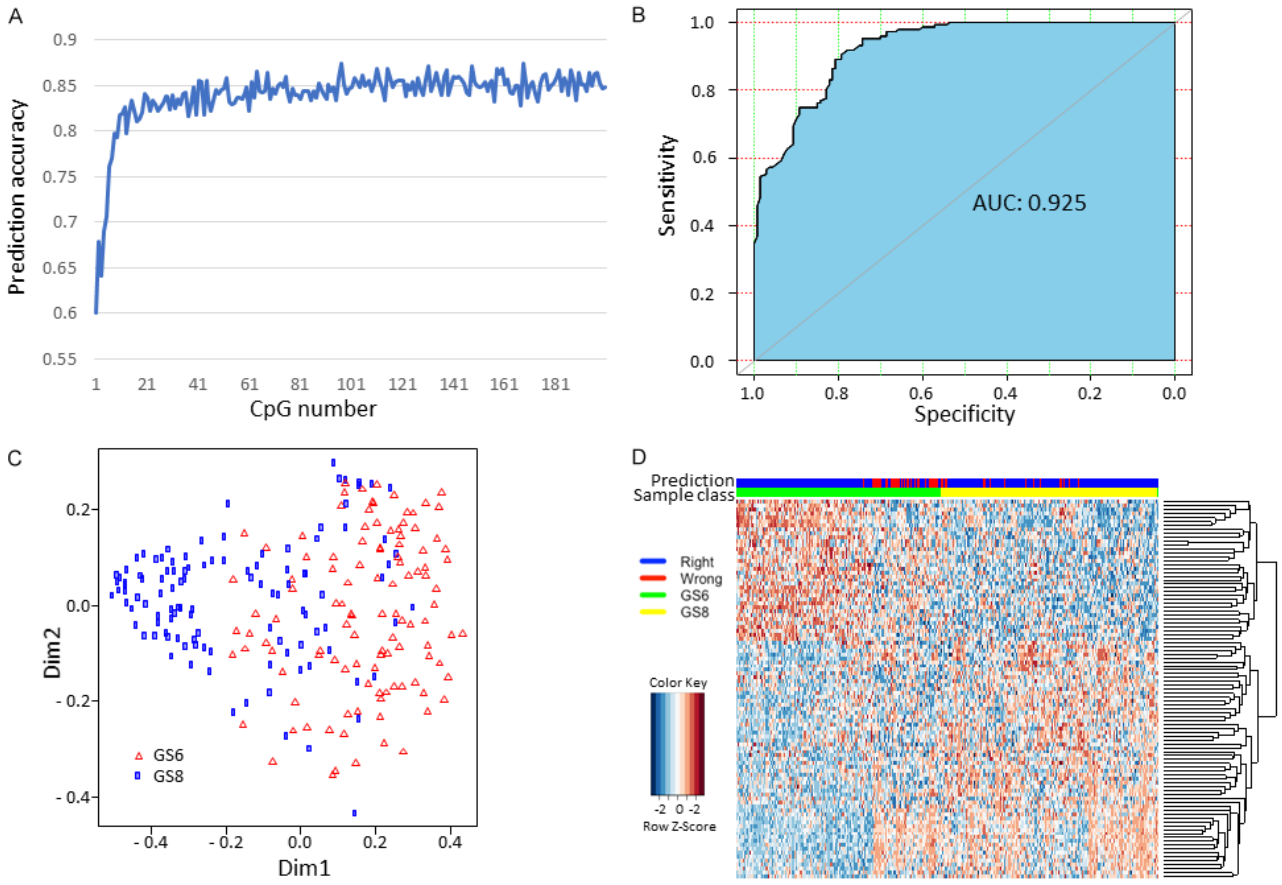


Figure 13. Leukocyte DNA methylation signature that differentiates GS≥8 patients from GS=6 patients.

A. Prediction accuracy based on the number of differentially methylated CpG probes (DMPs); B. The ROC and AUC of prediction model using top 97 DMPs; C. Multidimensional Scaling (MDS) plot indicating the ability of the model to cluster patients; D. Supervised clustering of GS≥8 and GS=6 patients.

Table 4.Top 97 differentially methylated CpG sites between GS=6 and GS≥8 patients

CpG ID	β value		*P* value	Chr	Position	Gene	CpG Location
	GS=6	GS≥8					
cg00111102	0.9175	0.9328	1.73E-05	20	60509975	CDH4	Body-shore
cg00216361	0.0491	0.0543	1.96E-05	3	115342527	GAP43	1st Exon-open sea
cg00419564	0.0223	0.0266	4.21E-07	1	153508860	S100A6	TSS200-shore
cg00567696	0.0492	0.0564	2.37E-06	6	46097521	ENPP4	TSS200-shore
cg00619978	0.527	0.565	4.13E-06	5	180046052	FLT4	Body-island
cg00843795	0.6578	0.7677	1.32E-05	7	105163736	PUS7	TSS1500-shore
cg00850868	0.6365	0.6548	9.03E-09	10	64437920	IGR	NA
cg01071346	0.0294	0.0335	5.61E-06	1	2480431	IGR	chr1:2477563-2478363
cg01077623	0.7255	0.7055	2.16E-05	7	55757733	FKBP9L	TSS1500-open sea
cg01466348	0.9389	0.9245	5.40E-07	2	161503843	IGR	NA
cg01890546	0.9143	0.9227	1.64E-06	7	884588	UNC84A	Body-shelf
cg02005490	0.9271	0.915	6.24E-06	5	1959850	IGR	NA
cg02048674	0.0437	0.0505	2.24E-06	19	49991517	RPL13AP5	Body-island
cg02383160	0.028	0.0331	1.23E-07	11	62496393	TTC9C	1st Exon-shore
cg02895995	0.0641	0.0741	4.94E-08	19	7554069	PEX11G	TSS200-shore
cg03014008	0.597	0.6156	1.93E-07	20	57463767	GNAS	3’ UTR-island
cg03354554	0.2366	0.2169	2.08E-05	11	9781412	IGR	chr11:9779592-9780470
cg03414732	0.0713	0.0597	3.06E-09	18	32870301	ZNF271	Body-island

cg04208114	0.0454	0.0554	4.16E-08	1	59012469	OMA1	TSS200-island
cg04250904	0.6706	0.6485	2.25E-07	19	12623422	ZNF709	5’ UTR-shore
cg04442328	0.1526	0.1675	1.34E-05	3	185304136	SENP2	1st Exon-island
cg04913913	0.1189	0.107	7.40E-07	6	31126599	CCHCR1	TSS1500-shore
cg05176964	0.9753	0.9701	3.67E-06	22	42910177	RRP7A	Body-island
cg06295548	0.8495	0.8741	3.28E-08	4	146296778	IGR	NA
cg06434972	0.0371	0.0426	3.48E-07	7	44122219	POLM	TSS200-island
cg06834240	0.1333	0.1507	4.07E-08	16	79632625	MAF	3’ UTR-island
cg07374247	0.0157	0.0187	3.00E-08	6	27860935	HIST1H2AM	1st Exon-shore
cg07872947	0.9402	0.9481	1.47E-06	2	1732172	PXDN	Body-open sea
cg08005992	0.108	0.1219	7.13E-06	11	31832959	PAX6	TSS200-island
cg08907257	0.8877	0.8974	1.34E-05	16	2223188	TRAF7	Body-shelf
cg09618381	0.8791	0.8566	6.12E-07	6	150379479	IGR	chr6:150378838-150379048
cg09910998	0.5714	0.581	1.18E-06	7	94285942	SGCE	TSS1500-island
cg10149161	0.0831	0.068	3.18E-10	11	64578067	MEN1	TSS200-island
cg10438391	0.2942	0.2522	2.42E-05	8	144631915	IGR	chr8:144631767-144631971
cg10446143	0.0558	0.0626	2.16E-05	21	44394730	PKNOX1	5’ UTR-island
cg10632144	0.9276	0.8967	1.27E-05	13	50252564	EBPL	Body-open sea
cg10797195	0.0457	0.0514	4.70E-06	1	45805338	MUTYH	5’ UTR-shore
cg10919522	0.2547	0.233	1.84E-05	14	74227441	C14orf43	5’ UTR-shore

cg11214243	0.0373	0.0424	2.36E-07	11	65405388	SIPA1	TSS200-shelf
cg11678250	0.49	0.4546	3.31E-06	7	136362483	IGR	NA
cg11956953	0.0881	0.0726	1.85E-05	17	27347092	IGR	chr17:27346853-27347222
cg12791243	0.0682	0.0609	1.01E-06	4	79698201	BMP2K	Body-shore
cg13038108	0.0267	0.0317	4.34E-07	4	39461155	LIAS	Body-shore
cg13785223	0.4713	0.4286	4.20E-06	13	114905788	IGR	NA
cg14235800	0.8887	0.8982	2.01E-05	9	104238593	C9orf125	Body-open sea
cg14323199	0.0517	0.0572	3.99E-06	17	60705839	MRC2	Body-island
cg14416269	0.2198	0.1914	4.71E-06	4	6271139	WFS1	TSS1500-shore
cg14420670	0.0328	0.0382	8.18E-08	6	29617961	IGR	chr6:29617765-29617974
cg14951488	0.1623	0.1538	1.73E-05	10	95256188	CEP55	TSS200-island
cg15248835	0.0413	0.0485	1.76E-05	8	9761171	LOC157627	TSS1500-island
cg15354625	0.9302	0.9372	4.06E-06	11	78381223	ODZ4	Body-open sea
cg15404375	0.9397	0.9458	2.02E-05	4	111866546	IGR	NA
cg15731816	0.0361	0.0408	6.70E-07	14	75230414	YLPM1	1st Exon-island
cg15896939	0.928	0.9352	1.35E-05	1	156030809	RAB25	TSS200-opensea
cg15935247	0.7404	0.7189	2.31E-06	17	56606842	4-Sep	TSS200-shelf
cg16311364	0.5388	0.4808	1.07E-06	10	46912902	FAM35B	Body-shore
cg16374753	0.962	0.9682	2.27E-06	X	79279642	TBX22	Body-open sea
cg16513883	0.8804	0.8923	2.90E-05	5	9295286	SEMA5A	Body-open sea

cg16619899	0.8632	0.8492	6.18E-06	8	915860	IGR	chr8:914817-915894
cg16925090	0.0692	0.079	4.46E-06	11	101785516	KIAA1377	TSS1500-shore
cg17098965	0.3396	0.3067	3.26E-07	20	52199520	ZNF217	5’ UTR-shore
cg17329834	0.8211	0.7991	4.98E-06	6	131380543	EPB41L2	5’ UTR-shelf
cg17392909	0.2192	0.251	4.35E-09	10	135187035	ECHS1	TSS200-island
cg17524854	0.0458	0.0515	8.21E-08	12	67663046	CAND1	TSS200-island
cg18050997	0.8858	0.897	6.34E-08	8	8176225	PRAGMIN	Body-island
cg18483322	0.0772	0.0845	3.01E-06	2	97523826	ANKRD39	TSS200-island
cg18651347	0.8273	0.8092	2.18E-05	7	70102632	AUTS2	Body-open sea
cg18725195	0.6238	0.655	5.79E-07	5	976058	IGR	NA
cg18943383	0.0329	0.0402	6.49E-09	6	27777858	HIST1H3H	1st Exon-island
cg19239278	0.781	0.7583	2.98E-06	19	19513162	GATAD2A	5’ UTR-shelf
cg19242459	0.9213	0.9302	1.99E-08	2	239006511	SCLY	Body-shelf
cg19757631	0.872	0.8529	6.65E-06	1	11118889	SRM	Body-shore
cg19864851	0.0275	0.0333	1.48E-08	10	75503847	SEC24C	TSS1500-shore
cg20153768	0.0334	0.0376	2.42E-05	6	26123228	HIST1H2AC	TSS1500-shore
cg20390613	0.0432	0.0513	1.86E-07	1	12678355	DHRS3	TSS1500-island
cg20539816	0.9396	0.9334	4.13E-07	17	5988249	WSCD1	Body-open sea
cg21636841	0.8416	0.854	1.43E-05	11	968731	AP2A2	Body-open sea
cg22028624	0.8741	0.8513	1.01E-07	11	70281091	CTTN	Body-open sea

cg22110517	0.9086	0.9168	4.36E-06	17	4800583	MINK1	3’ UTR-shore
cg22407822	0.657	0.679	7.54E-08	20	57463658	GNAS	3’ UTR-island
cg22716488	0.0387	0.0437	2.15E-06	6	35995431	MAPK14	TSS200-island
cg22826071	0.0364	0.0452	2.91E-08	19	344165	MIER2	Body-island
cg22961241	0.8822	0.8999	1.56E-06	6	32917502	HLA-DMA	Body-open sea
cg23496597	0.6552	0.6715	1.23E-06	20	57463725	GNAS	3’ UTR-island
cg23983453	0.4253	0.4789	8.55E-06	5	92925524	NR2F1	Body-shore
cg24337701	0.6228	0.593	1.87E-05	8	141275191	TRAPPC9	Body-open sea
cg24751378	0.7014	0.7205	2.26E-06	21	30396349	USP16	TSS1500-shore
cg25079743	0.0262	0.0311	7.23E-07	16	30441674	DCTPP1	TSS1500-island
cg25198967	0.8953	0.9066	7.74E-08	3	52325846	GLYCTK	Body-shelf
cg25554036	0.2521	0.2127	7.34E-07	4	6271136	WFS1	TSS1500-shore
cg25696807	0.7097	0.6766	6.39E-07	X	145109374	MIR891A	Body-open sea
cg25697492	0.1212	0.111	1.54E-05	19	2950919	IGR	chr19:2950359-2950962
cg25748441	0.0329	0.0383	2.47E-05	2	202122587	CASP8	5’ UTR-open sea
cg25806190	0.8185	0.7973	3.24E-06	2	232878174	DIS3L2	5’ UTR-open sea
cg26127025	0.9343	0.9201	1.11E-06	5	2703138	IGR	NA
cg26683137	0.3639	0.3234	7.98E-06	17	33447208	FNDC8	TSS1500-shore
cg27482619	0.0574	0.0663	7.74E-11	10	30818479	IGR	NA

Table 5. Selected patients' characteristics by training and testing set.

	TRAINING	TESTING	p-value
Age			0.615
45-54	22	20	
55-64	56	67	
65-74	57	50	
75-84	8	6	
Smoke			0.269
Never smoker	56	69	
Former smoker	76	60	
Current smoker	10	12	
unknown	1	2	
Gleason Score			0.999
<=6	70	70	
>=8	73	74	
Stage			0.57
T1	99	96	
T2	7	12	
T3	34	30	
T4	3	5	
PSA at diagnosis			0.471
<10 ng/ml	108	99	
>=10 and <20ng/ml	15	20	
>=20 ng/ml	20	24	
Biochemical recurrence			0.882
No	114	116	
Yes	29	27	

Re-classification of PCa patients with methylation biomarkers

As discussed before, Gleason Score alone is not accurate in predicting PCa aggressiveness. Here we demonstrated re-classification results of PCa patients from GS=6 and GS≥8 groups using top-97 CpG probes identified from the last section. With hierarchical clustering algorithm provided by hclust (*196*), we re-classified patients into 3 risk groups: 79 low risk, 62 mid risk, and 147 high risk. Among the low risk group, 93.6% patients were GS=6; in the mid risk group, 43.5% were GS=6 patients; in the high risk group, 25.9% were GS=6 patients (Figure 14). These results indicated that although most majority of GS≥8 patients would have poor prognoses, about 25% of them had similar the methylation pattern in the 97 CpG biomarkers. In fact, only 7 out of 38 (18.4%) low or mid risk GS≥8 patients had BCR, which was significantly lower comparing with high risk GS≥8 groups (41 out of 109, 37.6%. chi-squire p-value=0.048).

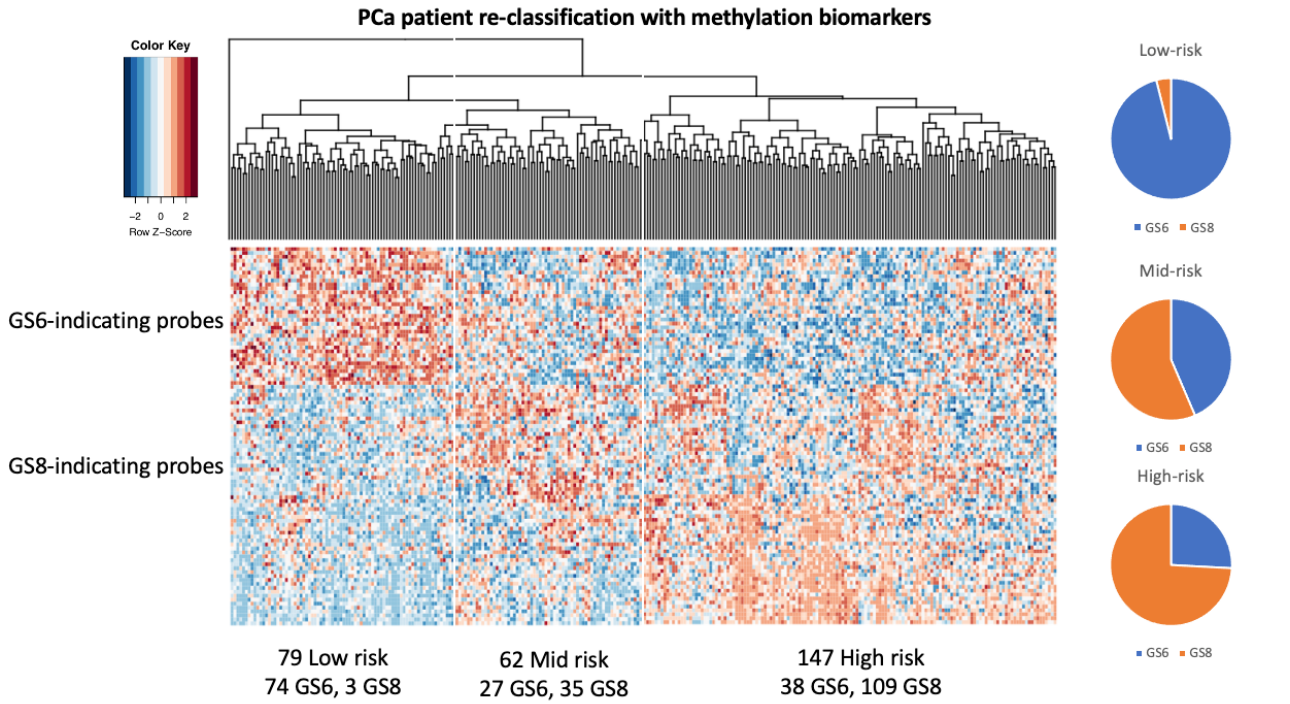


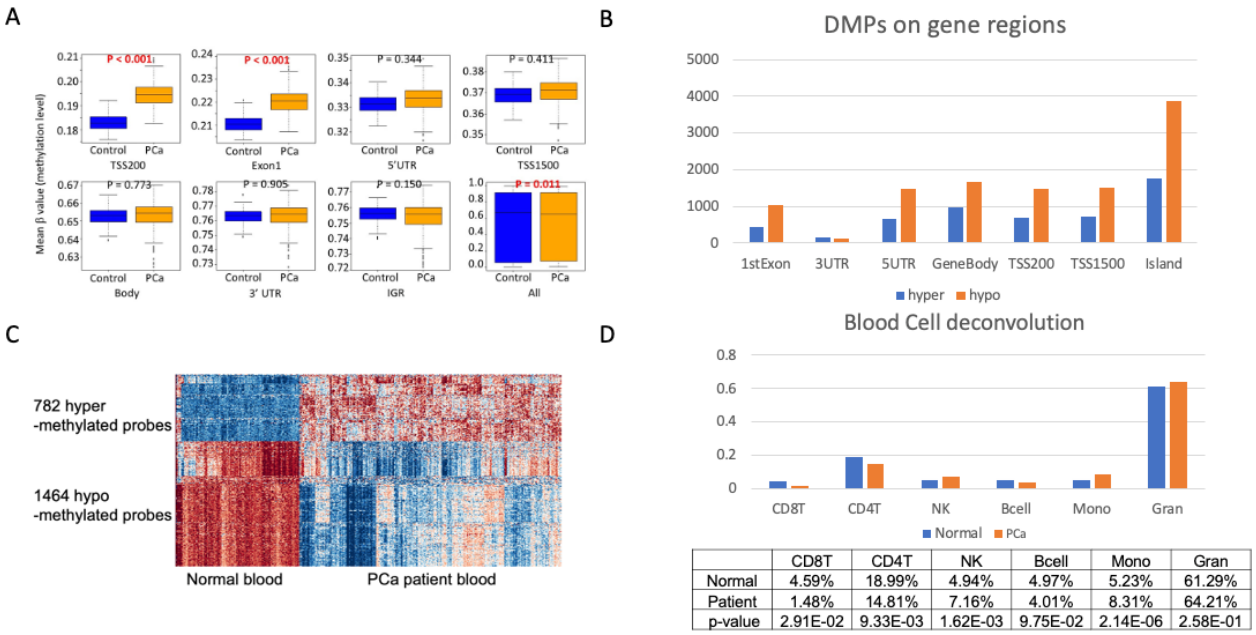
Figure 14. Re-classification of GS=6 and GS⩾8 patients with 97 biomarkers.

GS=6 and GS≥8 patients were re-clustered with hierarchical clustering algorithm. Based on the cluster result, we classified GS=6 and GS≥8 patients into 3 risk groups: 79 low-risk, 62 mid-risk, and 147 high-risk.

Comparison of leukocyte DNA methylation between PCa patients and healthy controls

We compared our data with a publicly available leukocyte 450K methylation dataset of healthy controls (GSE85210). There were 172 healthy men in the dataset. The mean β values of all the CpG sites were significantly lower in PCa patients than in healthy controls (P=0.011), indicating global hypomethylation of PCa patients. However, the mean β values of CpG sites in TSS-200 and Exon 1 regions were significantly higher in PCa patients than in healthy controls (P<0.001 for both) (Figure 15A). There were 15,941 differentially methylated CpG probes (DMPs) between PCa patients and normal men with FDR<0.05,

among which 6,232 were hypermethylated and 9,709 were hypomethylated in PCa patients compared to normal men, which again indicated a global hypomethylation in PCa patients. In a breakdown of significant hypermethylated and hypomethylated CpG sites by CpG locations, there were significantly more hypomethylated than hypermethylated CpG sites in transcriptionally active regions, in particular, TSS200, Exon 1, 5' UTR, CpG Islands, and TSS1500, whereas the numbers of significantly hypermethylated and hypomethylated CpG sites were similar in 3' UTR (Figure 15B). Heatmap of top 762 hypomethylated and 1464 hypomethylated probes indicated subgroups may exist within the patient group (Figure 15C). We estimated the frequencies of B cell, CD8+ and CD4+ T cell, natural killer cell, granulocytes, and monocytes using methylation profiles (Figure 15D). Unlike GS≥8 and GS=6 patients, the frequencies of major leukocyte subpopulations were significant different between PCa patients and normal men: in PCa patient blood, the amounts of CD8+ T cells (1.48% versus 4.59%, p-value = 2.9E-02) and CD4+ T cells (14.81% versus 18.99%, p-value = 9.3E-03) were significantly decreased, while the amounts of natural killer cells (7.16% versus 4.94%, p-value = 1.6E-03) and monocytes (8.31% versus 5.23%, p-value = 2.1E-06) were significantly increased. A report showed that the proportion of CD4+ T cells was lower in metastasis cervical cancer patients (*197*). Although increasing proportions of monocytes and natural killer cells indicating higher non-specific immune system activity, a lower proportion of CD8+ T and CD4+ T cells also inferred to impaired ability to remove specific PCa cells from the patient immune system.



	CD8T	CD4T	NK	Bcell	Mono	Gran
Normal	4.59%	18.99%	4.94%	4.97%	5.23%	61.29%
Patient	1.48%	14.81%	7.16%	4.01%	8.31%	64.21%
p-value	2.91E-02	9.33E-03	1.62E-03	9.75E-02	2.14E-06	2.58E-01

Figure 15. Overall leukocyte DNA hypermethylation in transcriptionally active regions in PCa patients compared to normal men.

A. Comparisons of mean β value of CpG sites by locations of CpG sites relative to gene structure; B. Comparisons of the total numbers of significantly hypermethylated and hypomethylated CpG sites by locations of CpG sites relative to gene structure; C. Heatmap of top 762 hypomethylated and 1464 hypomethylated probes between normal and PCa patient; D. Comparisons of the frequencies of major leukocyte subpopulations between GS≥8 and GS=6 patients. Abbreviations: TSS200: within 200 bp of the transcription start site (TSS); TSS1500: within 1500 bp of the TSS; UTR: untranslated region; IGR: intergenic regions. Student's T-tests were performed for each comparison.

Consistency between WGBS and 450k DNA microarray results

To understand the whole genome methylation profile for PCa patients, we randomly selected 4 samples from both GS=6 and GS≥8 patient pools and performed whole

genome bisulfite sequencing (WGBS). Similar to 450k DNA microarray results, we observed global hypomethylation in WGBS data as well: we detected 26,668 differentially hypermethylated CpGs (DMCs) and 317 hypermethylated regions (DMRs), while there were 38,468 hypomethylated DMCs and 564 DMRs. We also found a robust correlation for CpG methylation levels across 450k DNA microarray and WGBS, in both all mutant CpG sites and differentially methylated locations (Figure 16A). Visualization of differentially methylated regions (DMRs) and GSEA pathway enrichment analysis further validated our findings (Figure 16B and 16C). Finally, we filtered mutually differentially methylated CpG sites with FDR<0.001 and methylation difference (absolute difference >0.05 for DNA microarray, credible difference >0.3 for WGBS) and discovered 7 mutual most differentially methylated CpG sites: cg01539474 (H19 TSS region), cg01883208 (PCNXL2 gene body), cg02895509 (Intergenic region), cg07240043 (GNPAT gene body), cg13917504 (MEST TSS region), cg22860172 (PRDM16 gene body), and cg23496597 (GNAS 3‘UTR region) (Figure 16D). H19 SNPs were reported to correlate with aggressive PCa (*198*); deletion of PCNLX2 was reported in PCa (*199*); Based on data from the Human Protein Atlas, low expression of GNPAT in renal cancer patients has a significant lower survival rate; a study showed that PEG1/MEST expression was altered along with PCa development (*200*); PRDM16 was found overexpressed in PCa cell lines and associated with PCa invasion (*201*); Multiple studies demonstrated that GNAS played vital roles in PCa, especially in metastatic castration-resistant PCa (*202-204*).

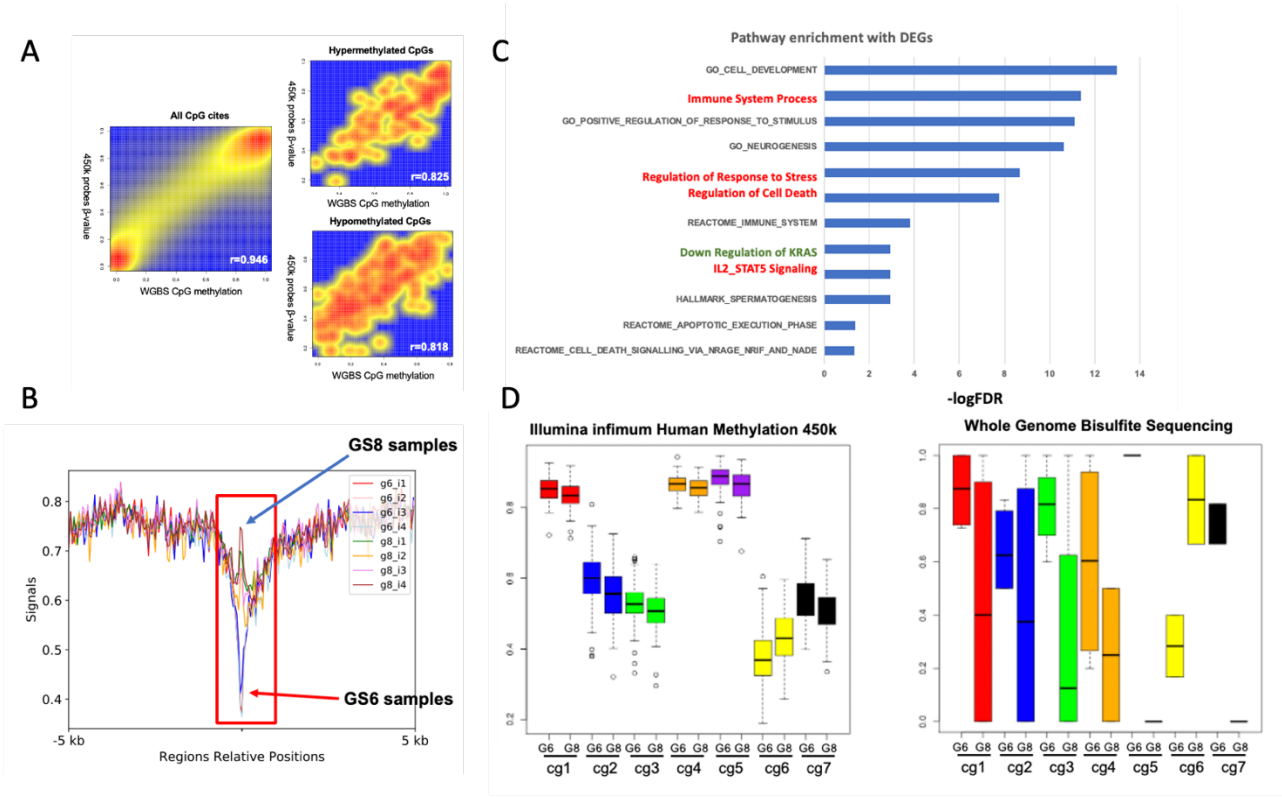


Figure 16. Crosstalk between WGBS and 450k DNA microarray results.

A. Correlation of the CpG sites' methylation level measured using WGBS (sequencing depth >=5) and the CpG sites' β-values measured by illumine 450k DNA methylation microarray. Correlation coefficient was shown as Pearson product-moment correlation coefficient. R=0.946 for all 238,722 filtered CpG sites, 0.825 for 482 mutual hypomethylated CpG sites, and 0.818 for 821 hypomethylated CpG sites. B. GSEA pathway analysis for WGBS promoter DMR related genes. Immune system, stress response, KRAS, IL2-STAT5 pathways were detected. C. Profile of differentially hypermethylated regions detected in WGBS samples. D. Boxplots of 7 most mutual differentially methylated CpG sites with both WGBS and 450k DNA microarray.

Conclusion and Discussion

The main purpose of this study was to identify intrinsic biological differences between clinically defined non-aggressive (GS=6) and aggressive (GS≥8) that may serve as predictors of aggressive PCa. We performed a genome-wide CpG methylation profiling of leukocyte DNA from 287 PCa patients with GS=6 and GS≥8. We found leukocyte DNA hypermethylation in transcriptionally active regions in aggressive PCa patients and identified a 97-CpG signature that could distinguish aggressive from non-aggressive PCa. To our knowledge, this is the first study to report leukocyte CpG methylation signature for the prediction of aggressive PCa.

We found the mean DNA methylation level was the lowest in the core promoter region (TSS-200), followed by Exon 1, 5' UTR, and TSS-1500, but considerably higher in the gene body, 3' UTR, and intergenic regions, which is consistent with literature reports of low methylation in the transcriptionally active regions, indicating open chromatin structure (*195*). More importantly, we observed hypermethylation of leukocyte DNA in GS≥8 patients compared to GS=6 patients in the most transcriptionally active regions (TSS200 and Exon 1). In gene set enrichment analysis, the top enriched pathways included RNA-binding, enzyme-binding, ribonucleotide binding, and regulation of gene expression. These findings indicate an overall down-regulation of gene expression in leukocytes of GS≥8 patients, likely affecting inflammatory response and immune function and contributing to the aggressive phenotypes. Likewise, when we used a publicly available dataset of leukocyte DNA methylation in healthy men and compared it to the data in our PCa patients, we observed hypermethylation of leukocyte DNA in PCa patients in the most transcriptionally

active regions (TSS200 and Exon 1), supporting an overall down-regulation of gene expression in leukocytes of PCa patients, particularly aggressive PCa, that affects inflammatory response and immune function and contributes to PCa development and progression. We also observed overall lower methylation of leukocyte DNA in PCa patients compared to healthy individuals, and in GS≥8 than in GS=6 patients. Global hypomethylation in tumor tissues is well-established cancer-promoting event (*119, 174, 175*). It has also been hypothesized that global DNA hypomethylation in leukocytes may be a cancer risk factor due to increased genomic instability (*119, 177*). There are some supporting evidence for this notion, but the data were not consistent (*119, 177*). Previous studies evaluating global DNA methylation and cancer risks mostly used methylation of short repetitive DNA sequences (e.g., LINE-1 and Alu) as surrogates to represent global DNA methylation level. In our study, we used the mean β value of all the assayed CpG sites, which provides a more accurate estimate of global DNA methylation level. Our data support the notion that global hypomethylation in leukocyte DNA contributes to the development and progression of PCa likely through general genomic instability.

Leukocyte DNA methylation is at the interphase between genetics and environment. It is under a strong influence of genetics and also has been linked to immune cell subpopulation, aging, and smoking. We did not observe significant differences in the immune cell subpopulations between GS=6 and GS≥8 patients, indicating that there were minimal immune cell turnovers between aggressive and non-aggressive PCa patients and the methylation level differences between GS=6 and GS≥8 patients were consistent across all immune cell types. The absolute methylation level difference (β value difference) of each individual CpG site between GS=6 and GS≥8 patients was modest, and the

prediction accuracy of our model reached a plateau of 85%. This limitation of predicting aggressive PCa using leukocyte DNA methylation is not surprising given the predominant background of normal immune cells. We only included GS=6 and GS≥8 patients in this analysis because they have distinct clinical phenotypes. At the same time, since DNA 450k microarray only covers about 2% of total genome CpGs and we have few WGBS samples, our findings may be limited to specific regions. This study design is intended to identify biological features that differentiate clinically defined aggressive diseases from non-aggressive diseases. GS=7 patients, on the other hand, have intermediate risks of progression and their outcomes are more heterogeneous and more difficult to predict.

Taken 450k DNA microarray and WGBS data together, we found good consistencies across these two sequencing platforms. Moreover, we discover a novel 7-CpG biomarker list which was mutually altered in both datasets.

In summary, we performed a large-scale DNA methylation profiling of leukocyte DNA in clinically defined aggressive and non-aggressive PC patients. We observed hypermethylation in transcriptionally active regions of aggressive PCa patients compared to non-aggressive PCa patients and global hypomethylation in PCa patients. We identified a 97-CpG methylation signature in leukocytes that is associated with aggressive PCa at diagnosis, and a 7-CpG PCa biomarker list across both sequencing platforms. Our study also provides biological insights into the modulation of the immune system by aggressive PCa.

CHAPTER III

MOABS-GALAXY: A WEB-BASED ONLINE TOOLKIT FOR BS-SEQ ANALYSIS

Introduction

DNA Cytosine methylation, an epigenetic modification on DNA, has various functional roles in development and disease (*205*). 5-methylcytosine (5mC) exists mostly at CpG dinucleotides throughout the genome, yet executes distinct functions in different genomic regions. Although DNA methylation is typically associated with gene transcription repression, the fundamental regulatory mechanism can be different at gene promoters, in gene bodies, or in repeated sequences. Methylation in promoters with increased CpG densities represses transcription and correlates with gene silencing potential (*206*). DNA methylation is enriched in gene bodies. Gene-body methylation is highly conserved in eukaryotes (*207*), and it is positively correlated with transcription (*208*). A recent study indicated that abnormal methylation in DNA repeat regions is directly linked to various inherited rare diseases, such as facial anomalies syndrome, Huntington's disease, and Fragile X syndrome (*209*). Therefore, it is important to study the mechanisms underlying the regulatory roles of DNA methylation.

Advanced technologies enable broad protocols to detect DNA methylation. In particular, bisulfite sequencing (BS-Seq) has emerged as a golden standard for genome-wide DNA methylation profiling at the single-base resolution. The most widely used protocols include Reduced Representation Bisulfite Sequencing (RRBS) (*210*) and Whole Genome

Bisulfite Sequencing (WGBS) (*211*). Utilizing the wealth of BS-Seq data, we proposed Model based Analysis of Bisulfite Sequencing data（MOABS) to detect differential methylation changes at a single-CpG resolution (*192*). Comprising mapping of short-reads via RRBSMAP (*212*), and BS-Seq data quality control using BSeQC (*213*), MOABS uses a beta-binomial hierarchical model to capture sampling and biological variation. Credible methylation difference (CDIF) incorporates biological significance into the statistical significance of differential methylation. MOABS revealed stable differentially methylated CpGs and regions with a fast, accurate, statistically powerful, and a biological relevant analysis of BS-Seq data. For example, MOABS facilitated the discovery of epigenetic dysregulation in aged hematopoietic stem cells from murine bone marrow compared with those from young mice, and identified changes such as hypermethylation at transcriptome binding sites of differentiation-promoting genes and hypomethylation at HSC maintenance genes (*214*). In addition to bisulfite sequencing to measure DNA 5mC, MOABS is also ideal for base-resolution bisulfite sequencing technologies to measure DNA hydroxymethylcytosine (5hmC), DNA N6-methyladenine (6mA), and RNA 5mC, such as oxBS-Seq (*215*), TAB-Seq (*216*), and CMS-IP-Seq (*217*).

The Galaxy platform is a scientific analysis platform in a web service format (*218*). The platform aims to stimulate accessible and reproducible scientific analyses for scientists in a broad area of research. Its web interface makes data-heavy analysis much easier, especially for scientists who need not possess extensive software development experience. The Galaxy platform consists of several complementary components. Its public Galaxy servers host thousands of high-quality software programs supported with robust computational power and large data storage space. Using these public Galaxy servers,

researchers can conduct their data analysis instantly without worrying about software installation and learning their many options. The Galaxy platform is especially helpful for those who may not have their own computational resources. Other components include the Galaxy framework and the Galaxy ToolShed. The Galaxy framework allows users to install their own Galaxy servers together with many Galaxy tools stored in the Galaxy ToolShed. The Galaxy platform now is widely used by tens of thousands of scientists in biomedical research communities.

To widen accessibility, many sequencing data analysis tools have incorporated front-end web services built upon the Galaxy platform. Their web services can be accessed via the URL https://usegalaxy.eu/root?toolid=<tool id>. For example, FastQC (*219*) (tool id: fastqc) and MultiQC (*220*) (tool id: multiqc) have implemented web interfaces for quality control of raw high-throughput sequencing data and generating well-presented summary reports. To map raw reads to a reference genome or a transcriptome, Bowtie2 (*221*) (tool id is bowtie2) and RNA-STAR (*222*) (tool id: rnastar) can be used for mapping RNA reads and DNA reads, respectively. There are also several Galaxy web interfaces for widely-used differential gene expression analysis tools, such as limma (*223*) (tool id: limmavoom), edgeR (*224*) (tool id: edger), and DESeq2 (*225*) (tool id: deseq2). In summary, over 5,500 web-based tools are available in the Galaxy ToolShed repository (*218*) up to 2018, the number of available tools is increasing over years, and valid tools has reached over 7,500 by January 2020 (https://galaxyproject.org/galaxy-project/statistics/#galaxy-project-tool-shed).

To broaden the accessibility of MOABS, we developed the MOABS-Galaxy web service. MOABS-Galaxy encapsulates MOABS with web interfaces utilizing the benefits of the Galaxy platform. Its web interface facilitates the ready analysis of BS-Seq data for

detecting differential methylation. MOABS-Galaxy not just strengthens the use of MOABS among scientists who are interested in DNA methylation but also contributes to the Galaxy software ecosystem for the Galaxy platform.

Materials and Methods

Implementation

MOABS is a consolidated solution for bisulfite experiment analysis, consisting of modules for reads alignment, single sample analysis, and multiple-sample comparative analysis (*192*). MOABS is written in the programming languages C++ and Perl. Even though MOABS source code has been originally publicly available, it required extensive efforts for users to compile the tools by themselves due to library dependencies, e.g., Boost libraries. To alleviate the installation burden, we had provided a download of static-complied binaries, but the sizes of static binaries became unnecessarily large compared to dynamically linked binaries. To improve the installation of MOABS, we have reorganized the source code using the GNU Autotools (*226*). The Autotools is a suite of programming tools to facilitate the portable distribution of source code. The Autotools enabled easy installation of MOABS across common platforms, especially Unix-like systems.

To facilitate the deployment of MOABS in Galaxy, MOABS was first deployed in Bioconda (*227*). Bioconda is a channel from the Conda package manager, and it aims to boost the accessibility and reproducibility of bioinformatics software. Utilizing this popular and powerful software management platform, the MOABS package in Bioconda frees users from laborious configurations and manual compiling, and it is now easily installable for the research community.

To distribute MOABS in the Galaxy platform, the MOABS-Galaxy web interface was added to the tools-iuc Github repository (https://github.com/galaxyproject/tools-iuc). This repository hosts a set of high-quality Galaxy tools specifically maintained by the Intergalactic

Utilities Commission (https://galaxyproject.org/iuc). Galaxy tools in the tools-iuc repository are also automatically uploaded to the Test and Main Galaxy Tool Shed, which stores public tools available for the Galaxy platform.

Availability

MOABS source code and its manual is hosted on Github at https://github.com/sunnyisgalaxy/moabs. The MOABS package in Bioconda is accessible at https://anaconda.org/bioconda/moabs. The MOABS-Galaxy web service is freely usable in the Galaxy public server at https://usegalaxy.eu/root?toolid=moabs. The support email is moabsmsuite@googlegroups.com. The discussion group is at https://groups.google.com/d/forum/moabsmsuite.

Results

The User-friendly interface of MOABS-Galaxy

The MOABS-Galaxy web service incorporates interfaces to configure input and output parameters. The reference genome sequences are available as a cached genome FASTA on the Galaxy server, e.g., hg38, or users may upload a customized reference genome (Figure 17A). The input read files should have two groups for comparison, e.g., a tumor group and a healthy control group of the same tissue types (Figure 17B). The “group” concept is for the web interface purpose. Users are responsible for identifying biological groups for a meaningful biological comparison. Reads in each group can be combined sequencing libraries, i.e., single-end reads and paired-end reads. Replicates can be specified for each group in order to reduce the false positive discovery of methylation dynamics by considering the reproducibility between different batches or samples of the same group. MOABS-Galaxy also provides interfaces for users to configure 15 parameters for BSMAP, 10 parameters for MCALL, and 9 parameters for MCOMP (submodules of MOABS). For example, users can specify MCOMP parameters for minimum CpG sequencing depth, minimum absolute credible difference (CDIF), and DMC or DMR cutoffs (Figure 17C).

Five sections of results will be generated, including 1) The alignment files generated by BSMAP: BAM files containing mapped reads information. BAM files are compressed binary version of SAM files that are used to represent aligned sequences. 2) The methylation calling files produced by MCALL: files containing methylation calling information with single CpG resolution. 3) The comparison result file: a tabular style file containing CpG information

of statistical test results between two comparison groups. 4) the DMC loci file: a TXT file for DMCs. 5) the DMR file: a TXT file for DMRs. Detailed interpretation of these result files is explained in the web interface homepage at https://usegalaxy.eu/root?toolid=moabs (Figure 17D).

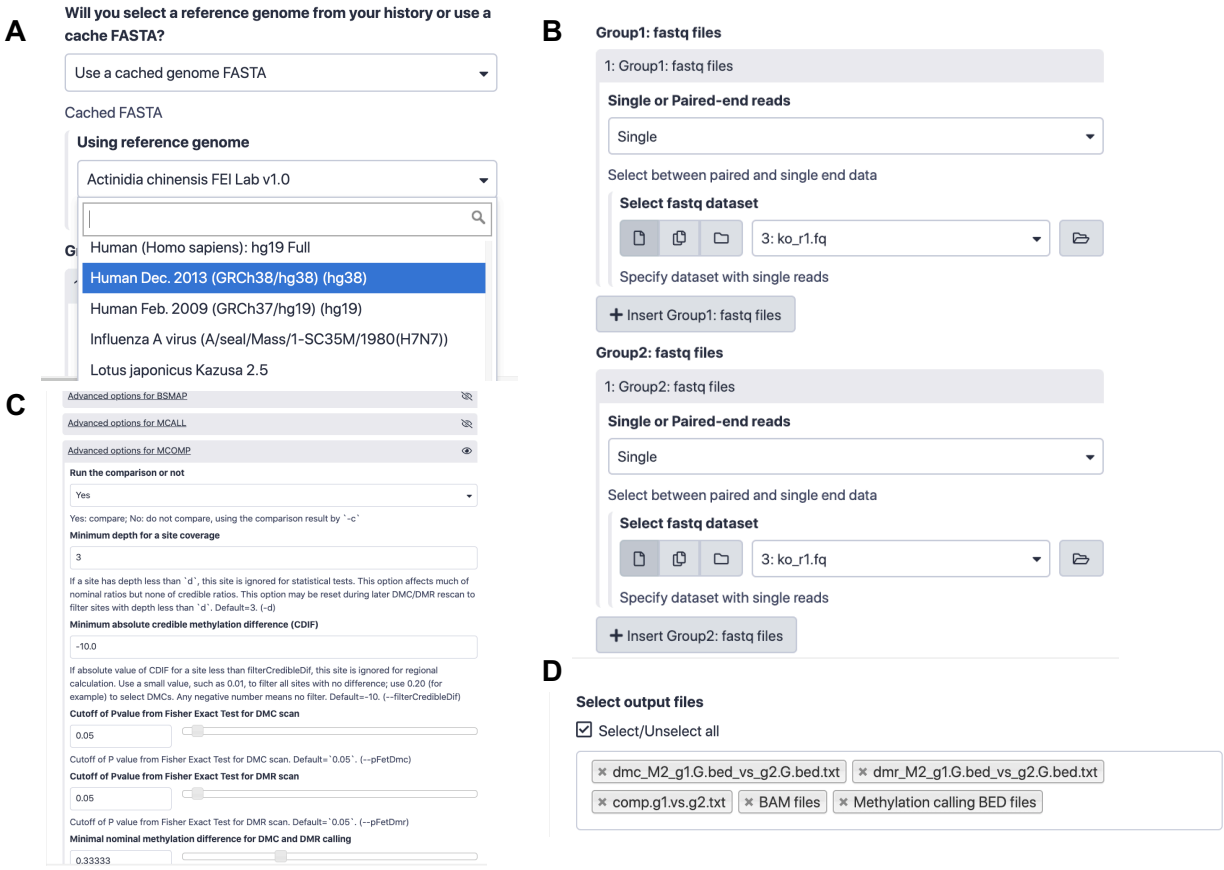


Figure 17. MOABS-Galaxy web interfaces.

(A) The interface for the reference genome. It can be a cached genome FASTA or a FASTA file in history. (B) Input bisulfite-sequencing reads for two groups. The FASTQ files can be a combined library layout such as single or paired-end reads. (C) Advanced options for BSMAP, MCALL, and MCOMP. (D) Options to generate selected results files. Five result files can be generated, including mapping BAM files, the comparison result file, the methylation calling result file, the DMC location file, and the DMR region file.

Quick test analysis

To allow users to glimpse how MOABS-Galaxy works and familiarize them with the output format, we prepared a quick test analysis that can be performed on the web interface. Users should go to https://usegalaxy.eu and register a user account with an email address. One user account is granted 250 GB of disk usage for data storage. Based on data availability and total size, users may choose different options from local file upload (local small files), FTP upload (local big files), URL upload, and SRA upload. For the test run, users may download the example test data from the amazon AWS URL at https://s3.amazonaws.com/deqiangsun/software/moabsgalaxy/Testrun.zip. After downloading the test run data, click on the "upload" button on the top left of the main page in galaxy (Figure 18A). In the pop-up interface, select "choose local file" and choose the 3 files downloaded before. Press "Start". The file upload speed depends on the user web service speed (Figure 18B).

Next, users should access the MOABS-Galaxy web interface by visiting https://usegalaxy.eu/root?toolid=moabs. In the MOABS-Galaxy interface, select "Use a genome FASTA from history" for "Will you select a reference genome from your history or use a cache FASTA?" and select chr1.fa in the following section. For input files, keep "Single" and select "s1r1.fq" and "kor1.fq" for the 2 groups. For a test run, leave all advanced options as default. In the "Select output files" section, check "Select/Unselect all" (Figure 18C). Now users are all set for a test run. Press the "Execute" button. The test run will require ~5 minutes. When the run is done, users should find new files in green in the history panel on the right side of the screen. Users may freely download the output files and perform further analysis with the results (Figure 18D).

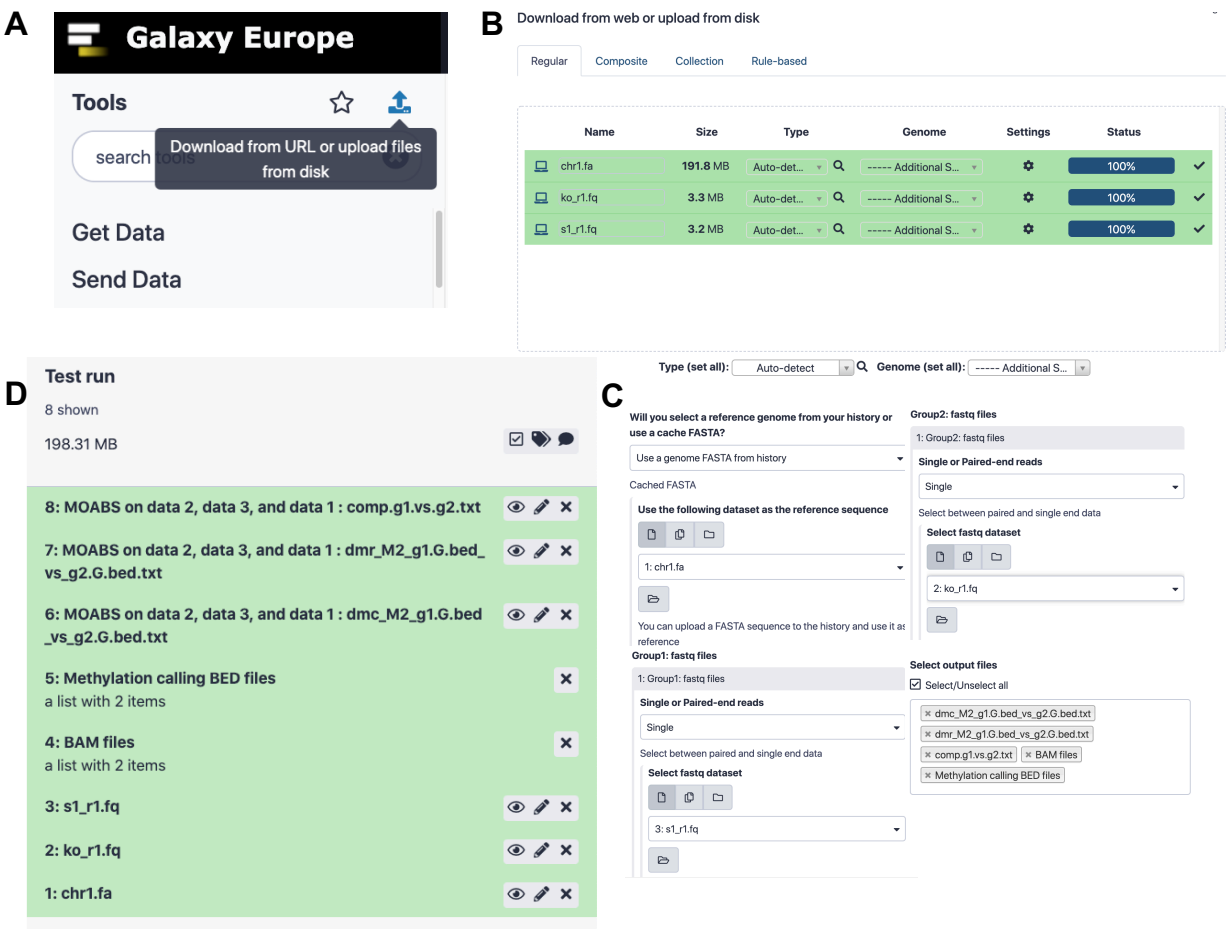


Figure 18. A test analysis with MOABS-Galaxy.

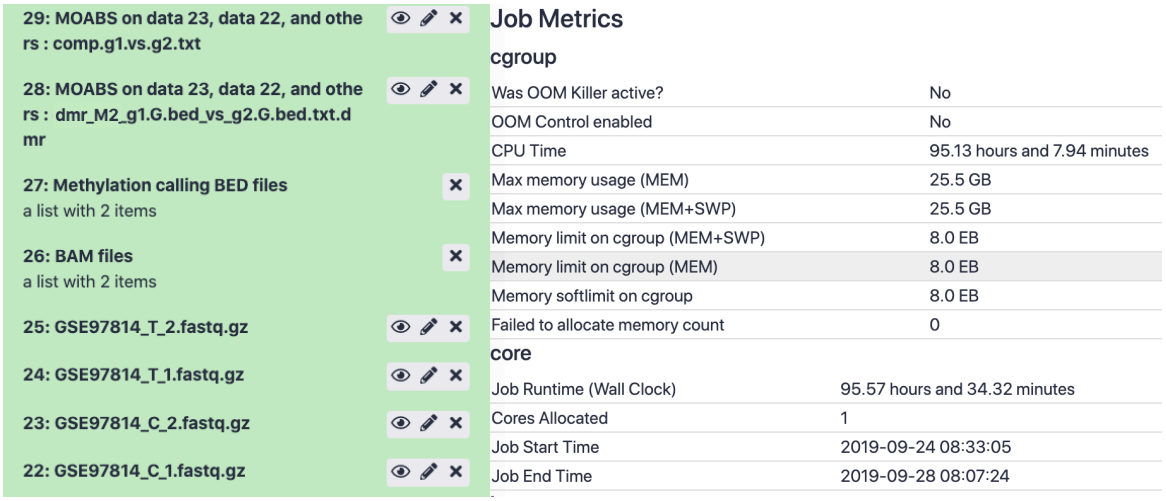
(A) Data upload page in Galaxy, in this tutorial, we exhibited a local upload as an example. To view the complete guidelines for uploading a dataset to galaxy.org, view: https://galaxyproject.org/tutorials/upload/. (B) Uploaded test dataset page in Galaxy. For the test run, the reference genome (chr1.fa) and two sequencing fastq files (s1r1.fq, kor1.fq) should be correctly uploaded. (C) MOABS-Galaxy configuration page for the test run. Users should select “Use a genome fasta from history” and select chr1.fa, “Single” for both data groups 1 and 2, and select s1r1.fq, kor1.fq as inputs respectively. (D) The result page of the quick test run on MOABS-Galaxy. Input FASTQ files and four results files are listed in the user’s history (Green section on the left). Each file in history has an incremental ID with its name. A small reference (ID: 1) is provided for a shorter test run time. The two input FASTQ files have single IDs (ID:2 and 3). The resulting BAM files (ID:

12) and the methylation calling BED files (ID: 13) are file lists (two items in each list). The DMC file (ID:14), the DMR file (ID:15), and the comparison file (ID:16) are stored as single files.

A Use case for WGBS data

To demonstrate the utility of MOABS-Galaxy, we analyzed a public WGBS dataset (GSE97814) using the web service. This dataset was analyzed to evaluate the methylation changes of the global DNA methylome upon CRISPR-Cas SunTag-directed DNMT3A treatment (*228*). WGBS FASTQ files were uploaded to UseGalaxy.edu. MOABS-Galaxy generated a list of differentially methylated regions between the HOAX5-guided-edited sample and the control (Figure 19A). Increased methylation in the *HOAX5* loci following dCas9-SunTag-DNMT3A1 treatment using HOXA5 guide RNAs was detected by MOABS-Galaxy (*228*). Hence, MOABS-Galaxy was able to detect reported DMRs both upstream and downstream of the *HOAX5* loci. UCSC genome browser tracks visualized these increased methylation levels (Figure 19B). These results demonstrated the high consistency of MOABS-galaxy with other analysis approaches for WGBS data.

A



B

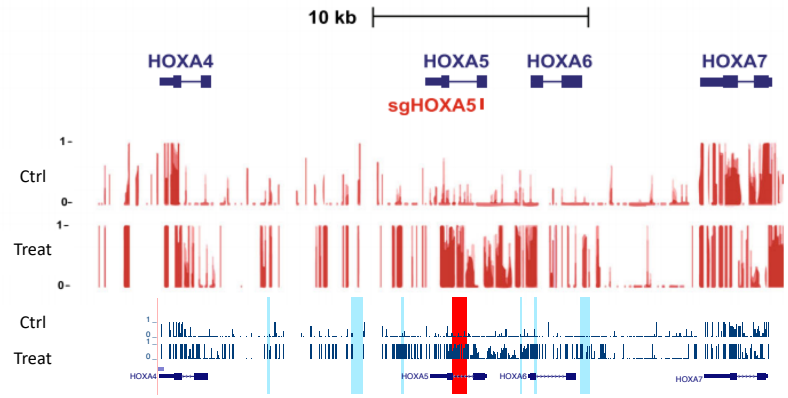


Figure 19. A WGBS analysis with MOABS-Galaxy.

(A) The result page of the use case on MOABS-Galaxy. Input FASTQ files and four results files are listed in the user's history (Green section on the left). Each file in history has an incremental ID with its name. Four input FASTQ files have single IDs (from ID:22 to ID:25). The resulting BAM files (ID: 26) and the methylation calling BED files (ID: 27) are file lists (two items in each list). The comparison file (ID:28) and the DMR file (ID:29) are stored as single files. On the right are the running statistics of the analysis. (B) UCSC genome browser tracks of the methylation levels in the HOXA5 locus. Methylation changes were reported in HOXA5 specific regions (red tracks above), and our MOABS-Galaxy analysis

identified DMRs (highlighted regions) from upstream to downstream of the HOXA5 locus in the blue tracks below.

Use case for RRBS data

To exhibit the versatility of MOABS-Galaxy for various types of DNA methylation sequencing data from disparate platforms, we analyzed a public RRBS dataset (GSE80761) using the web service (*229*). MOABS-Galaxy mapping and methylation calling results were almost identical to the original article reports (Figure 20A), such as the bisulfite conversion ratio, total CpG number, CpG number with coverage, CpG overall depth, and global CpG methylation ratio. Furthermore, the CpG methylation patterns categorized by methylation level were close to those reported in the article (Figure 20B). We observed binomial distributions while hypomethylated (0-10%) CpGs consisted of ~60% of the total CpGs. These results showed the high consistency of MOABS-Galaxy with other approaches for RRBS data analysis.

A

	Mixed	Mixed - MOABS	B6	B6 - MOABS
Number of reads	34644517	35868585	49783046	53089897
Mapping effciency	67.90%	64.60%	67.70%	62.50%
Conversion rate	99.88%	99.86%	99.89%	99.87%
Total # of CpG sequenced	73450633	67205202	101405046	101707390
# of CpGs covered > 10x	1351160 (1.84%)	1049144 (1.56%)	1542759 (1.53%)	1232470 (1.21%)
Mean CpG coverage depth	47	42	62	57
CpG methylation	27.30%	27.53%	28.10%	28.11%
non-CpG methylation	0.30%	0.20%	0.20%	0.20%

B

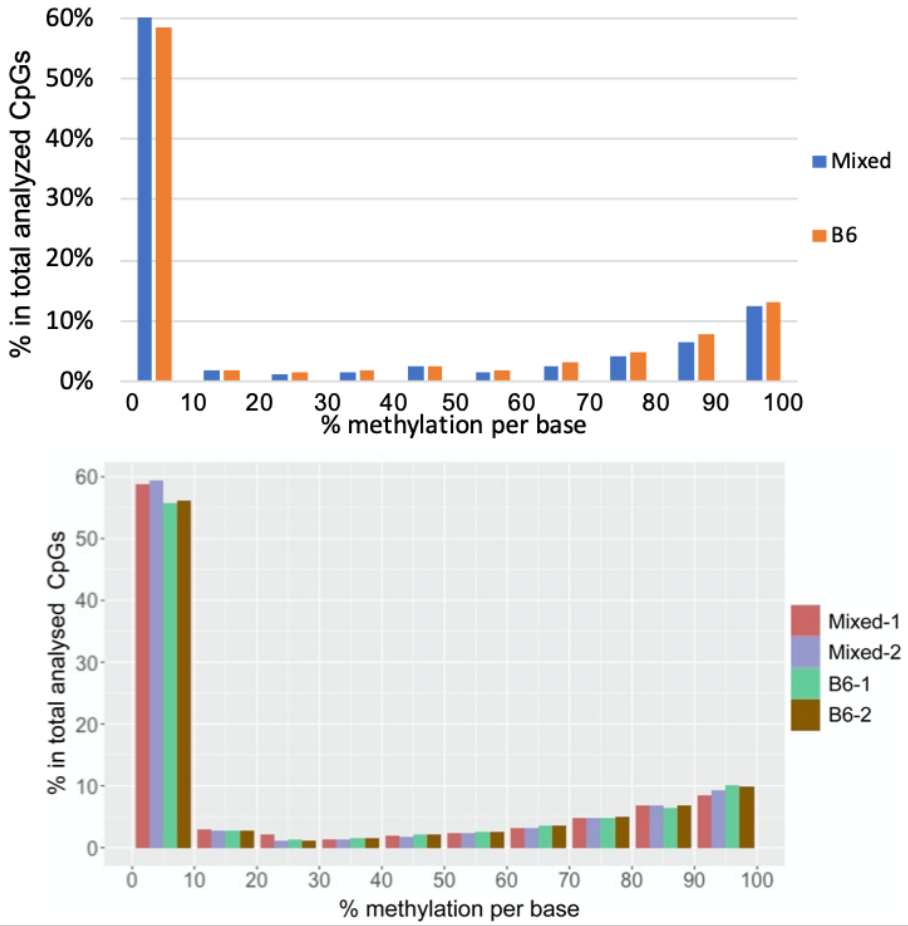


Figure 20. An RRBS analysis with MOABS-Galaxy.

(A) Table for mapping and methylation calling statistics. Original statistics were directly from the original article Table 1(*229*). Columns 1 and 3 are original statistics, while

columns 2 and 4 are results from MOABS-Galaxy. (B) CpG methylation patterns are categorized by methylation ratio. All CpGs were categorized into 10 bins (from 0% to 100% methylated). The barplot above shows the average CpG methylation percentage of each categorized CpG group in Mixed or B6 mice. The original barplot is shown below with all 4 samples.

Advanced options

To fully wrapper MOABS functions, we provided more than 30 advanced adjustable options for users to tailor the analysis to their requirements. For the mapping module BSMAP, users may adjust 15 parameters in 4 categories: trimming options, quality control options, RRBS options, and mapping report options (Table 3). For the methylation calling module MCALL, users may adjust 10 parameters in 4 categories: quality cut-off options, trimming options, RRBS option, and report style options (Table 4). For the methylation comparing module MCOMP, users may adjust 9 parameters in 4 categories: quality control options, cut-off by p-value options, cut-off by methylation difference options, and DMR identification options (Table 5). With those advanced options available, MOABS-Galaxy can serve as an integrated workflow with great versatility for various whole genome bisulfite sequencing data projects.

Advanced option list for BSMAP module		
	Advanced Option	Description
Trimming options		
	Quality threshold in trimming	Trim low quality reads with q-scores less than X. Max value: 40, default value: 0.
	3' adapter sequence to trim	Provide the sequence that the user would like to trim for 3' end.
Quality control options		
	Base quality	Provide sequencing reads type, Illumina or Sanger (default).
	Maximum number of Ns in a read to filter out	Filter out reads with more than X Ns. Default value: 5.
	Minimal insert size allowed in paired-end mapping	The minimal insert size value allowed for paired-end mapping reads. Default value: 28.
	Maximal insert size allowed in paired-end mapping	The maximal insert size value allowed for paired-end mapping reads. Default value: 1000.
	Mismatch rate/bases	The maximal mismatch rate or mismatch bases number in a single read. Default value: 8%.
	Maximum number of equal best hits to count:	The maximal number of records to report when there are multiple mapping occurs for a read. Default value: 1000.
	Random Seed	Seed for random number generation used in selecting multiple hits. Other seed values generate a pseudo random number based on read index number, to allow reproducible mapping results. Default value:0.
RRBS options		
	Seed size	The longer seed size, the faster speed. Default value for WGBS mode: 16. Default value for RRBS mode: 12.
	Restriction enzyme digestion sites for RRBS mode	This option activates the RRBS mapping mode and sets restriction enzyme digestion sites. Enzyme cleavage site marked by '-', example: -D C-CGG for MspI digestion.
Mapping report options		
	Mapping for 4 strands	Specify mapping strand options. Yes: map SE or PE reads to all 4 strands, i.e. ++, +-, -+, --; No: only map to 2 forward strands, i.e. BSW(++) and BSC(-+).
	How to report repeat hits	How to report reads that have multiple hits on the reference. 0=none (only report unique hit/pair); 1=select random one when there are multiple hits; 2=report all hits (slow, not recommended). Default value: 0.
	Print corresponding reference sequences	Print corresponding reference sequences in mapping records, a `RS:` tag will be added in record attributes. Default value: No.
	Report unmapped reads	Report unmapped reads. Default value: No.

Table 6. Advanced options for the BSMAP module.

There are 15 advanced options available for the BSMAP module, which can be categorized into 4 bins: trimming options, quality control options, RRBS options, and mapping report options.

Advanced option list for MCALL module		
	Advanced Option	Description
Quality cut-off options		
	Specify the quality score system	Specify the quality score system, Available options: Sanger, Solexa, or Illumina. Default value: auto-detection.
	Threshold for cytosine quality score	Filter out CpG with low quality score. Default value: 20.
	Threshold for the next base quality score	Possible values: -1 makes the program not to check if next base matches reference; any positive integer or zero makes the program to check if the next base matches reference and reaches this score threshold. Default value: 3.
	Minimal fragment size for properly mapped reads	The 9th field in the BAM file is the fragment size of the mapping, and non-properly-paired reads have 0 at the 9th field. This option is set to require properly paired and sufficiently large fragment size. Default value: 0.
	Minimal fragment size for multiply matched reads	Same as the option above but this option is only applicable to reads with flag 0x100 set as 1, i.e., reads multiply mapped. Default value: 0.
Trimming options		
	Bases to trim end-repair sequences from +-/--	Trim end-repair sequences from the beginning of +-/-- reads from Pair End WGBS Sequencing. Default value: 3.
	Bases to trim end-repair sequences from ++/-+	Trim end-repair sequences from the beginning of ++/-+ reads from Pair End WGBS Sequencing. Default value: 3.
RRBS option		
	How to trim end-repair sequence for RRBS reads	Trim end-repair sequences for RRBS reads. Default value: 2.
Report style options		
	Count once or twice the overlap sequence of two pairs	For paired-end sequencing, count once or twice for overlapped sequences in CpG methylation measurements. Default value: once.
	Generates CpG/A/C/T methylation	Measure methylation for CpG, or CpA/CpC/CpT. Default value: CpG.

Table 7. Advanced options for the MCALL module.

There are 10 advanced options available for the MCALL module, which can be categorized into 4 bins: quality cut-off options, trimming options, RRBS option, and report style options.

Advanced option list for MCOMP module		
Quality control options		
	Run the comparison or not	Run MCOMP module or not. Default value: Yes.
	Minimum depth for a site coverage	Filter out CpG sites with less than X depth. Default value: 3.
Cut-off by p-value options		
	Cutoff of Pvalue from Fisher Exact Test for DMC scan	Filter out DMCs with p-value less than X. Default value: 0.05.
	Cutoff of Pvalue from Fisher Exact Test for DMR scan	Filter out DMRs with p-value less than X. Default value: 0.05.
Cut-off by methylation difference options		
	Minimum absolute credible methylation difference (CDIF)	If the absolute CDIF for a site is less than X, this site is ignored for regional calculation. Default value: -10.
	Minimal nominal methylation difference for DMC and DMR calling	Filter out DMCs with methylation difference less than X. Default value: 0.3333.
	Minimal credible methylation difference for DMC calling	Filter out DMCs with CDIF less than X. Default value: 0.2.
DMR identification options		
	Minimum number of DMCs in a DMR	The minimal DMC number in a DMR. Default value: 3.
	Maximum distance between two consective DMCs for a DMR	The maximal distance between 2 consective DMCs allowed for a DMR. Default value: 300.

Table 8. Advanced options for the MCOMP module.

There are 9 advanced options available for the MCOMP module, which can be categorized into 4 bins: quality control options, cut-off by *p*-value options, cut-off by methylation difference options, and DMR identification options.

Conclusion and Discussion

We have developed the MOABS-Galaxy web service for DNA methylation analysis using bisulfite sequencing data such as WGBS, RRBS, and CMS-IP-Seq data. MOABS-Galaxy is deployed as a Galaxy tool in the Galaxy Tool Shed, and it is publicly available on the Galaxy server. The web interface of MOABS-Galaxy enables convenient DNA methylation data analysis which was demonstrated by our use case. It is especially useful for researchers who may have limited computational resources and minimal bioinformatics skills. MOABS-Galaxy constitutes a useful extension to the Galaxy tool ecosystem and will promote the study of DNA methylation in biomedical research.

CHAPTER IV

CONCLUSION AND FUTURE DIRECTION

We proposed a novel test by utilizing the DNA microarray and a powerful WGBS technology for PBL DNA from PCa patients, as well as big data analysis methods such as machine learning to predict PCa aggressiveness. The following figure shows the working flowchart of our approach.

The blood sample can be obtained from any person without knowing cancer history. Once the patient blood sample is collected, we will extract PBL DNA from white cells. The next steps are library preparation, whole genome bisulfite sequencing and data analysis based on specific methylation markers. The analysis will generate a report on the prediction of cancer status with high sensitivity and specificity (Figure 2).

Figure 21. Schematic diagram for PCa screening with patient PBL DNA methylation biomarker.

With this work, we may discover a novel diagnostic method for PCa in the future: we can extract methylation information from patient peripheral blood, which will be utilized to predict cancer existence and cancer aggressiveness if cancer exists. Moreover, our

study also provided a better understanding of the comparison between two different methylation sequencing platforms, DNA microarray and WGBS. Future studies are needed to determine whether leukocyte DNA methylation can predict more aggressive clinical behavior in GS=7 patients.

In this study, a random forest model handled data with high dimensionality properly. When the training data have many features, a random forest model performed well since the bagged tree are working with a subset of the features. More importantly, a random forest model would have relatively higher accuracy before any feature selection (high resistance on noise). These characters of random forest make it a good fit for our project since we trained the model with many CpG probes' β-value from the 450k DNA microarray platform, and the most majority of the probes (>99.7%) were not biomarkers (noises). After training, random forest model provided feature importance based on mean Gini decrease or impurity decrease. This was crucial for our project since we expect to select a small set of probes as methylation biomarkers, and the rank of probe importance gives us direct evidence to filter the candidates. Random forest algorithm utilized unbiased estimate to assess the generalization error, which offered the model stronger generalization ability. Moreover, the random forest algorithm was easy to perform, requires low computational resources since the trees are independent to each other (parallelizable). Numerous publications utilized the random forest algorithm to select gene symbols or classify patient samples (*230-236*).

MOABS-Galaxy web service offers integrated, user-friendly solution for various types of bisulfite sequencing analysis to the public. We are proud of the speed, accuracy, statistical power, and biological relevance of MOABS, and excited for releasing this tool by

joining the galaxy.eu toolshed. More importantly, with increased user experience, we expect to hear feedbacks from the galaxy user community and further optimize the MOABS-Galaxy program.

REFERENCES

1. R. L. Siegel, K. D. Miller, H. E. Fuchs, A. Jemal, Cancer Statistics, 2021. *CA Cancer J Clin* **71**, 7-33 (2021).
2. J. L. Jahn, E. L. Giovannucci, M. J. Stampfer, The high prevalence of undiagnosed prostate cancer at autopsy: implications for epidemiology and treatment of prostate cancer in the Prostate-specific Antigen-era. *Int J Cancer* **137**, 2795-2802 (2015).
3. A. R. Zlotta *et al.*, Prevalence of prostate cancer on autopsy: cross-sectional study on unscreened Caucasian and Asian men. *J Natl Cancer Inst* **105**, 1050-1058 (2013).
4. K. J. Bell, C. Del Mar, G. Wright, J. Dickinson, P. Glasziou, Prevalence of incidental prostate cancer: A systematic review of autopsy studies. *Int J Cancer* **137**, 1749-1757 (2015).
5. J. C, P. Y, B. Sf, B. Rj, "More men die with prostate cancer than because of it" - an old adage that still holds true in the 21st century. *Cancer Treat Res Commun* **26**, 100225 (2021).
6. O. Husson *et al.*, Patients with prostate cancer continue to have excess mortality up to 15 years after diagnosis. *BJU Int* **114**, 691-697 (2014).
7. W. Hamilton, D. J. Sharp, T. J. Peters, A. P. Round, Clinical features of prostate cancer before diagnosis: a population-based, case-control study. *Br J Gen Pract* **56**, 756-762 (2006).

8. C. Mettlin, F. Lee, J. Drago, G. P. Murphy, The American Cancer Society National Prostate Cancer Detection Project. Findings on the detection of early prostate cancer in 2425 men. *Cancer* **67**, 2949-2958 (1991).
9. B. Lojanapiwat, W. Anutrakulchai, W. Chongruksut, C. Udomphot, Correlation and diagnostic performance of the prostate-specific antigen level with the diagnosis, aggressiveness, and bone metastasis of prostate cancer in clinical practice. *Prostate Int* **2**, 133-139 (2014).
10. W. J. Catalona *et al.*, Comparison of Digital Rectal Examination and Serum Prostate Specific Antigen in the Early Detection of Prostate Cancer: Results of a Multicenter Clinical Trial of 6,630 Men. *J Urol* **197**, S200-S207 (2017).
11. S. Loeb *et al.*, Overdiagnosis and overtreatment of prostate cancer. *Eur Urol* **65**, 1046-1055 (2014).
12. *Surveillance, Epidemilogy, and EndResults (SEER) Program* (2016).
13. A. C. Society. (2021).
14. U. S. F. a. D. Administration. (2020).
15. J. C. Weinreb *et al.*, PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. *Eur Urol* **69**, 16-40 (2016).
16. J. Tosoian, S. Loeb, PSA and beyond: the past, present, and future of investigative biomarkers for prostate cancer. *ScientificWorldJournal* **10**, 1919-1931 (2010).
17. T. Nordstrom, O. Akre, M. Aly, H. Gronberg, M. Eklund, Prostate-specific antigen (PSA) density in the diagnostic algorithm of prostate cancer. *Prostate Cancer Prostatic Dis* **21**, 57-63 (2018).

18. K. Ito *et al.*, Free/total PSA ratio is a powerful predictor of future prostate cancer morbidity in men with initial PSA levels of 4.1 to 10.0 ng/mL. *Urology* **61**, 760-764 (2003).
19. F. Strittmatter *et al.*, Detection of prostate cancer with complexed PSA and complexed/total PSA ratio - is there any advantage? *Eur J Med Res* **16**, 445-450 (2011).
20. G. M. Oremek, N. Sapoutzis, F. Eden, D. Jonas, Complexed PSA in routine diagnosis. *Anticancer Res* **23**, 975-977 (2003).
21. W. Horninger *et al.*, Complexed prostate-specific antigen for early detection of prostate cancer in men with serum prostate-specific antigen levels of 2 to 4 nanograms per milliliter. *Urology* **60**, 31-35 (2002).
22. H. U. Ahmed *et al.*, Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet* **389**, 815-822 (2017).
23. M. Ahdoot *et al.*, MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis. *N Engl J Med* **382**, 917-928 (2020).
24. B. Delahunt, R. J. Miller, J. R. Srigley, A. J. Evans, H. Samaratunga, Gleason grading: past, present and future. *Histopathology* **60**, 75-86 (2012).
25. N. Chen, Q. Zhou, The evolving Gleason grading system. *Chin J Cancer Res* **28**, 58-64 (2016).
26. M. A. Bjurlin *et al.*, Update of the Standard Operating Procedure on the Use of Multiparametric Magnetic Resonance Imaging for the Diagnosis, Staging and Management of Prostate Cancer. *J Urol* **203**, 706-712 (2020).

27. L. Rodgers, C. J. Peer, W. D. Figg, Diagnosis, staging, and risk stratification in prostate cancer: Utilizing diagnostic tools to avoid unnecessary therapies and side effects. *Cancer Biol Ther* **18**, 470-472 (2017).

28. M. K. Buyyounouski *et al.*, Prostate cancer - major changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin* **67**, 245-253 (2017).

29. J. L. Mohler *et al.*, Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* **17**, 479-505 (2019).

30. N. Borley, M. R. Feneley, Prostate cancer: diagnosis and staging. *Asian J Androl* **11**, 74-80 (2009).

31. F. C. Hamdy *et al.*, 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. *N Engl J Med* **375**, 1415-1424 (2016).

32. N. Suardi *et al.*, Nerve-sparing approach during radical prostatectomy is strongly associated with the rate of postoperative urinary continence recovery. *BJU Int* **111**, 717-722 (2013).

33. J. U. Stolzenburg *et al.*, Effect of surgical approach on erectile function recovery following bilateral nerve-sparing radical prostatectomy: an evaluation utilising data from a randomised, double-blind, double-dummy multicentre trial of tadalafil vs placebo. *BJU Int* **116**, 241-251 (2015).

34. A. L. Siegel, Pelvic floor muscle training in males: practical applications. *Urology* **84**, 1-7 (2014).

35. A. Straczynska *et al.*, The Impact Of Pelvic Floor Muscle Training On Urinary Incontinence In Men After Radical Prostatectomy (RP) - A Systematic Review. *Clin Interv Aging* **14**, 1997-2005 (2019).

36. G. L. Lu-Yao, S. L. Yao, Population-based study of long-term survival in patients with clinically localised prostate cancer. *Lancet* **349**, 906-910 (1997).

37. A. Taylor, M. E. Powell, Intensity-modulated radiotherapy--what is it? *Cancer Imaging* **4**, 68-73 (2004).

38. N. Nabavizadeh *et al.*, Image Guided Radiation Therapy (IGRT) Practice Patterns and IGRT's Impact on Workflow and Treatment Planning: Results From a National Survey of American Society for Radiation Oncology Members. *Int J Radiat Oncol Biol Phys* **94**, 850-857 (2016).

39. J. Skowronek, Brachytherapy in the therapy of prostate cancer - an interesting choice. *Contemp Oncol (Pozn)* **17**, 407-412 (2013).

40. J. Skowronek, Current status of brachytherapy in cancer treatment - short overview. *J Contemp Brachytherapy* **9**, 581-589 (2017).

41. Z. Wang *et al.*, The efficacy and safety of radical prostatectomy and radiotherapy in high-risk prostate cancer: a systematic review and meta-analysis. *World J Surg Oncol* **18**, 42 (2020).

42. A. S. Kibel *et al.*, Survival among men with clinically localized prostate cancer treated with radical prostatectomy or radiation therapy in the prostate specific antigen era. *J Urol* **187**, 1259-1265 (2012).

43. L. Incrocci, Radiotherapy for prostate cancer and sexual health. *Transl Androl Urol* **4**, 124-130 (2015).

44. B. G. Vanneste *et al.*, Chronic radiation proctitis: tricks to prevent and treat. *Int J Colorectal Dis* **30**, 1293-1303 (2015).

45. P. E. Teloken, M. Parker, N. Mohideen, J. P. Mulhall, Predictors of response to sildenafil citrate following radiation therapy for prostate cancer. *J Sex Med* **6**, 1135-1140 (2009).

46. R. W. Ross *et al.*, Efficacy of androgen deprivation therapy (ADT) in patients with advanced prostate cancer: association between Gleason score, prostate-specific antigen level, and prior ADT exposure with duration of ADT effect. *Cancer* **112**, 1247-1253 (2008).

47. A. Liede *et al.*, International survey of androgen deprivation therapy (ADT) for non-metastatic prostate cancer in 19 countries. *ESMO Open* **1**, e000040 (2016).

48. M. Hussain *et al.*, Intermittent versus continuous androgen deprivation in prostate cancer. *N Engl J Med* **368**, 1314-1325 (2013).

49. J. M. Crook *et al.*, Intermittent androgen suppression for rising PSA level after radiotherapy. *N Engl J Med* **367**, 895-903 (2012).

50. N. D. Shore, E. D. Crawford, Intermittent androgen deprivation therapy: redefining the standard of care? *Rev Urol* **12**, 1-11 (2010).

51. P. Ghadjar *et al.*, Use of androgen deprivation and salvage radiation therapy for patients with prostate cancer and biochemical recurrence after prostatectomy. *Strahlenther Onkol* **194**, 619-626 (2018).

52. E. Schaeffer *et al.*, NCCN Guidelines Insights: Prostate Cancer, Version 1.2021. *J Natl Compr Canc Netw* **19**, 134-143 (2021).

53. S. E. Cotter *et al.*, Salvage radiation in men after prostate-specific antigen failure and the risk of death. *Cancer* **117**, 3925-3932 (2011).

54. B. J. Trock *et al.*, Prostate cancer-specific survival following salvage radiotherapy vs observation in men with biochemical recurrence after radical prostatectomy. *JAMA* **299**, 2760-2769 (2008).

55. S. A. Boorjian *et al.*, Radiation therapy after radical prostatectomy: impact on metastasis and survival. *J Urol* **182**, 2708-2714 (2009).

56. L. L. Pisters *et al.*, Locally recurrent prostate cancer after initial radiation therapy: a comparison of salvage radical prostatectomy versus cryotherapy. *J Urol* **182**, 517-525; discussion 525-517 (2009).

57. G. L. Lu-Yao *et al.*, Survival following primary androgen deprivation therapy among men with localized prostate cancer. *JAMA* **300**, 173-181 (2008).

58. E. D. Crawford *et al.*, Prostate-specific antigen 1.5-4.0 ng/mL: a diagnostic challenge and danger zone. *BJU Int* **108**, 1743-1749 (2011).

59. J. R. Prensner, M. A. Rubin, J. T. Wei, A. M. Chinnaiyan, Beyond PSA: the next generation of prostate cancer biomarkers. *Sci Transl Med* **4**, 127rv123 (2012).

60. J. A. Cohn *et al.*, Primary care physician PSA screening practices before and after the final U.S. Preventive Services Task Force recommendation. *Urol Oncol* **32**, 41 e23-30 (2014).

61. N. D. S. E. David Crawford, Daniel P. Petrylak, Leonard G. Gomella, Neil H. Baum, Francisco G. La Rosa, Michael Leapman, Tim Langford, Paul Arangua, Jessica Myers-Schecter. (2017).

62. L. S. Marks *et al.*, PCA3 molecular urine assay for prostate cancer in men undergoing repeat biopsy. *Urology* **69**, 532-535 (2007).

63. H. Nakanishi *et al.*, PCA3 molecular urine assay correlates with prostate cancer tumor volume: implication in selecting candidates for active surveillance. *J Urol* **179**, 1804-1809; discussion 1809-1810 (2008).

64. P. R. Carroll *et al.*, NCCN Guidelines Insights: Prostate Cancer Early Detection, Version 2.2016. *J Natl Compr Canc Netw* **14**, 509-519 (2016).

65. S. A. Tomlins *et al.*, Urine TMPRSS2:ERG Plus PCA3 for Individualized Prostate Cancer Risk Assessment. *Eur Urol* **70**, 45-53 (2016).

66. M. Auprich *et al.*, Contemporary role of prostate cancer antigen 3 in the management of prostate cancer. *Eur Urol* **60**, 1045-1054 (2011).

67. K. C. Cary, M. R. Cooperberg, Biomarkers in prostate cancer surveillance and screening: past, present, and future. *Ther Adv Urol* **5**, 318-329 (2013).

68. M. Truong, B. Yang, D. F. Jarrard, Toward the detection of prostate cancer in urine: a critical analysis. *J Urol* **189**, 422-429 (2013).

69. A. Haese *et al.*, Multicenter Optimization and Validation of a 2-Gene mRNA Urine Test for Detection of Clinically Significant Prostate Cancer before Initial Prostate Biopsy. *J Urol* **202**, 256-263 (2019).

70. D. J. Parekh *et al.*, A multi-institutional prospective trial in the USA confirms that the 4Kscore accurately identifies men with high-grade prostate cancer. *Eur Urol* **68**, 464-470 (2015).

71. S. Punnen *et al.*, A Multi-Institutional Prospective Trial Confirms Noninvasive Blood Test Maintains Predictive Value in African American Men. *J Urol* **199**, 1459-1463 (2018).

72. L. J. Sokoll *et al.*, A prospective, multicenter, National Cancer Institute Early Detection Research Network study of [-2]proPSA: improving prostate cancer detection and correlating with cancer aggressiveness. *Cancer Epidemiol Biomarkers Prev* **19**, 1193-1200 (2010).

73. S. Loeb, W. J. Catalona, The Prostate Health Index: a new test for the detection of prostate cancer. *Ther Adv Urol* **6**, 74-77 (2014).

74. X. Wang *et al.*, Autoantibody signatures in prostate cancer. *N Engl J Med* **353**, 1224-1235 (2005).

75. M. Schipper, G. Wang, N. Giles, J. Ohrnberger, Novel prostate cancer biomarkers derived from autoantibody signatures. *Transl Oncol* **8**, 106-111 (2015).

76. R. Henrique, C. Jeronimo, Molecular detection of prostate cancer: a role for GSTP1 hypermethylation. *Eur Urol* **46**, 660-669; discussion 669 (2004).

77. J. D. Kronz, C. H. Allan, A. A. Shaikh, J. I. Epstein, Predicting cancer following a diagnosis of high-grade prostatic intraepithelial neoplasia on needle biopsy: data on men with more than one follow-up biopsy. *Am J Surg Pathol* **25**, 1079-1085 (2001).

78. M. S. Geybels *et al.*, Epigenomic profiling of DNA methylation in paired prostate cancer versus adjacent benign tissue. *Prostate* **75**, 1941-1950 (2015).

79. K. J. Wojno *et al.*, Reduced Rate of Repeated Prostate Biopsies Observed in ConfirmMDx Clinical Utility Field Study. *Am Health Drug Benefits* **7**, 129-134 (2014).

80. W. Aubry *et al.*, Budget impact model: epigenetic assay can help avoid unnecessary repeated prostate biopsies and reduce healthcare spending. *Am Health Drug Benefits* **6**, 15-24 (2013).

81. E. A. Klein *et al.*, A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur Urol* **66**, 550-560 (2014).

82. K. K. Badani *et al.*, Effect of a genomic classifier test on clinical practice decisions for patients with high-risk prostate cancer after surgery. *BJU Int* **115**, 419-429 (2015).

83. M. L. Tsai *et al.*, Utility of Oncotype DX Risk Assessment in Patients With Invasive Lobular Carcinoma. *Clin Breast Cancer* **16**, 45-50 (2016).

84. E. S. Antonarakis *et al.*, AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. *N Engl J Med* **371**, 1028-1038 (2014).

85. H. I. Scher *et al.*, Nuclear-specific AR-V7 Protein Localization is Necessary to Guide Treatment Selection in Metastatic Castration-resistant Prostate Cancer. *Eur Urol* **71**, 874-882 (2017).

86. H. I. Scher *et al.*, Assessment of the Validity of Nuclear-Localized Androgen Receptor Splice Variant 7 in Circulating Tumor Cells as a Predictive Biomarker for Castration-Resistant Prostate Cancer. *JAMA Oncol* **4**, 1179-1186 (2018).

87. O. Health Quality, Prolaris Cell Cycle Progression Test for Localized Prostate Cancer: A Health Technology Assessment. *Ont Health Technol Assess Ser* **17**, 1-75 (2017).

88. J. Cuzick *et al.*, 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 (2011).

89. M. B. Warf *et al.*, Analytical validation of a proliferation-based molecular signature used as a prognostic marker in early stage lung adenocarcinoma. *Biomark Med* **9**, 901-910 (2015).

90. S. J. Freedland *et al.*, Prognostic utility of cell cycle progression score in men with prostate cancer after primary external beam radiation therapy. *Int J Radiat Oncol Biol Phys* **86**, 848-853 (2013).

91. J. Cuzick *et al.*, Validation of an RNA cell cycle progression score for predicting death from prostate cancer in a conservatively managed needle biopsy cohort. *Br J Cancer* **113**, 382-389 (2015).

92. N. Erho *et al.*, Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* **8**, e66855 (2013).

93. E. A. Klein *et al.*, Decipher Genomic Classifier Measured on Prostate Biopsy Predicts Metastasis Risk. *Urology* **90**, 148-152 (2016).

94. C. Meldrum, M. A. Doyle, R. W. Tothill, Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin Biochem Rev* **32**, 177-195 (2011).

95. H. Chen *et al.*, Adding genetic risk score to family history identifies twice as many high-risk men for prostate cancer: Results from the prostate cancer prevention trial. *Prostate* **76**, 1120-1129 (2016).

96. C. A. Conran *et al.*, Population-standardized genetic risk score: the SNP-based method of choice for inherited risk assessment of prostate cancer. *Asian J Androl* **18**, 520-524 (2016).

97. T. J. Hoffmann *et al.*, A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. *Cancer Discov* **5**, 878-891 (2015).

98. V. N. Giri *et al.*, Role of Genetic Testing for Inherited Prostate Cancer Risk: Philadelphia Prostate Cancer Consensus Conference 2017. *J Clin Oncol* **36**, 414-424 (2018).

99. R. Na *et al.*, Germline Mutations in ATM and BRCA1/2 Distinguish Risk for Lethal and Indolent Prostate Cancer and are Associated with Early Age at Death. *Eur Urol* **71**, 740-747 (2017).

100. M. S. Lucia *et al.*, Pathologic characteristics of cancers detected in The Prostate Cancer Prevention Trial: implications for prostate cancer detection and chemoprevention. *Cancer Prev Res (Phila)* **1**, 167-173 (2008).

101. A. Bird, The essentials of DNA methylation. *Cell* **70**, 5-8 (1992).

102. J. E. Dodge, B. H. Ramsahoye, Z. G. Wo, M. Okano, E. Li, De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene* **289**, 41-48 (2002).

103. R. Lister *et al.*, Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).

104. Z. D. Smith, A. Meissner, DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**, 204-220 (2013).

105. R. Lister *et al.*, Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).

106. T. R. Haines, D. I. Rodenhiser, P. J. Ainsworth, Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Dev Biol* **240**, 585-598 (2001).

107. M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261-282 (1987).

108. R. S. Illingworth *et al.*, Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* **6**, e1001134 (2010).

109. S. Saxonov, P. Berg, D. L. Brutlag, A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412-1417 (2006).

110. T. Chen, E. Li, Structure and function of eukaryotic DNA methyltransferases. *Curr Top Dev Biol* **60**, 55-89 (2004).

111. G. C. Vieira, M. F. D'Avila, R. Zanini, M. Depra, V. L. da Silva Valente, Evolution of DNMT2 in drosophilids: Evidence for positive and purifying selection and insights into new protein (pathways) interactions. *Genet Mol Biol* **41**, 215-234 (2018).

112. M. Tahiliani *et al.*, Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935 (2009).

113. S. Ito *et al.*, Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-1133 (2010).

114. Y. F. He *et al.*, Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307 (2011).

115. A. Merlo *et al.*, 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat Med* **1**, 686-692 (1995).

116. J. R. Dobosy, J. L. Roberts, V. X. Fu, D. F. Jarrard, The expanding role of epigenetics in the development, diagnosis and treatment of prostate cancer and benign prostatic hyperplasia. *J Urol* **177**, 822-831 (2007).

117. V. A. Moyer, U. S. P. S. T. Force, Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med* **157**, 120-134 (2012).

118. W. A. Schulz, M. J. Hoffmann, Epigenetic mechanisms in the biology of prostate cancer. *Semin Cancer Biol* **19**, 172-180 (2009).

119. R. Zelic *et al.*, Global DNA hypomethylation in prostate cancer development and progression: a systematic review. *Prostate Cancer Prostatic Dis* **18**, 1-12 (2015).

120. M. Fraser *et al.*, Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359-364 (2017).

121. G. Gurioli *et al.*, GSTP1 methylation in cancer: a liquid biopsy biomarker? *Clin Chem Lab Med* **56**, 702-717 (2018).

122. E. E. Holmes *et al.*, PITX3 promoter methylation is a prognostic biomarker for biochemical recurrence-free survival in prostate cancer patients after radical prostatectomy. *Clin Epigenetics* **8**, 104 (2016).

123. M. K. Kirby *et al.*, Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns. *BMC Cancer* **17**, 273 (2017).

124. M. O. Hoque, DNA methylation changes in prostate cancer: current developments and future clinical implementation. *Expert Rev Mol Diagn* **9**, 243-257 (2009).

125. M. O. Hoque *et al.*, Quantitative methylation-specific polymerase chain reaction gene patterns in urine sediment distinguish prostate cancer patients from control subjects. *J Clin Oncol* **23**, 6569-6575 (2005).

126. C. S. Cooper, C. S. Foster, Concepts of epigenetics in prostate cancer development. *Br J Cancer* **100**, 240-245 (2009).

127. W. G. Nelson *et al.*, Abnormal DNA methylation, epigenetics, and prostate cancer. *Front Biosci* **12**, 4254-4266 (2007).

128. Y. Han, J. Xu, J. Kim, X. Wu, J. Gu, LINE-1 methylation in peripheral blood leukocytes and clinical characteristics and prognosis of prostate cancer patients. *Oncotarget* **8**, 94020-94027 (2017).

129. Y. Han, J. Xu, J. Kim, X. Wu, J. Gu, Methylation of subtelomeric repeat D4Z4 in peripheral blood leukocytes is associated with biochemical recurrence in localized prostate cancer patients. *Carcinogenesis* **38**, 821-826 (2017).

130. S. R. Sturgeon *et al.*, White blood cell DNA methylation and risk of breast cancer in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). *Breast Cancer Res* **19**, 94 (2017).

131. J. Xu *et al.*, Methylation of global DNA repeat LINE-1 and subtelomeric DNA repeats D4Z4 in leukocytes is associated with biochemical recurrence in African American prostate cancer patients. *Carcinogenesis*, (2019).

132. L. Li *et al.*, DNA methylation in peripheral blood: a potential biomarker for cancer molecular epidemiology. *J Epidemiol* **22**, 384-394 (2012).

133. I. Florath, K. Butterbach, H. Muller, M. Bewerunge-Hudler, H. Brenner, Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum Mol Genet*, (2013).

134. H. Heyn *et al.*, DNA methylation contributes to natural human variation. *Genome Res* **23**, 1363-1372 (2013).

135. I. Illumina, Infinium® HumanMethylation450 BeadChip. (2012).

136. J. H. Luo *et al.*, Genome-wide methylation analysis of prostate tissues reveals global methylation patterns of prostate cancer. *Am J Pathol* **182**, 2028-2036 (2013).

137. Y. Wang *et al.*, Genome-Wide Methylation Patterns in Androgen-Independent Prostate Cancer Cells: A Comprehensive Analysis Combining MeDIP-Bisulfite, RNA, and microRNA Sequencing Data. *Genes (Basel)* **9**, (2018).

138. Y. P. Yu *et al.*, Whole-genome methylation sequencing reveals distinct impact of differential methylations on gene transcription in prostate cancer. *Am J Pathol* **183**, 1960-1970 (2013).

139. K. Chiam, C. Ricciardelli, T. Bianco-Miotto, Epigenetic biomarkers in prostate cancer: Current and future uses. *Cancer Lett* **342**, 248-256 (2014).

140. C. Jeronimo, R. Henrique, Epigenetic biomarkers in urological tumors: A systematic review. *Cancer Lett* **342**, 264-274 (2014).

141. P. Cairns *et al.*, Molecular detection of prostate cancer in urine by GSTP1 hypermethylation. *Clin Cancer Res* **7**, 2727-2730 (2001).

142. M. Esteller, Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **8**, 286-298 (2007).

143. L. Delgado-Cruzata *et al.*, DNA methylation changes correlate with Gleason score and tumor stage in prostate cancer. *DNA Cell Biol* **31**, 187-192 (2012).

144. W. Goering, M. Kloth, W. A. Schulz, DNA methylation changes in prostate cancer. *Methods Mol Biol* **863**, 47-66 (2012).

145. C. Chao, M. Chi, M. Preciado, M. H. Black, Methylation markers for prostate cancer prognosis: a systematic review. *Cancer causes & control : CCC* **24**, 1615-1641 (2013).

146. A. S. Perry, R. W. Watson, M. Lawler, D. Hollywood, The epigenome as a therapeutic target in prostate cancer. *Nature reviews. Urology* **7**, 668-680 (2010).

147. C. Jeronimo *et al.*, Epigenetics in prostate cancer: biologic and clinical relevance. *Eur Urol* **60**, 753-766 (2011).

148. C. Haldrup *et al.*, DNA methylation signatures for prediction of biochemical recurrence after radical prostatectomy of clinically localized prostate cancer. *J Clin Oncol* **31**, 3250-3258 (2013).

149. C. J. Marsit *et al.*, DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. *J Clin Oncol* **29**, 1133-1139 (2011).

150. J. T. Bell *et al.*, Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet* **8**, e1002629 (2012).

151. V. K. Cortessis *et al.*, Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. *Hum Genet* **131**, 1565-1589 (2012).

152. D. C. Koestler *et al.*, Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. *Cancer Epidemiol Biomarkers Prev* **21**, 1293-1302 (2012).

153. C. P. Lange *et al.*, Genome-scale discovery of DNA-methylation biomarkers for blood-based detection of colorectal cancer. *PLoS One* **7**, e50266 (2012).

154. P. C. Tsai, T. D. Spector, J. T. Bell, Using epigenome-wide association scans of DNA methylation in age-related complex human traits. *Epigenomics* **4**, 511-526 (2012).

155. E. S. Wan *et al.*, Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. *Hum Mol Genet* **21**, 3073-3082 (2012).

156. D. C. Koestler *et al.*, Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics : official journal of the DNA Methylation Society* **8**, 816-826 (2013).

157. K. B. Michels *et al.*, Recommendations for the design and analysis of epigenome-wide association studies. *Nature methods* **10**, 949-955 (2013).

158. N. S. Shenker *et al.*, Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. *Hum Mol Genet* **22**, 843-851 (2013).

159. L. Wang *et al.*, Methylation markers for small cell lung cancer in peripheral blood leukocyte DNA. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer* **5**, 778-785 (2010).

160. Z. Xu *et al.*, Epigenome-wide association study of breast cancer using prospectively collected sister study samples. *J Natl Cancer Inst* **105**, 694-700 (2013).

161. M. Campan *et al.*, Genome-scale screen for DNA methylation-based detection markers for ovarian cancer. *PLoS One* **6**, e28141 (2011).

162. X. Xu *et al.*, A genome-wide methylation study on obesity: differential variability and differential methylation. *Epigenetics : official journal of the DNA Methylation Society* **8**, 522-533 (2013).

163. L. P. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am J Hum Genet* **88**, 450-457 (2011).

164. J. L. Garcia-Gimenez *et al.*, Epigenetic biomarkers: Current strategies and future challenges for their use in the clinical laboratory. *Crit Rev Clin Lab Sci* **54**, 529-550 (2017).

165. C. Glinge *et al.*, Stability of Circulating Blood-Based MicroRNAs - Pre-Analytic Methodological Considerations. *PLoS One* **12**, e0167969 (2017).

166. L. Peiro-Chova *et al.*, High stability of microRNAs in tissue samples of compromised quality. *Virchows Arch* **463**, 765-774 (2013).

167. J. E. Joo *et al.*, The use of DNA from archival dried blood spots with the Infinium HumanMethylation450 array. *BMC Biotechnol* **13**, 23 (2013).

168. C. L. Relton, F. P. Hartwig, G. Davey Smith, From stem cells to the law courts: DNA methylation, the forensic epigenome and the possibility of a biosocial archive. *Int J Epidemiol* **44**, 1083-1093 (2015).

169. M. A. Liss, Infection: prostate biopsy-infection and prior fluoroquinolone exposure. *Nat Rev Urol* **8**, 592-594 (2011).

170. R. K. Nam *et al.*, Increasing hospital admission rates for urological complications after transrectal ultrasound guided prostate biopsy. *J Urol* **189**, S12-17; discussion S17-18 (2013).

171. Y. Wang *et al.*, Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord. *Proc Natl Acad Sci U S A* **112**, 9704-9709 (2015).

172. A. R. Rastinehad *et al.*, Improving detection of clinically significant prostate cancer: magnetic resonance imaging/transrectal ultrasound fusion guided prostate biopsy. *J Urol* **191**, 1749-1754 (2014).

173. M. S. Geybels *et al.*, PTEN loss is associated with prostate cancer recurrence and alterations in tumor DNA methylation profiles. *Oncotarget* **8**, 84338-84348 (2017).

174. A. Eden, F. Gaudet, A. Waghmare, R. Jaenisch, Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**, 455 (2003).

175. J. Breivik, G. Gaudernack, Genomic instability, DNA methylation, and natural selection in colorectal carcinogenesis. *Semin Cancer Biol* **9**, 245-254 (1999).

176. H. D. Woo, J. Kim, Global DNA hypomethylation in peripheral blood leukocytes as a biomarker for cancer risk: a meta-analysis. *PLoS One* **7**, e34615 (2012).

177. M. Barchitta, A. Quattrocchi, A. Maugeri, M. Vinciguerra, A. Agodi, LINE-1 hypomethylation in blood and tissue samples as an epigenetic marker for cancer risk: a systematic review and meta-analysis. *PLoS One* **9**, e109478 (2014).

178. J. M. Flanagan *et al.*, Platinum-Based Chemotherapy Induces Methylation Changes in Blood DNA Associated with Overall Survival in Patients with Ovarian Cancer. *Clin Cancer Res* **23**, 2213-2222 (2017).

179. P. A. Dugue *et al.*, DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. *Int J Cancer* **142**, 1611-1619 (2018).

180. L. Dong, H. Ren, Blood-based DNA Methylation Biomarkers for Early Detection of Colorectal Cancer. *J Proteomics Bioinform* **11**, 120-126 (2018).

181. S. Ambatipudi *et al.*, DNA methylation derived systemic inflammation indices are associated with head and neck cancer development and survival. *Oral Oncol* **85**, 87-94 (2018).

182. J. E. Joo *et al.*, Heritable DNA methylation marks associated with susceptibility to breast cancer. *Nat Commun* **9**, 867 (2018).

183. B. T. Joyce *et al.*, DNA Methylation of Telomere-Related Genes and Cancer Risk. *Cancer Prev Res (Phila)* **11**, 511-522 (2018).

184. P. A. Dugue *et al.*, Heritable methylation marks associated with breast and prostate cancer risk. *Prostate* **78**, 962-969 (2018).

185. L. M. FitzGerald *et al.*, Genome-Wide Measures of Peripheral Blood Dna Methylation and Prostate Cancer Risk in a Prospective Nested Case-Control Study. *Prostate* **77**, 471-478 (2017).

186. K. H. Barry *et al.*, Prospective study of DNA methylation at chromosome 8q24 in peripheral blood and prostate cancer risk. *Br J Cancer* **116**, 1470-1479 (2017).

187. J. Sandoval *et al.*, Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692-702 (2011).

188. T. J. Morris *et al.*, ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428-430 (2014).

189. M. J. Aryee *et al.*, Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-1369 (2014).

190. A. E. Teschendorff *et al.*, A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-196 (2013).

191. E. A. Houseman *et al.*, DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

192. D. Sun *et al.*, MOABS: model based analysis of bisulfite sequencing data. *Genome Biol* **15**, R38 (2014).

193. I. Diboun, L. Wernisch, C. A. Orengo, M. Koltzenburg, Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genomics* **7**, 252 (2006).

194. W. Koehrsen. Random Forest Simple Explanation (2017).

195. U. Baron *et al.*, Epigenetic immune cell counting in human blood samples for immunodiagnostics. *Sci Transl Med* **10**, (2018).

196. F. Murtagh, P. Legendre, Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* **31**, 274-295 (2014).

197. B. C. Sheu *et al.*, Reversed CD4/CD8 ratios of tumor-infiltrating lymphocytes are correlated with the progression of human cervical carcinoma. *Cancer* **86**, 1537-1543 (1999).

198. J. C. Hu *et al.*, Impact of H19 Polymorphisms on Prostate Cancer Clinicopathologic Characteristics. *Diagnostics (Basel)* **10**, (2020).

199. M. Kluth *et al.*, Prevalence of chromosomal rearrangements involving non-ETS genes in prostate cancer. *Int J Oncol* **46**, 1637-1642 (2015).

200. D. J. Joseph R. Dobosy, Malije Onwueme, Josh Desotelle, Vivian Fu, Expression of the imprinted human Peg1/MEST gene is altered in senescent prostate cells and prostate cancer tissues. *Proc Amer Assoc Cancer Res* **46**, (2005).

201. S. Zhu *et al.*, PRDM16 is associated with evasion of apoptosis by prostatic cancer cells according to RNA interference screening. *Mol Med Rep* **14**, 3357-3361 (2016).

202. H. Matthaei *et al.*, GNAS codon 201 mutations are uncommon in intraductal papillary neoplasms of the bile duct. *HPB (Oxford)* **14**, 677-683 (2012).

203. A. J. Parish *et al.*, GNAS, GNAQ, and GNA11 alterations in patients with diverse cancers. *Cancer* **124**, 4080-4089 (2018).

204. D. Robinson *et al.*, Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **162**, 454 (2015).

205. M. V. C. Greenberg, D. Bourc'his, The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* **20**, 590-607 (2019).

206. M. Weber *et al.*, Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* **39**, 457-466 (2007).

207. A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916-919 (2010).

208. X. J. Yang *et al.*, Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell* **26**, 577-590 (2014).

209. C. Francastel, F. Magdinier, DNA methylation in satellite repeats disorders. *Essays Biochem* **63**, 757-771 (2019).

210. A. Meissner *et al.*, Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* **33**, 5868-5877 (2005).

211. M. Kernaleguen *et al.*, Whole-Genome Bisulfite Sequencing for the Analysis of Genome-Wide DNA Methylation and Hydroxymethylation Patterns at Single-Nucleotide Resolution. *Methods Mol Biol* **1767**, 311-349 (2018).

212. Y. Xi *et al.*, RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics* **28**, 430-432 (2012).

213. X. Lin *et al.*, BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics* **29**, 3227-3229 (2013).

214. D. Sun *et al.*, Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell Stem Cell* **14**, 673-688 (2014).

215. M. J. Booth *et al.*, Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc* **8**, 1841-1851 (2013).

216. M. Yu *et al.*, Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368-1380 (2012).

217. Y. Huang, W. A. Pastor, J. A. Zepeda-Martinez, A. Rao, The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat Protoc* **7**, 1897-1908 (2012).

218. E. Afgan *et al.*, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* **46**, W537-W544 (2018).

219. S. Andrews. (Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010).

220. P. Ewels, M. Magnusson, S. Lundin, M. Kaller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016).

221. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

222. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

223. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).

224. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

225. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

226. J. Calcote, *Autotools: a practitioner's guide to GNU autoconf, automake, and libtool*. (No Starch Press, 2019).

227. B. Gruning *et al.*, Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **15**, 475-476 (2018).

228. Y. H. Huang *et al.*, DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biology* **18**, (2017).

229. C. Zhang, Y. Hoshida, K. C. Sadler, Comparative Epigenomic Profiling of the DNA Methylome in Mouse and Zebrafish Uncovers High Interspecies Divergence. *Front Genet* **7**, 110 (2016).

230. L. T. Zhou *et al.*, Feature selection and classification of urinary mRNA microarray data by iterative random forest to diagnose renal fibrosis: a two-stage study. *Sci Rep* **7**, 39832 (2017).

231. J. Naue *et al.*, Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. *Forensic Sci Int Genet* **31**, 19-28 (2017).

232. M. B. Kursa, Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics* **15**, 8 (2014).

233. A. Anaissi, P. J. Kennedy, M. Goyal, D. R. Catchpoole, A balanced iterative random forest for gene selection from microarray data. *BMC Bioinformatics* **14**, 261 (2013).

234. R. Diaz-Uriarte, S. Alvarez de Andres, Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).

235. K. Moorthy, M. S. Mohamad, Random forest for gene selection and microarray data classification. *Bioinformation* **7**, 142-146 (2011).

236. H. Tariq, E. Eldridge, I. Welch, An efficient approach for feature construction of high-dimensional microarray data by random projections. *PLoS One* **13**, e0196385 (2018).