Epigenetic Study of Plasma Circulating DNA in Prostate Cancer

by

Andrew Kwan

A thesis submitted in conformity with the requirements for the degree of Master of Science
Graduate Department of Pharmacology and Toxicology
University of Toronto

© Copyright by Andrew Kwan (2015)
Abstract

Malignant cells exhibit epigenetic differences compared to normal cells, which can also be detected in the cell-free circulating DNA (cirDNA) of bodily fluids. DNA modification studies in cirDNA may lead to the development of non-invasive cancer biomarkers. This thesis investigates plasma cirDNA modifications of prostate cancer patients, benign prostate hyperplasia patients, and men without any known prostate disease. A modification-sensitive restriction enzyme approach was used to enrich the modified DNA fractions, which were interrogated on CpG island- and Affymetrix tiling microarrays. Our initial discovery screen identified a number of differentially modified genes and non-coding loci between PCa patients and controls; a subset of candidate loci were verified using bisulfite pyrosequencing. We applied machine-learning techniques to develop a multi-locus biomarker set, which exhibited promising predictive performance in distinguishing cancer patients from unaffected controls. Our results suggest that epigenomics of cirDNA is a promising source for diagnostic non-invasive biomarkers of solid tumors.
Acknowledgments

I would like to start by expressing my gratitude to my supervisor, Dr. Arturas Petronis, for giving me the opportunity to work under his mentorship and guidance for the past few years. It is under his tutelage that I have grown both as a person and a scientist. I will also be forever indebted to the supportive post-doctoral fellows that I have had the chance to work with, Dr. Rene Cortese and Dr. Tarang Khare. Their friendship and care for me – extending well outside of the lab setting – have inspired me to develop mentorship skills and has truly made the lab feel like a second home. To each member of the Krembil Epigenetics Laboratory, the memories have been truly unforgettable and I thank each of you for your friendship on this journey. My thanks also go out to my committee members, Dr. Albert Wong and Dr. Jose Nobrega for their invaluable advice throughout my program.

Further, I would like to acknowledge our collaborators, Dr. Paul Boutros, Emilie Lalonde, and Cindy Yao for performing the bioinformatics on our experimental work. I would also like to acknowledge Vanier CGS for providing me with the financial support to pursue my degree.

Last but certainly not least, I owe all of my successes to my family and friends for their love and unending support throughout these years. I would not have been able to accomplish any of this without all of you.
Table of Contents

Acknowledgments ................................................................................................................................................ iii
Table of Contents ................................................................................................................................................ iv
List of Tables ....................................................................................................................................................... vii
List of Figures ...................................................................................................................................................... viii
List of Appendices ............................................................................................................................................... ix
Abbreviations ...................................................................................................................................................... x

Chapter 1 Introduction ...................................................................................................................................... 1

1 Introduction ...................................................................................................................................................... 1
   1.1 Statement of research problem ..................................................................................................................... 1
   1.2 Review of Literature .................................................................................................................................... 2
      1.2.1 Epigenetics ........................................................................................................................................... 2
      1.2.2 Cancer biomarkers ............................................................................................................................... 6
      1.2.3 Circulating DNA (cirDNA) .................................................................................................................. 8
      1.2.4 Prostate cancer ...................................................................................................................................... 17
   1.3 Study objectives and rationale for hypotheses ............................................................................................. 25

Chapter 2 Materials and Methods .................................................................................................................. 26

2 Materials and Methods ................................................................................................................................ 26
   2.1 Samples ....................................................................................................................................................... 26
      2.1.1 Prostate cancer pilot study ..................................................................................................................... 26
      2.1.2 Prostate cancer tiling microarray study ................................................................................................. 26
   2.2 DNA extraction ............................................................................................................................................ 27
   2.3 DNA modification amplicon preparation ..................................................................................................... 27
      2.3.1 Principle of DNA modification detection ............................................................................................... 27
      2.3.2 DNA blunting ....................................................................................................................................... 29
4.1.4 Changes in repetitive DNA elements in cancer .................................................. 77
4.1.5 Primary PCa tumor analysis ............................................................................... 78
4.2 PCa tiling microarray study .................................................................................. 79
  4.2.1 DNA modification profiling on high resolution tiling arrays ......................... 79
  4.2.2 Predictive performance of epigenetic signature ........................................... 80
  4.2.3 Gene ontology analysis .................................................................................. 82
4.3 Limitations and recommendations for future studies ........................................... 84
4.4 Summary of Findings ......................................................................................... 86
References .................................................................................................................. 88
Appendices .................................................................................................................. 116
Copyright Acknowledgements ................................................................................. Error! Bookmark not defined.
List of Tables

Table 1. Prostate cancer tiling array study cohort description .............................................. 27
Table 2. Primer sequences for pyrosequencing analysis ......................................................... 36
Table 3. Primer sequences for pyrosequencing analysis ......................................................... 36
Table 4. Loci selected for fine mapping of modified CpG positions ....................................... 50
Table 5. Summary of pyrosequencing results ........................................................................ 51
List of Figures

Figure 1. Cell-free nucleic acids in the blood. ................................................................. 10

Figure 2. Principle of DNA modification detection technology in plasma cirDNA. ............. 28

Figure 3. Method optimization for cirDNA target preparation........................................... 31

Figure 4. Building of an “epigenetic signature” for prostate cancer screening... ................. 42

Figure 5. Differentially modified regions in cirDNA in prostate cancer and control individuals. 46

Figure 6. Differentially modified regions in cirDNA in prostate cancer and control individuals. 47

Figure 7. Fine mapping of modified cytosines in the \textit{RNF219} gene in prostate cancer and control individuals.......................................................... 52

Figure 8. DNA modification in the \textit{RNF219} gene in two independent sample sets............. 53

Figure 9. Prediction accuracy of the cirDNA modification level in the \textit{RNF219} locus .......... 55

Figure 10. Clusters of differential microarray signals identified in pericentromeric regions...... 57

Figure 11. Machine-learning analysis of microarray-based cirDNA modification profiles........ 59

Figure 12. Overlap between cirDNA analysis and public primary prostate tissue analyses. ...... 61

Figure 13. Differential DNA modification and mRNA expression of \textit{ARHGAP6} and \textit{CDK6} in primary patient tissue................................................................. 62

Figure 14. cirDNA modification profiling in PCa and control individuals. ......................... 64

Figure 15. cirDNA modification profiling in PCa and control individuals (continued)......... 66

Figure 16. Epigenetic signature as diagnostic biomarker for PCa........................................ 70

Figure 17. Epigenetic signature as prognostic biomarker in PCa....................................... 71
List of Appendices

Figure S1. Flowchart of machine learning analysis................................................................. 111

Table S1: Binding motifs for transcription factors associated to regions of differential cirDNA modification................................................................. 113

Table S2: Genes exhibiting cirDNA modification differences and steady-state mRNA and DNA modification differences in primary tumours.................. ................................................................. 120
Abbreviations

5-carboxylyctosine (5-caC)
5-formylcytosine (5-fC)
5-hydroxymethylcytosine (5-hmC)
5-methylcytosine (5mC)

Adenomatous polyposis coli (APC)

Alpha-methylacyl-CoA racemase (AMACR)

Area under the curve (AUC)

Balanced Accurate rate (BAC)

Benign prostatic hyperplasia (BPH)

Bridging Integrator 1 (BIN1)

Breast cancer 1, early onset gene (BRCA1)

BMI1 polycomb ring finger oncogene (BMI1).

C- myelocytic leukaemia (c-MYC)

Circulating DNA (cirDNA)

Copy number variation (CNV)

Cyclin-dependent kinase 6 (CDK6),

Database for Annotation, Visualization and Integrated Discovery (DAVID)

Death-associated protein kinase 1 (DAPK)

DeoxyriboNucleotide TriPhosphate (dNTP)
Digital rectal examinations (DRE)

Discs large homolog 2 (DLG2)

DNA methyltransferase enzymes (DNMTs)

Erythroblastosis related gene (ERG)

E-twenty-six (ETS)

European Randomized Study of Screening for Prostate Cancer (ERSPC)

False negative (FN)

False positive (FP)

False-positive rate (FPR)

Gene ontology (GO)

Gleason Score (GS)

Glutathione S-transferase pi 1 (GSTP1)

Glutathione S-transferases (GSTs)

Guanine nucleotide binding protein (G protein) gamma 7 (GNG7)

Heparanase 2 (HPSE2)

Hidden markov model (HMM)

KIAA1539 protein (KIAA1539)

Leave-one-out cross-validation (LOOCV)

Linear mixed-effect (LME)

Long interspersed element 1 (LINE1 or L1)
Loss of heterozygosity (LOH)

Low in Lung Cancer 1 (LLCI)

Maternally expressed 3 (MEG3)

Methylation-specific PCR (MSP)

Microsatellite instability (MSI)

NAD-dependent protein deacetylase sirtuin-2 (SIRT2),

Negative predictive value (NPV)

NudC domain containing 3 (NUDCD3)

Growth factor independent-1 (GFI1)

Polymerase chain reaction (PCR)

Positive predictive value (PPV)

Princess Margaret Cancer Centre (PM)

Prostate cancer (PCa)

Prostate cancer antigen 3 (PCA3),

Prostate specific antigen (PSA)

Protocadherin beta 1 (PCDHB1)

Pyrophosphate (PPI).

Pituitary tumor-transforming protein 1 (PTTG1)

Quantitative methylation specific PCR (qMSP)

Quantitative PCR (qPCR)
Ras association domain family member 1 (RASSF1A)

Receiver operating characteristic (ROC)

Recurrence-free survival (RFS)

Ring finger protein 219 (RNF219)

Rho GTPase-activating protein 6 isoform 4 (ARHGAP6)

S-adenosyl-methionine (SAM)

Ten-eleven translocation (TET)

Transcription Factor 3 (TCF3)

Transient receptor potential cation channel, subfamily V, member 6 (TRPV6)

Transmembrane protease serine 2 (TMPRSS2)

Trans-rectal ultrasonography (TRUS)

True negative (TN)

True positive (TP)

True positive rate (TPR)

Tumour node metastasis (TNM)
Chapter 1
Introduction

1 INTRODUCTION

1.1 Statement of research problem

According to the American National Cancer Institute, cancer is defined as a disease that is characterized by uncontrolled proliferation of abnormal cells. Affecting nearly 14 million new individuals in 2012, cancer was responsible for causing 8.2 million deaths – accounting for 15% of all deaths that year (Bray et al., 2013). Cancer undoubtedly remains one of the key health problems that many research groups are struggling to address. Currently, detection of cancer in its early stages represents the most effective way to improve prognosis, as this allows confined tumors to be more effectively treated and cured (Bremnes et al., 2005). As a result, there is a pressing need for accurate screening tests of asymptomatic populations to aid in patient management and to decrease mortality.

In current clinical settings, cancer diagnosis and prognosis depends on the histological assessment of cancer cells from a biopsy performed after positive screening indications. In prostate cancer (PCa), detection of elevated prostate specific antigen (PSA) levels in serum remains the most common marker used for the screening of prostate cancer, though recently, a number of questions have been raised concerning its diagnostic performance and clinical value (Arneth, 2009; Barqawi et al., 2012; Bell et al., 2014). With increased PSA levels also found in benign conditions such as prostatitis and benign prostatic hyperplasia (BPH), many patients are subjected to unnecessary prostate biopsies that are expensive, uncomfortable, and carry non-trivial morbidities (Smith et al., 2002). In addition to this, a biopsy is often used to infer prognostic information of a patient – with the Gleason score histological grading considered to be the gold standard in prostate cancer. This scoring system has demonstrated close to a 5-fold difference in recurrence-free survival (RFS) between patients with 3+4 and 4+3 Gleason scores (Chan et al., 2000; Lau et al., 2001; Rasiah et al., 2003). However, while the Gleason score has shown to be an effective tool for determining PCa prognosis, it is a subjective test that requires highly invasive biopsies that can lead to a number of complications including bacterial infection and persistent hematuria – occurring in 47% of cases (Rodriguez and Terris, 1998).
Consequently, there is a very clear and concrete need to identify a set of minimally invasive biomarkers for PCa diagnosis and prognosis to be used in clinical laboratories.

1.2 Review of Literature

1.2.1 Epigenetics

The classic definition of ‘epigenetics’ is the study of mitotically and/or meiotically heritable changes in gene function that do not involve alterations in DNA sequence (Russo et al., 1996). Epigenetic aberrations are often linked with diseases ranging from imprinting disorders, such as Prader Willi syndrome (Robertson, 2005), to complex diseases – as is the case with most human cancers. In order to further study the role of epigenetics in cancer development, we must first understand the nature of epigenetic controls acting across the genome. There are four main interrelated types of epigenetic regulation that act synergistically, including DNA modifications, histone modifications, nucleosome positioning, and RNA interference. Among these, DNA modifications, such as methylation, and histone modifications are the most intensively studied for their roles in regulating gene expression as well as their involvement in diseases.

1.2.1.1 The effects of DNA methylation and other modifications

DNA methylation occurs in genomes of both eukaryotes and prokaryotes, and involves a covalent addition of a methyl group (CH3) to the 5’ position of the cytosine pyrimidine ring or the 6th nitrogen of the adenine purine ring. This modification is catalyzed through an enzymatic reaction that uses the action of DNA methyltransferases and S-adenosyl-methionine (SAM) as a methyl group donor. While DNA methylation can occur on both cytosine and adenine bases in prokaryote systems, methylation in eukaryotes seems to occur only on cytosine bases, most commonly (but not exclusively) preceding a guanine in CpG dinucleotides (Jaenisch and Bird, 2003). An exception to this can be found in embryonic stem cells and neurons, where DNA methylation can occur in CpH contexts (with H representing nucleotides A, C or T) (Lister et al., 2009a). CpG methylation is typically found in adult somatic tissues, where the methylation status is associated with a repressed chromatin state and an inhibition of gene expression (Holliday, 2006).

In mammals, DNA methylation is established and maintained by DNA methyltransferase enzymes (DNMTs). These enzymes are typically classified as de novo methyltransferases in the case of DNMT3A and DNMT3B, or maintenance methyltransferases (DNMT1) (Jones, 2012).
DNMT1 is the most abundant DNA methyltransferase in mammalian cells and localizes to replication foci to copy pre-existing methylation patterns onto new DNA strands during replication. As a result, DNMT1 has been shown to be vital for normal embryonic development, imprinting, and X chromosome inactivation, with DNMT1 knockout mice exhibiting embryonic lethality as a result of passive demethylation over time (Li et al., 1992). In contrast, DNMT3A and DNMT3B act independent of replication and show affinity for both hemi- and fully unmethylated DNA. Together, these enzymes work to produce the precise de novo DNA methylation patterns observed during embryogenesis and early in development; knockout models of DNMT3a in mice have also exhibited lethality (Li et al., 1992).

The result of cytosine methylation is a base known as 5-methylcytosine (5mC). However, in addition to DNA methylation, there are a number of different cytosine modifications that have been recently discovered. It was previously thought that DNA methylation could only be reversed by inhibiting the maintenance enzyme during cell divisions, which is generally a passive process. However, active DNA demethylation has also been reported to maintain the balance of DNA methylation levels throughout the genome. One mechanism of active demethylation involves a family of ten-eleven translocation (TET) enzymes that oxidize 5-mCs to 5-hydroxymethylcytosines (5-hmC). It has been suggested that these TET enzymes can then convert 5-hmC to 5-formylcytosine (5-fC) and subsequently 5-carboxylcytosine (5-caC), which can finally be converted to its unmodified form. Since most of the techniques used in epigenetic studies of cancer have not differentiated 5-mC from other forms of modified cytosines (Nestor et al., 2010), the term “DNA modification” will be used throughout this thesis to refer to such modifications.

In normal mammalian cells, between 70% and 80% of all CpGs are modified, which may function to silence repetitive elements (Ehrlich et al., 1982). However, there exist specific CpG-dense areas of DNA (CpG islands) where the CG dinucleotides are unmodified. Generally, CpG islands are defined as DNA sequences greater than 200-500 bases in length, with more than 50% GC content, and an observed-to-expected CpG ratio greater than 0.6. Using this definition, nearly 70% of CpG islands are associated with the first exons of human genes and more than half are within promoter regions (Bird et al., 1985). However, while CpG islands are dense with CpG dinucleotides, they are predominantly unmodified. This is in contrast to CpG island shores that occur outside of CpG islands, in relatively close proximity (~2kb), which exhibit greater
variability of DNA modification (Irizarry et al., 2009). In fact, CpG shores exhibit tissue-specific modifications, with their patterns shown to be sufficient for distinguishing between different tissues (Irizarry et al., 2009).

As mentioned above, the majority of CpG islands are unmodified during development and in differentiated tissues, with exceptions involving silenced alleles for imprinted genes and genes encompassed within regions of X-chromosome inactivation (Feil and Berger, 2007). This indicates the importance of DNA modifications for providing epigenetic control during normal mammalian development in both germ-line and tissue specific gene expression. In this regard, it has been shown that repetitive elements, such as ALU and LINE repeats scattered throughout the genome, exhibit high levels of modifications to prevent chromosomal instability through the silencing of non-coding DNA and transposable DNA elements (Watanabe and Maekawa, 2010).

There are two main processes by which DNA modifications silence gene expression. The first is a structural effect, whereby the modification of cytosine bases can inhibit necessary binding proteins, such as transcription factors, from associating with their respective DNA recognition sites (Sasai and Defossez, 2009). The second is the effect of modification marks on eliciting a repressive response by proteins that recognize the modification and recruit additional co-repressor molecules, such as chromatin remodelling proteins and histone deacetylases that alter surrounding chromatin into its repressed state to silence transcription (Sansom et al., 2007). It is clear that DNA modifications play an important role in proper biological differentiation and functioning, as aberrant patterns have been observed in a wide range of human diseases, including imprinting disorders such as Beckwith-Wiedemann syndrome (Robertson, 2005), and complex diseases – of which the most intensively studied is cancer (Feinberg, 2007).

1.2.1.2 Altered DNA modification in cancer

The presence of aberrant modification patterns in cancer was first discovered in 1983 (Feinberg and Vogelstein, 1983). Since then, a number of studies have demonstrated the role of epigenetics, particularly DNA modifications, in cancer initiation and progression. The cancer epigenome is characterized by global changes in DNA modifications, such as genome-wide loss of modified cytosines accompanied by a gain of such modified cytosines at CpG islands (Lopez et al., 2009). While the underlying mechanisms initiating these global changes are still unknown,
reports of their presence very early on in cancer development suggests that their roles in the
dysregulation of gene expression may be involved in carcinogenesis (Sharma et al., 2010).

A genome-wide loss of modified cytosines is mostly observed in repetitive regions of
both benign and malignant tumor types, with the degree of loss increasing with the progression
of a benign lesion to an invasive malignancy (Esteller, 2008a). There are a number of different
mechanisms to explain the link between a loss of modified cytosines and cancer cell
development, since a loss of modification in repetitive elements of the genome results in
genomic instability through chromosomal rearrangements. This has been confirmed in studies
that have shown an association between the loss of DNMTs and chromosomal instability in
human cancer cells (Karpf and Matsui, 2005). Since transposons are typically silenced through
DNA modifications, a loss of modification in cancer allows these elements to cause
translocations, insertions, and deletions that can disrupt the genome (Suter et al., 2004). In
addition to this, a decrease in cytosine modification can also result in the activation of normally
silenced proto-oncogenes, such as C- myelocytic leukaemia (c-MYC) – a growth promoting
factor (Cui et al., 2002).

However, the most well characterized epigenetic change in carcinogenesis is the de novo
modification of CpG islands around promoter regions of genes, a change that correlates with
transcriptional repression. This occurrence of increased cytosine modification in CpG islands, in
tumors, has been found to occur as frequently as genetic alterations in tumor suppressor genes
(Jones and Baylin, 2002). The silencing of tumor-suppressor genes through this epigenetic
change constitutes a major event in the development of the majority of cancer types (Esteller,
2008a), as this inactivation has the ability to affect molecular pathways involved with all aspects
of cell processes. This phenomenon has been observed in DNA repair (hMLH1, MGMT), cell-
cycle regulation (CDKN2A), and other roles that contribute to the hallmarks of cancer, including
apoptosis, transformation, signal transduction, angiogenesis, and metastasis (Widschwendter and
Jones, 2002). However, while some tumor-suppressor genes exhibit a gain of cytosine
modification across several cancer types, others have been found to be tumor-specific –their
profiles representing possible markers to differentiate between cancers (Esteller et al., 2001).

Together, the finding of aberrant genome-wide losses of modified cytosines, as well as
increased DNA modification of tumor-suppressor genes in cancer have resulted in further
investigations to study the epigenetic features of cancer. This avenue of research to elucidate the link between abnormal DNA modifications and different cancer stages represents a promising clinical application for the development of novel biomarkers for diagnosis, prognosis, and drug response (Esteller, 2008a).

1.2.2 Cancer biomarkers

The National Cancer Institute defines a biomarker as a biological molecule found in blood, other bodily fluids, or tissues that can signify the presence of a condition or disease (Biomarkers Definitions Working Group., 2001). As a result, biomarkers are often classified by the function they perform, including: diagnostic roles in classifying individuals as diseased or healthy, prognostic roles in providing information regarding the disease course or outcome, or predictive roles in assessing response to treatment (Kumar et al., 2006). While conventional methods of these assessments require histological examination of cell morphology, molecular markers have gained recent popularity in clinical settings with the development of high-throughput technologies. In the context of human neoplasia, novel cancer biomarkers can include DNA, RNA, proteins, or processes such as apoptosis, angiogenesis, or proliferation that can be measured quantitatively or qualitatively (Hayes et al., 1996).

Of these, DNA modification as a biomarker has several advantages over other molecular biomarkers. As described earlier, a number of DNA epigenetic changes have been shown to occur early on in carcinogenesis and therefore serve as potential biomarkers for early cancer development. Further, higher density of modified nucleotides in the promoters of tumor suppressor genes have been found in a variety of human cancers, with reports indicating distinct modification profiles associated with different malignancies – allowing the potential identification of the specific tissue of origin from biopsies or bodily fluids.

In practice, DNA is also easier to handle as it is more stable than both proteins and RNA, under conditions where clinical samples are collected and stored (Esteller, 2004). Compared with proteins, DNA represents a biomarker that is easily amplifiable and detectable by PCR-based techniques, meaning that very small amounts of DNA are required for these assays (Sidransky, 1997). Moreover, PCR amplification-based tests also allow the detection of as few as one cancer cell (or genome copy) in a background of thousands of wild-type cells, allowing cancer detection even prior to imaging techniques or traditional pathology analysis. As a result, modified DNA
sequences have the potential for use in diagnostic and prognostic scenarios, forming a basis for a robust and informative test for the management of cancer (Cairns, 2007).

1.2.2.1 Characteristics of an ideal biomarker

An ideal biomarker should provide information on patient management, and will only be of clinical value if it is accurate, reproducible, and easy to measure and interpret by clinicians. These accuracy measures are assessed using sensitivity, specificity, and predictive value. Sensitivity, also known as the true positive rate (TPR), represents the probability of a test returning positive, indicating disease presence in the patient (Nendaz and Perrier, 2004). In contrast, specificity is the true negative rate and represents the probability of a negative test result, indicating disease absence in the patient. While a highly sensitive test will result in a positive result in nearly all patients with the disease (true positives, TP), it may also include a number of positives in patients without the disease (false positives, FP). As a result, a biomarker used in screening should exhibit a high level of both sensitivity and specificity, to ensure that healthy patients correctly return with negative test results (true negatives, TN), reducing the number of false positives and negatives. In clinical practice, it is also important to measure how a test performs in predicting the likelihood of the disease (Nendaz and Perrier, 2004). Positive and negative predictive values (PPV and NPV) calculate the likelihood of having disease given a particular test result. While sensitivity can be used to compare the usefulness of two diagnostic tests, positive predictive values help clinicians decide how to manage a patient after a diagnostic test comes back positive.

Discoveries of novel biomarkers must be validated in independent studies to prove their ability to improve patient management in a consistent manner. During validation, a number of performance characteristics are typically assessed, including sensitivity, specificity, and reproducibility. As such, the course of biomarker development is complex and challenging, with many current biomarkers failing to satisfy the required characteristics. Moreover, the demonstration of clinical value requires biomarkers to be tested in generations of prospective data. In complex and heterogeneous diseases such as cancer, it is more likely that markers will be used in combination, as opposed to the use of a single biomarker that can detect all cancer subtypes and stages with optimum sensitivity and specificity. The use of multiple biomarker panels have been assessed in a number of diseases including cardiovascular risk assessment (de
Lemos and Lloyd-Jones, 2008) as well as cancers (Yurkovetsky et al., 2006), and has exhibited improved results.

1.2.3 Circulating DNA (cirDNA)

The analysis of blood-based biomarkers is an appealing option for cancer management, allowing non-invasive assessment, screening, disease classification and monitoring. CirDNA represents such a non-invasive biomarker, as it can be easily isolated from human plasma, serum, and other bodily fluids (Utting et al., 2002). The presence of cell-free circulating nucleic acids in human blood plasma was first reported by Mandel and Metais in 1948 (Mandel and Metais, 1948). In 1977, Leon et al. reported increased levels of cirDNA in the circulation of cancer patients when compared to healthy controls (Leon et al., 1977). However, it would be another decade before it was shown that cirDNA exhibited the same molecular alterations as DNA derived from related tumoral tissue (Stroun et al., 1989). In the following years, tumor-specific aberrations including mutations in oncogenes and tumor suppressor genes (Wang et al., 2004), microsatellite instability (MSI) (Shaw et al., 2000), and epigenetic alterations (Fujiwara et al., 2005) have been identified. These findings provide evidence that cirDNA is released into circulation by tumors and have bolstered the potential of cirDNA as a non-invasive tool for cancer management (Bremnes et al., 2005; Pathak et al., 2006).

1.2.3.1 Origins and characteristics of cirDNA

CirDNA in plasma is a double stranded molecule of low molecular weight and can be found in the form of a nucleoprotein complex (Stroun et al., 1987). However, electrophoresis of cirDNA on low-percentage agarose gels reveals its highly fragmented nature, with a size distribution ranging mainly from 70-200 bp, but also containing larger fragments up to 21 kb (Jahr et al., 2001). These size differences have been found to vary between samples (Jahr et al., 2001; Stroun et al., 1987).

Since the discovery of cirDNA, the question regarding its origin remains poorly understood. A number of different mechanisms have since been proposed to explain the shedding of cirDNA, including apoptosis, necrosis, direct release, and release from macrophages after the absorption of necrotic cells (Figure 1) (Diehl et al., 2005, 2008). In the apoptotic pathway, cells release degraded DNA in a process of cellular shrinkage and blebbing for further clearance by phagocytes and other cells (Cline and Radic, 2004). Apoptotic cell death is often characterized
by a typical pattern of small nucleosomal units of 180-200bp, requiring specific endonucleases to perform the inter-nucleosome cleavage of genomic DNA. This is in contrast to necrosis, which appears more as the accidental release and non-specific degradation of DNA, producing much larger fragments (Stroun et al., 2001). Another possible mechanism includes the active release of DNA into the circulation; this occurs via an intracellular fusion between the plasma membranes and multi-vesicular bodies to form bioactive vesicles known as exosomes (Schorey and Bhatnagar, 2008). While the percentage of contribution from each mechanism to the total amount of cirDNA in plasma is still unknown, studies suggest that the contribution of cirDNA by exosomes is relatively insignificant compared to that of apoptosis and necrosis (Choi et al., 2004).
Figure 1. Cell-free nucleic acids in the blood.

Mutations, DNA modifications, alterations of DNA integrity and microsatellite instability can be detected in cell-free cirDNA in blood. Tumor-related cirDNA, which circulates in the blood of cancer patients, is released by tumor cells in different forms and at different levels. DNA can be shed as both single-stranded and double-stranded DNA. The release of DNA from tumor cells can be through various cell physiological events such as apoptosis, necrosis and secretion. The physiology and rate of release is still not well understood; tumor burden and tumor cell proliferation rate may have a substantial role in these events. Individual tumor types can release more than one form of cirDNA.

Permission to reprint figure and legend can be found in Appendix (Schwarzenbach et al., 2011a)
There is no consensus concerning the mechanism of cirDNA release, with certain studies reporting the presence of smaller fragment sizes indicating an apoptotic origin (Wu et al., 2002b), and others finding larger cirDNA fragments of primarily necrotic origin (Diehl et al., 2005). The contributions of both apoptosis and necrosis may play equally important roles, with recent studies demonstrating evidence of both (Jahr et al., 2001). However, regardless of its mechanism of release, consistently elevated cirDNA levels across various cancer types, accompanied with tumor-associated molecular alterations, suggest their potential as biomarkers for the disease.

1.2.3.2 Other bodily fluids

Although plasma and serum cirDNA is most frequently used for clinical applications, the presence of cirDNA can also be detected and quantified in other bodily fluids such as urine, saliva, synovial fluid, and cerebrospinal fluid (Wagner, 2012). In particular, urinary analysis of cirDNA has attracted much attention in the examination of cirDNA originating from circulation, as well as cirDNA produced from the urinary tract and renal tissues. While very low amounts of cirDNA have been shown to pass through the kidneys (Botezatu et al., 2000), other studies have demonstrated that the detected cirDNA contain sufficient levels of sequences originating from non-renal and non-urinary tract tissues for potential diagnostic applications in colorectal cancer (Su et al., 2004). Furthermore, the detection of fetal Y-chromosomal cirDNA in maternal urine, as well as the presence of donor-derived cirDNA in the urine of patients that receive sex-mismatched hematopoietic stem cell transplants (Hung et al., 2009) demonstrate other possible uses of urinary cirDNA in providing a non-invasive diagnostic alternative.

1.2.3.3 CirDNA as a biomarker for cancer

Currently, the assessment of cirDNA has garnered the most interest in clinical practices of oncology and prenatal diagnostics. However, studies continue to reveal new areas for which cirDNA monitoring can have a positive impact. These biomedical interests include transplantation medicine, traumatology, cardiovascular care, and the monitoring of autoimmune and microbial diseases (Tsang and Lo, 2007; Wagner, 2012). In the majority of the aforementioned categories, the amount and concentration changes of cirDNA levels have been observed to be a hallmark characteristic of disease severity (Butt and Swaminathan, 2008;
Wagner, 2012), however DNA mutations and locus specific epigenetic changes of cirDNA have been found to be more suited for diagnostic purposes (Schwarzenbach et al., 2011a).

1.2.3.3.1 Quantitative changes in cancer

CirDNA has been found to be present in healthy subjects in concentrations ranging from 0 to 100 ng/mL of blood, with a mean level of 30 ng/mL (Anker and Stroun, 2000). In contrast to this, cirDNA occurs at much higher concentrations in cancer patients, exhibiting levels between 0 to >1000 ng/mL of blood, with an average of 180 ng/mL (Shapiro et al., 1983). Concentrations of cirDNA are influenced by a number of cancer dependent factors, including tumor size, stage, location, and other risk- and prognosis-related parameters; with results showing significantly higher levels of cirDNA detected in plasma of cancer patients with metastases (Jahr et al., 2001; van der Vaart and Pretorius, 2008). A study by Diehl et al. estimates that for a patient with a primary tumor weighing 100ng (corresponding to 3 x 10³ tumor cells), up to 3.3% of tumor DNA can enter the blood every day (Diehl et al., 2005).

Elevated levels of cirDNA in circulation have been detected in cancer patients when compared with healthy controls as early as the late 1970s using immunoassays (Leon et al., 1977). Since that time, other studies have detected increased levels of cirDNA in serum and plasma across different cancer types, including those of the breast (Sunami et al., 2008), ovaries, uterus, cervix (Leon et al., 1977), colon (Shapiro et al., 1983), pancreas (Shapiro et al., 1983), lung (Fleischhacker and Schmidt, 2007), and prostate (Allen et al., 2004). A number of these studies have suggested the simple quantification of cirDNA as a sufficient biomarker to confirm the presence of cancer or disease-free status, and/or relapse after curative surgery (Kim et al., 2014; Sozzi et al., 2003). In cases of post-treatment monitoring of cancer progression, decreased post-treatment cirDNA levels reflect treatment success and predict remission, while a persistent elevation of cirDNA levels indicate the presence of residual cancer cells, a failure to respond to treatment or a systemic spread of the disease (Jung et al., 2010; Schwarzenbach et al., 2011a).

However, there are contrasting studies that have raised considerations that should be taken in interpreting these results. With a certain level of cirDNA background observable in healthy individuals, it is important to achieve a better understanding of the mechanisms of cirDNA release and clearance to establish a baseline that allows consistent discrimination between healthy and diseased individuals (Lo et al., 1999). A number of studies have also found
elevated levels of cirDNA in patients with other physiological conditions such as physical trauma, inflammatory disorders, myocardial infarctions, and even patients with benign lesions (Chang et al., 2003; Jung et al., 2010; Lo et al., 2000; Zhong et al., 2007). As a result, the quantification of cirDNA concentrations alone as a biomarker does not seem to be sufficient for clinical use due to the overlapping DNA concentrations found in healthy individuals with those in patients with benign and malignant disease, and the failure to evaluate its cancer specificity.

1.2.3.3.2 Qualitative DNA changes in cancer

A milestone in the history of cirDNA research occurred with the discovery of tumor-specific alterations in cirDNA (Mayall et al., 1999). These qualitative changes detected in cirDNA corresponding with those from tumor tissues, can include microsatellite instability (MSI), loss of heterozygosity (LOH), DNA integrity, genetic mutations, and epigenetic alterations. These findings have driven forward research in this field as they provide a basis for the development of new approaches by which cancer-specific cirDNA changes may be used as non-invasive biomarkers.

In cancer, tumor-specific genetic alterations such as microsatellite instability (MSI), and loss of heterozygosity (LOH) have been observed in cirDNA. Microsatellites represent highly polymorphic repetitive sequences of DNA in which short motifs (usually one – six base pairs in length) are repeated 5-100 times. Microsatellites can exhibit abnormal lengths, with expansions of these repeats found to contribute to the inheritance of nearly 30 developmental and neurological disorders, typically involving DNA processes such as replication, repair, recombination, and transcription. With MSI associated with several cancer subtypes, and detectable in cirDNA; the study of MSI may aid in the practice of cancer diagnosis and progression. However, studies have shown that such testing depends on a small number of known microsatellite loci or mismatch repair genes, which may present as a challenge to use this application (Jin et al.; Lu et al., 2013).

Loss of heterozygosity (LOH) on the other hand occurs when, due to mutations, normal function of an allele is lost. In cancer, LOH is often observed in tumor suppressor genes and has been suggested to play a role in contributing to neoplastic transformation. As such, LOH analysis of alleles at specific chromosomes of cirDNA holds value for the early diagnostic and prognostic evaluation of primary tumors. A number of studies have demonstrated the potential of LOH as a
marker for tumor recurrence, even correlating with tumor status (Butt and Swaminathan, 2008; Jin et al.; Perkins et al., 2012). However, it is worth noting that LOH at different loci are found in low molecular weight fraction, and as cirDNA is present in both high and low molecular weight fractions – the fractionation of cirDNA may be important for achieving reproducible results (Gormally et al., 2007; Holdenrieder et al., 2005). Further, inflammatory processes can also result in an increase of apoptosis and an accumulation of small fragmented DNA in blood, which can mask the detection of LOH (Schulte-Hermann et al., 1995).

Since both MSI and LOH have been linked with roles in carcinogenesis, a number of studies have attempted to uncover these alterations in cirDNA. In a study conducted by Beau-Faller et al., a panel of 12 microsatellite markers were tested – with alterations detected in 88% of the plasma samples from lung cancer patients. This result was accompanied with the finding that all control samples were negative for such change (Beau-Faller et al., 2003). However, despite the concordance between tumor-related microsatellite alterations detected in plasma cirDNA and in DNA from corresponding primary tumors, discrepancies have also been found (Fleischhacker and Schmidt, 2007). These contradictory data have been partly attributed to technical problems and the masking of tumor-specific cirDNA in blood by DNA released from normal cells. Since the proportion of tumor DNA in cirDNA can be quite low, such microsatellite modifications may be difficult to detect in a high background of wild-type DNA (Coulet et al., 2000; Hibi et al., 1998; Kopreski et al., 1997).

The assessment of cirDNA integrity is another assay that has garnered interest in recent years. The measurement of cirDNA integrity involves the assessment of the ratio of longer to shorter DNA fragments (Umetani et al., 2006a), particularly of non-coding DNA such as the repeat sequences of ALU and long interspersed element 1 (LINE1 or L1). While ALU and LINE1 sequences are often referred to as “junk DNA”, there is mounting evidence that suggests their importance in vital physiological events, including DNA transcription, repair, and transposon-based activity (Bennett et al., 2008; Wolff et al., 2010). Furthermore, LINE1 is found to be quite abundant – comprising approximately 18% of the human genome, and can be detected in cirDNA in a variety of sizes. A number of studies focusing on DNA integrity have associated a greater cirDNA integrity with cancer (Salani et al., 2007; Umetani et al., 2006b). This is a result of DNA release from tumor cells undergoing non-apoptotic or necrotic cell death, with fragment sizes ranging between 200 and 400 bp in size. Since ALU and LINE1 sequences
are interspersed throughout the genome, sensitivity of the test may be enhanced at the cost of the specificity for any individual cancer type. Using PCR, studies have shown the sensitivity of DNA integrity assays in monitoring the progression of early stage breast cancer, as well as the presence of micrometastasis (Umetani et al., 2006a). Moreover, though further validation of these assays will be required to assess their utility in specific cancers, results from studies conducted in prostate, ovarian, testicular, and nasopharyngeal cancer have shown signs of early promise (Chan et al., 2008; Ellinger et al., 2009; Salani et al., 2007; Umetani et al., 2006b).

The characteristics of different tumors, including their location, morphology, metastatic potential, and response to various therapeutic interventions can be deduced from their tumor genomic signature. Differences in the histology of the same tumor type in the same organ (e.g. breast adenocarcinoma) or even variations in tumor types within the same organ (e.g. stromal gastric tumors vs gastric carcinomas) can be indicated by their genetic patterns (Heyn et al., 2013; Zauber et al., 2013). These variations are often a result of new gene combinations, such as chromosomal crossovers, and other genetic changes. In the context of cancer, there are a number of hotspots for chromosome instability that co-localize with cancer-associated genomic rearrangements (Palumbo et al., 2013).

Recent studies of tumor genome sequencing have confirmed the presence of multiple subpopulations (clones) of cells within one tumor, each subpopulation exhibiting different genetic profiles. The subsequent identification of genetic variations, including copy number variations (CNV) and somatic mutations have become a well-accepted approach to uncover critical genes important for the development and therapeutic resistance of cancer (Duell et al., 2010). In particular, genetic alterations of genes associated with DNA repair (Long et al., 2013), angiogenesis (Dorjgochoo et al., 2013), immune system, and inflammation (Dorjgochoo et al., 2013) have been found to affect the progression and survival of cancer patients. An example of this includes genetic variants in growth factor signaling, such as EGFR, ERBB2, and FGF1, different variations of which are associated with varying survival in breast cancer (Slattery et al., 2013). With a number of these genetic variations detected in solid tumors, studies are currently being performed to detect their presence in cirDNA.
1.2.3.3 Tumor-specific epigenetic alterations in cirDNA

Epigenetic modifications have been shown to play key roles in a number of cellular processes, including imprinting, chromatin remodelling, gene silencing, X chromosome inactivation, as well as carcinogenesis (Taback et al., 2004). Several studies have revealed concordant differential modification patterns in both primary tumor tissues and corresponding plasma or serum cirDNA in a number of genes across various malignancies, including breast (Silva et al., 1999), ovarian (Melnikov et al., 2009), cervical (Widschwendter et al., 2004), and lung cancer (Usadel et al., 2002). As a result, the detection of modified cirDNA represents one of the most promising approaches for risk assessment in patients across different cancer types.

Since aberrant modification patterns of CpG islands rarely occur in non-neoplastic or normally differentiated cells, cirDNA released from tumor cells can be detected even in the presence of wild-type DNA (Kim and Mirkin, 2013; Sunami et al., 2009; Wong et al., 2001). Goessl et al. was able to detect modified glutathione S-transferase pi 1 (GSTP1) cirDNA corresponding to 200 cancer cells in a background of $2.2 \times 10^7$ normal cells (Goessl et al., 2000). This is a clear advantage over assays that focus on microsatellite and mutation analyses, and represents great potential for clinical applications. Currently, specific gene modifications can be detected by sodium bisulfite treatment of DNA, which converts unmodified cytosines to uracil, while leaving the modified cytosines unchanged. This allows the modified DNA to be studied by DNA sequencing (Kristensen and Hansen, 2009), real-time PCR (Fleischhacker and Schmidt, 2007), or by methylation-specific PCR (MSP), with primers selective for modified and non-modified sequences (Herman, 2004).

Using these methods, a number of differentially modified genes in cirDNA have been identified across different cancer types. While epigenetic modifications may not be unique for a single cancer type, there are a number of tumor suppressor genes that are frequently downregulated and silenced through modifications in certain tumors (Ellinger et al., 2009; Taback et al., 2006). One example includes the gain of modified cytosines in the promoter regions of the GSTP1 and adenomatous polyposis coli (APC) genes, which have been shown to represent important epigenetic events in the carcinogenesis of prostate and colorectal cancer respectively (Ellinger et al., 2009; Taback et al., 2006). Moreover, one study looking at a panel of epigenetic markers found modification levels of 29% in APC, 56% in Ras association domain family member 1 (RASSF1A), and 35% in death-associated protein kinase 1 (DAPK) of 35 breast
cancer patients, while no modification was observed in any of the genes of 20 healthy controls and eight patients with benign breast disease (Dulaimi et al., 2004).

1.2.4 Prostate cancer

1.2.4.1 Epidemiology and risk factors

Prostate cancer (PCa) is the most common malignancy affecting men in the later stages of life (Siegel et al., 2014). Worldwide, prostate cancer is the second most frequently diagnosed cancer as well as the sixth leading cause of death in males with cancer. In Canada alone, an estimate of 24,000 new cases were diagnosed in 2014 resulting in nearly 4,100 deaths (Bell et al., 2014). Approximately one in six men will be diagnosed with prostate cancer during his lifetime, and about one in 36 will die as a result of the disease (Siegel et al., 2014).

Several risk factors for prostate cancer have been identified, and while no preventable risk factors have been determined, the most commonly identified ones include age, race, and family history. Advanced age is generally accepted as the leading risk factor for prostate cancer. Rarely detected in men younger than age 40, the median age of diagnosis for prostate cancer is 66 years of age, with the highest rate of incidence occurring in individuals aged 65-74 years old (Howlader et al.). It has been estimated that 75% of all prostate cancers are diagnosed in men over the age of 65. As a result, the prevalence of prostate cancer has been found to increase with age, with autopsy data suggesting an age-specific prevalence rate as high as 90% in men aged 70 to 90 years (Haas and Sakr, 1997). While prostate-cancer related deaths are more common in older men, data identifying indolent prostate cancer suggests that more men will die with prostate cancer, rather than from prostate cancer (Dunn and Kazer, 2011). However, with an increasingly aging population due to prolonged life expectancy, prostate cancer will continue to be a major health concern.

Incidence rates have also been found to vary considerably across different ethnic groups, with race being the second most common risk factor for developing prostate cancer. Studies have shown that African-American exhibit the highest rates of developing prostate cancer. In 2009, approximately 27,130 cases were diagnosed – accounting for more than 30% of all cancers diagnosed in this population (American Cancer Society, 2009). Further, when comparing between ethnic groups, incidence rates were almost 60% higher in African populations than in Caucasian ones, while Asian populations exhibited the lowest prevalence rates (Magi-Galluzzi et
While the role of race and ethnicity in the development of prostate cancer remains unclear, the varying incidence rates within the same ethnic groups suggests that genetic, environmental, and dietary factors may play a significant role in an individual’s likelihood of developing prostate cancer (Freedman et al., 2006). Prostate cancer risk has been shown to increase two-fold for men with a first degree relative with prostate cancer (Centers for Disease Control and Prevention, 2010). Furthermore, this risk is further increased if the number of relatives diagnosed increases and if the relatives were diagnosed at an age younger than 60 (Hogle, 2009); this is thought to be due to an inheritance of cancer predisposing genes (Grönberg, 2003).

**1.2.4.2 Classification and prognosis of PCa**

Classification in the context of PCa, and for cancer in general, is performed to characterize the features of the malignancy and to assess the extent of the cancer’s spread throughout the body. Proper staging require clinicians’ identification of particular traits including the size of the primary tumor, the presence and location of secondary tumors, the depth of penetration affecting adjacent tissues and lymph nodes, as well as the effect on afflicted organs (Yano et al., 2007). In contrast to this, grading systems such as the Gleason Score for PCa are part of the biopsy process, and are used to measure the extent of malignant transformation in cells obtained from a tissue sample (Yano et al., 2007). In order to allow for consistency between labs where biopsies are graded, and between clinicians where patients are staged, a standard system is required.

**1.2.4.2.1 Gleason grading system**

The Gleason scoring system was first developed by the pathologist Donald F. Gleason (Gleason, 1966), and remains one of the most commonly used prognostic predictors in PCa. In this system, tumor tissue from biopsies or prostatectomies can be assessed and assigned a Gleason score (GS) based upon its microscopic appearance. This score represents the sum of the two most common patterns of tumor growth (grades 1 to 5), and can range between 2 and 10. For example, a Gleason score of 4+3 = 7 means that pattern 4 was the most abundant area and pattern 3 was the second must abundant. Tumors associated with higher GS behave more aggressively and have been linked with worse prognosis (Albertsen et al., 1998).
1.2.4.2 Tumor node metastasis (TNM) system

The tumor node metastasis (TNM) evaluates a particular cancer based on the status of the primary tumor (T), the degree of spread among regional lymph nodes (N), and the presence and degree of distant metastasis (M). Each category is then further divided into subcategories: with the extent of tumor ranging from T1 to T4, with higher values indicating increased invasiveness and a greater involvement of the prostate and surrounding structures. The node category can be classified as no sample (X), no positive nodes (0), or an involvement of lymph nodes (1). Finally, degree of metastasis can be categorized as no spread (0), or (1) with further subcategories (1a-c) to describe the extent of PCa progression in the body (Edge and Byrd, 2010). The creation of this standardized staging system has been instrumental in providing a basis for survival prediction, treatment selection, and patient stratification in clinical trials. It has also served as an effective method to allow uniform communication among healthcare providers (Ludwig and Weinstein, 2005).

1.2.4.3 Early detection and diagnosis

Early detection of PCa has been shown to improve survival rates of patients by informing better clinical decisions and allowing for more appropriate treatments (Bill-Axelson et al., 2008). While early PCa can present without symptoms, and is relatively slow growing, it still presents as a lethal disease at more advanced stages. As a result of PCa’s long latency period and its high cure rate, this disease presents as an excellent candidate for screening strategies that attempt to identify the disease in its early stages (Strope and Andriole, 2010).

Current diagnosis and “informed” treatment decisions for PCa most commonly involve the use of digital rectal examinations (DRE), trans-rectal ultrasonography (TRUS), prostate specific antigen (PSA), and subsequent biopsies for histopathological staging (Bangma et al., 2007). However, each procedure has demonstrated its own shortcomings, and in practice, has resulted in an over-treatment of low-risk patients (Schröder et al., 2009), unnecessary biopsies, and radical prostatectomies (Moyer and U.S. Preventive Services Task Force, 2012; Schröder, 2011).

Despite the DRE’s poor sensitivity for PCa, especially in the early phases, it remains one of the most commonly utilized tools for the detection and screening of this malignancy (Basler and Thompson, 1998). The DRE has shown to exhibit an overall accuracy of only 59%, and
represents a tool that is highly subjective to inter-observer variability (Smith and Catalona, 1995). Furthermore, one of DRE’s primary limitations remain that it can only detect sufficiently large tumors located in regions of the gland palpable from the rectum (Mahon, 2005). On the other hand, the DRE is an inexpensive and normally well tolerated tool that has the ability to detect other conditions such as BPH, and may be particularly helpful for identifying PCa among men with normal PSA levels (<4ng/mL) – a situation that constitutes about 20% of occurring prostate malignancies (Basler and Thompson, 1998).

The transrectal ultrasonography (TRUS) is another tool that is often used to detect PCa by applying ultrasonic sound waves to visualize across the rectal wall into the prostate. While more invasive and expensive to perform compared to the simple DRE, it has been shown to outperform DRE in the detection of early stage PCa (Men et al., 2001) as it does not depend on the clinician’s subjective sense of touch for palpating the gland. While TRUS can be recommended in place of the biopsy, it has been shown to exhibit limited sensitivity and specificity for PCa detection on its own (Punnen and Nam, 2009).

1.2.4.3.1 Prostate specific antigen

Prostate specific antigen (PSA) is a serine protease belonging to the Kallikrein family and is produced by prostate epithelial cells. PSA was first reported in 1979, where it was purified from prostate extracts by Wang et al (Wang et al., 1979). Its discovery led to a number of key studies that showed PSA’s potential utility as a PCa marker: with results reporting its presence in serum (Papsidero et al., 1980), as well as its elevated levels in PCa patients (Chu et al., 1977; Kuriyama et al., 1980). Similar to the DRE test, serum PSA screening for PCa is relatively inexpensive. However, the biggest difference between the two tests lies in its sensitivity for PCa, with PSA exhibiting 93% sensitivity at a serum PSA threshold level of 4ng/mL compared with 59% for DRE (Hoffman, 2011). However, specificity is poor at 33%, as evidenced by the fact that PSA elevations have been associated with other factors such as age, trauma, and benign conditions such as BPH (Hoffman et al., 2002). While the positive predictive value of the PSA test depends on the threshold employed, the recommended cutoff of 4ng/mL results in a PPV of 37% and a NPV of 91% (Hoffman, 2011).

Clinically, the American Urological Association and the American Cancer Society recommends annual PSA screening for all men over 50 years old. In situations where the patient
presents with risk factors such as race or known relatives diagnosed with PCa, testing is recommended to start as early as age 40 (Carroll et al., 2001). However, regular screening for PCa with the PSA test has faced many controversies. As described earlier, since the serum PSA test has been shown to exhibit a low positive predictive value – with nearly 60% of biopsies being negative – this typically results in repeat PSA measurements and biopsies. As a result, the US preventative Services Task Force has argued that the potential benefits of regular PSA screening fails to outweigh the complications associated with unnecessary biopsies (Chou et al., 2011). In particular, with the recommended cutoff of 4ng/mL exceeded in 15% of healthy men as well as in elderly that show elevated levels due to age and benign prostate conditions, such as BPH and prostatitis, this results in over-diagnosis and subsequent over-treatment (Telesca et al., 2008).

Furthermore, since PCa is generally low-grade and slow-growing, it often requires more than a decade upon first diagnosis to draw conclusions about the benefits associated with this form of screening. In a large study conducted by the European Randomized Study of Screening for Prostate Cancer (ERSPC) that involved over 182,000 men from 50 to 74 years of age, a median follow-up of 11 years demonstrated that the relative reduction in mortality in the screening group was 0.10 deaths per 1000 person-years. That is, to prevent one death due to PCa at 11 years of follow-up, 1,055 patients would have to undergo screening and 37 cancers would need to be detected (Schröder et al., 2012). From an economic perspective, studies have also shown that more than five million dollars would need to be spent on screening to prevent one death from PCa (Shteynshlyuger and Andriole, 2011). Together, these findings have fuelled a continued conflict surrounding the benefits of regular PSA screening and have led many to question the value of PSA as an early detection tool (Carter and Isaacs, 2004; Croswell et al., 2011), further driving the search for novel improved markers for PCa (Madu and Lu, 2010).

1.2.4.4 Current and emerging biomarkers for PCa screening

Other types of biomarkers have emerged as potential candidates for PCa screening. With the development of novel proteomic and genomic technologies, the search for potential clinical biomarkers have made great and important strides. While proteomics have contributed significantly to the identification of tumor-specific biomarkers for PCa in serum (Goo and Goodlett, 2010), it continues to struggle with challenges of varying protein concentrations, an inability to detect low-level proteins in a background of high-abundance proteins, extreme inter-
individual variability, and interfering compounds via increased levels of salts (Wu and Yates, 2003). Furthermore, lack of reproducibility suggests that research efforts in genomics and epigenomics may be more productive towards uncovering novel biomarkers.

1.2.4.4.1 DNA and RNA Markers

Over the last two decades, a number of genetic alterations have been found to occur during the initiation and progression of PCa. As a result, the use of genomic analyses for studying cancer biomarkers has received much attention. In a recent study conducted by the International Practical Consortium, over 70 PCa susceptibility loci have been identified to account for up to 30% of familial risk (Eeles et al., 2013). Consequently, the identification of PCa associated genetic variants will not only further improve our understanding of the disease etiology, but may also have clinical implications for the early detection, diagnosis, and treatment of prostate cancer.

The following markers represent promising candidates that still require extensive validation; these include alpha-methylacyl-CoA racemase (AMACR) mRNA levels, which have been shown to exhibit a nine-fold elevation in 88% of the PCa tissue samples when compared with benign tissue (Jiang et al., 2004); prostate cancer antigen 3 (PCA3) mRNA overexpression – found in nearly 95% of PCa biopsies when compared with BPH patients and healthy controls; and the TMPRSS2:ERG gene fusion test. This gene fusion is a PCa-specific genomic rearrangement that joins a heavily androgen-regulated transmembrane protease, serine 2 (TMPRSS2) gene with the proto-oncogene ERG (erythroblastosis related gene) – a member of the E-twenty-six transcription factor family. Constituting 90% of all gene fusions found in PCa, TMPRSS2:ERG has also been found in approximately 50% of all prostate tumors (Tomlins et al., 2005). As a urinary biomarker, this gene fusion displays an impressive specificity of 93%, but a limited sensitivity of 37% (Hessels and Schalken, 2013). However, in a multi-marker panel with PCA3, this has been shown to increase diagnostic accuracy from an area under the curve (AUC) of 0.65 for PCA3 alone to 0.77 when TMPRSS2:ERG is added (Stephan et al., 2013).

1.2.4.4.2 Epigenetic markers

As described earlier, abnormal epigenetic changes, such as loss of modified cytosines in oncogenes and their increase in density in tumor-suppressor gene are common anomalies observed in PCa. These include GSTP1 (Lee et al., 1994), RASSF1A, and adenomatous polyposis
coli (APC) (Liu et al., 2011a; Maruyama et al., 2002). While a number of other genes have also been observed to be modified in prostate cancer, only a few genes have been identified as potential targets for early detection. This is due to the fact that many of the genes exhibit modification differences insufficient for clinically meaningful sensitivity, while others are also present in BPH, which decreases the marker’s specificity.

1.2.4.4.2.1 Epigenetic changes in prostate tumor

One of the most well studied genes for early PCa detection is GSTP1. Glutathione S-transferases (GSTs) are a family of enzymes with protective functions as part of the detoxification process. Utilizing reduced gluthathione, GSTs reduce reactive substrates such as free radicals and aldehydes to prevent harmful oxidation of nucleic acids or proteins within a cell (Lee et al., 1994). While one might expect an upregulation of GSTs during times of cellular stress to protect its genetic integrity, GSTP1 – an important member of this family of enzymes – has demonstrated a loss of expression through increased modification during the development of PCa. While not a gatekeeper to PCa, GSTP1 is recognized as a caretaker gene where its loss of expression is in many ways characteristic of the malignancy (Nelson et al., 2003). The increase of GSTP1 modification is present in 90% of PCa tissues but rarely appears in BPH, making this gene a solid DNA-based biomarker for PCa (Brooks et al., 1998). In addition to this, GSTP1 has been used in combination with other genes in multi-marker assays. Using quantitative methylation specific PCR (qMSP) to measure relative levels of GSTP1 to MYOD1 modification in 21 prostate biopsy samples from patients that had demonstrated elevated PSA levels, this panel showed 91% sensitivity and 100% specificity. This assay correctly predicted the biopsies’ histological diagnoses in 10/11 cases, and correctly confirmed negative PCa results from the 10 negative biopsies (Jerónimo et al., 2001). In another study where histological classification was compared with the GSTP1 assay using qMSP, it was found that histology alone detected PCa with 64% sensitivity and 100% specificity, whereas a combination of the two improved the sensitivity to 75% (Harden et al., 2003).

Other epigenetic markers have also been found to be useful for the early detection of PCa, including RASSF1A, RARβ2, APC, AR, LGALS3, and TIG1 (Tokumaru et al., 2004). These markers are most often used in combination with GSTP1, with a study using a four-gene panel of GSTP1, RARβ2, APC, and TIG1 demonstrating a 97% sensitivity (59/61 PCa cases correctly identified), representing a 33% improvement over the use of histological classification alone.
Further, it was shown that a combination of markers displayed a much better sensitivity than any of the markers alone, with $TIG1$, $APC$, $RAR\beta 2$, and $GSTP1$ detecting prostate carcinoma with a sensitivity of 70%, 79%, 89%, and 75%, respectively (Tokumaru et al., 2004).

1.2.4.4.2.2 Epigenetic changes in cirDNA of PCa patients

The findings of several studies have shown that PCa-specific epigenetic changes can be detected in cirDNA of serum and urine (Hoque et al., 2004), which provides opportunity as non-invasive methods for early detection. The first studies conducted to assess PCa epigenetic changes in plasma involved the analysis of $GSTP1$ promoters exhibiting an increase in cytosine modification. $GSTP1$ was found to be modified in 72% of PCa patients ($n=32$), but absent in patients with BPH (Goessl et al., 2000). In subsequent studies, this group also measured the utility of multiple serum-based markers, evaluating the modification levels of $GSTP1$, $TIG1$, $PTGS2$ and the Reprimo gene (Goessl et al., 2001). Importantly, they demonstrated concordant modification changes of cirDNA with those found in PCa tissue, representing the tumor-specificity of the markers. Further, each of the four genes displayed higher modification frequencies in PCa tumor tissue than in BPH, with healthy controls exhibiting 0% modification across all four markers (Ellinger et al., 2008a).

As a result of the prostate’s proximity to the urethra, cirDNA from normal and diseased prostate cells are shed into urine and detection of epigenetic changes could represent a non-invasive method for diagnosing PCa. However, while a pilot study was able to detect an increase of $GSTP1$ modification in urine of PCa patients, it only appeared in less than one third of the urine samples (27%) (Cairns et al., 2001). In another study, post-prostatic massage urine samples were studied for the presence of $GSTP1$ modification in men referred for diagnostic biopsies to compare its accuracies in screening for PCa (Rouprêt et al., 2007). An increase in $GSTP1$ modification in urine after massage displayed 75% sensitivity and 98% specificity for PCa when compared with the biopsies’ levels of 91% sensitivity and 88% specificity (Rouprêt et al., 2007). This study indicates that detection of increased $GSTP1$ modification in urine may present as a potential biomarker for PCa screening or as a potential companion marker for PSA to help increase specificity. However, larger prospective screening studies are necessary to fully understand its utility in clinical settings for PCa diagnosis. Meanwhile, as $GSTP1$ is detected in less than one third of PCa patient urine without invasive pre-assay prostate massage, the assessment of epigenetic changes in plasma may represent a more promising way forward.
Furthermore, since most of the earlier studies have focused on the assessment of \textit{GSTP1} and a handful of other selected genes, there is a very likely chance that more informative markers may have been missed.

\section*{1.3 Study objectives and rationale for hypotheses}

As reviewed above, epigenetic DNA modifications are one of the most common molecular alterations found in human cancer, and occur during the early stages of tumorigenesis (Esteller, 2008a; Herman and Baylin, 2003; Nelson et al., 2009). Since these aberrant patterns have been found in malignant cells as well as in plasma, serum, and other bodily fluids of cancer patients (Chan and Lo, 2002), the investigation of cirDNA may serve as a non-invasive approach to obtaining diagnostic and prognostic information to improve patient management.

The potential of DNA modification as a clinical biomarker has gained much attention in recent years, with a number of promising epigenetic biomarkers reported in plasma cirDNA – including \textit{SEPT9} and \textit{GSTP1} for colorectal cancer and prostate cancer, respectively. However, while a single gene approach can be productive, epigenome-wide studies have more potential to make a clinical impact through the identification of sets of individual markers. These epigenetic signatures are likely to exhibit higher sensitivity and specificity as compared to the traditional markers. Since previous research has only tested a limited number of selected genes, and therefore limit the scope of epigenetic studies; the objectives of this thesis are the following:

1. Develop a high throughput microarray-based technique for the interrogation of cirDNA modification profiles.

2. To use the microarray technique for identification of cirDNA modification differences in PCa patients and controls, and validate the identified markers using bisulfite pyrosequencing.

3. In addition to locus-by-locus comparisons, to employ the use of machine learning algorithms for the identification of PCa epigenetic signatures.
Chapter 2
Materials and Methods

2 MATERIALS AND METHODS

2.1 Samples

2.1.1 Prostate cancer pilot study

The primary sample set consisted of 19 prostate cancer patients, 20 BPH patients, and 20 control individuals (sample set 1) that were recruited in several hospitals in Novosibirsk, Russia. Prostate cancer and BPH patients were males matched by age (68.7±6.4 and 68.5±9.2 years old for prostate cancer and BPH patients, respectively). The additional control group consisted of younger males (46.3±6.5 years) with no history of any prostate disease. The secondary sample set (sample set 2) consisted of 20 prostate cancer patients (68.7±6.8 years old) and 18 individuals (69.1±7.2 years old) diagnosed with BPH recruited at the Vilnius University Hospital, Vilnius, Lithuania. All individuals were Caucasians. All the participants provided written informed consent and the research protocol has been approved by the research ethic boards from CAMH (Toronto, Canada), Vilnius University Hospital (Vilnius, Lithuania) and the Institute of Chemical Biology and Fundamental Medicine (Novosibirsk, Russia). All prostate cancer patients across both sample sets presented tumors confined to the prostate or adjacent tissue without spreading to lymph nodes or reported metastasis (T2–3N0Mx).

2.1.2 Prostate cancer tiling microarray study

In this study, 100 control individuals and 100 prostate cancer patients were recruited from the Princess Margaret Cancer Centre (PM), Toronto, Canada. Within the cancer patients, 50 had Gleason score = 6 and 50 had Gleason score > 6. Prostate cancer patients and control individuals were age-matched; all prostate cancer patients were Caucasians, as were 66% of control individuals. Clinical and demographic characteristics of both groups are summarized in Table 1. All of the participants provided written informed consent and the research protocol has been approved by the research ethic boards from Centre for Addiction and Mental Health (CAMH) (Toronto, Canada) and PM (Toronto, Canada). Pathological and clinical investigation was carried out at the PM. All prostate tumors were confined to the prostate or adjacent tissue without spreading to lymph nodes or reported metastasis (T2–3N0Mx).
2.2 DNA extraction

Blood samples from all individuals were collected and the plasma fraction separated by centrifugation and frozen at -80°C until DNA isolation. In our PCa pilot study, total cirDNA was isolated from 1 mL of plasma using the GF-1 Nucleic Acid Extraction Kit (Vivantis, Selangor, Malaysia) in the first sample set and QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) in the second sample set, according to manufacturers’ instructions. In our PCa tiling array study, CirDNA was isolated from 500 µL of plasma using QIAamp DNA Blood Mini Kit Circulating Nucleic Acids kit (Qiagen, Mississauga, ON, Canada) according to the manufacturers’ instructions. Isolated total plasma cirDNA was stored at -20°C until use.

2.3 DNA modification amplicon preparation

2.3.1 Principle of DNA modification detection

In our studies, the genomic fraction consisting of modified DNA is enriched and subsequently interrogated on microarray platforms (Figure 2). Briefly, cirDNA extracted from plasma are enzymatically blunt ended for adaptor ligation. Following this, the samples are digested using modification-sensitive enzymes that do not cut in the presence of methyl-cytosines. As a result, the only fragments that remain intact after enzymatic digestion are those that contain modified CpG sites. Subsequent PCR, with primers complementary to the universal adaptors, amplifies only the fragments bearing such modifications – enriching the modified fraction. Further, tumor cirDNA is expected to be degraded and shorter than the DNA fragments released from WBC cells ruptured during the storage and processing of blood samples. For this reason, we selected PCR conditions that favor amplification of shorter (<1.5 kbp) cirDNA fragments over larger (>1.5 kbp) ones originating from white blood cells.

Table 1. Prostate cancer tiling array study cohort description.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Median Age</th>
<th>Percent of Caucasians</th>
<th>Median PSA (ng/mL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>100</td>
<td>64 ± 6.3</td>
<td>66</td>
<td>6.54</td>
</tr>
<tr>
<td>Gleason Score = 6</td>
<td>50</td>
<td>61 ± 4.8</td>
<td>100</td>
<td>4.74</td>
</tr>
<tr>
<td>Gleason Score &gt; 6</td>
<td>50</td>
<td>64 ± 5.1</td>
<td>100</td>
<td>7.44</td>
</tr>
</tbody>
</table>
DNA samples isolated from plasma consist of fragmented cirDNA originated from apoptosis/necrosis in tumor cells (right) and larger size genomic DNA originated from circulating cells (i.e. lymphocytes) or other cellular sources (left). First, universal adaptors (magenta boxes) are ligated to the end of all DNA molecules. Next, samples are digested with DNA modification sensitive restriction enzymes. These enzymes will cut only at unmodified CpG positions (white lollipops) but not in modified CpG positions (black lollipops). Digested DNA is then amplified using primers that bind to the universal adaptors (green arrows). During the PCR reaction, DNA polymerase extends primers (dashed green lines) according to its processivity and the optimized reaction conditions. After the enrichment of the modified fraction of the cirDNA, PCR products will be obtained only from undigested short templates, mainly from tumor cirDNA, that have ligated adaptors at both sides containing modified CpG positions (enriched modified fraction) or lacking restriction sites (uninformative fraction). In longer templates (as expected from genomic DNA), the DNA polymerase cannot extend primers in the distance between 5’ and 3’ adaptors and therefore they will not be amplified.
2.3.2 DNA blunting

The cirDNA in our studies were subjected to enzyme treatment to create blunt double-stranded ends. Fifty nanograms of total cirDNA was incubated with 1X NEB Buffer 2, 100 µM dNTPs, 60 units of T4 DNA polymerase (NEB, Pickering, ON, Canada), and 2 ng of BSA in 112.2 µL final volume. This mixture was incubated at 12ºC for 20 minutes, and subsequently transferred to ice. The blunted DNA was further purified with a phenol-chloroform DNA isolation technique, using 120 µL of phenol:chloroform:isoamyl alcohol (25:24:1), followed by an ethanol precipitation. After precipitation, samples were dissolved in 25 µL of water.

2.3.3 Adaptor ligation

In both our PCa pilot and tiling array studies, blunt universal adaptors were prepared by annealing two oligonucleotides: oligo1: GCGGTGACCCGGGAGATCTGAATTC and oligo2: GAATTCAGATC. The adaptors were prepared in a reaction of 100 µL consisting of two oligo sequences at 40 µM in 1 M Tris (pH 7.9); the solution was heated to 95ºC for 5 minutes, incubated at 70ºC for 2 minutes, cooled at 25ºC for 2 minutes, and finally incubated for 16 hours at 4ºC. The annealed adaptors were stored at -20ºC prior to use. During the ligation process, blunted samples were incubated at 16ºC for 8 hours in a reaction volume of 50 µL containing 1X NEB ligation buffer, 0.1 pmol of adaptors, and 200 units of T4 DNA ligase.

2.3.4 DNA modification-sensitive restriction enzyme digestion

Next, the adaptor-ligated DNA was digested using a cocktail of DNA modification-sensitive restriction enzymes (HpaII, HinP1I and HpyCH4IV), which do not cut when the corresponding restriction sites contain modified nucleotides, 5-mC. Using half of the adaptor-ligated template, ten units of each enzyme were used with 3 µL of 10X NEB Buffer 1, in a total reaction volume of 56 µL. The samples were incubated for 8 hours at 37ºC then heated to 65ºC for 20 minutes to deactivate the enzymes.
2.3.5 Adaptor-mediated PCR

2.3.5.1 Prostate cancer pilot study

The enzymatically digested DNA template was amplified in 25 µL of final volume under the following conditions: 1× PCR buffer (Sigma-Aldrich, Oakville, ON, Canada), 3 mM MgCl₂ (Sigma-Aldrich), 275 µM aminoallyl dNTP mix (Ambion, Austin, TX, USA), 1.6 µM oligo 1 and 25 units Taq DNA polymerase (NEB). The PCR program started with an initial extension at 72°C for 5 min followed by 24 cycles of denaturation at 95°C for 1 min, annealing at 94°C for 40 s and elongation at 72°C for 2.5 min, with a final extension step at 72°C for 5 min. We have previously determined that such conditions enable the amplification of cirDNA isolated from plasma and mechanically fragmented DNA (Figure 3). Amplification was checked by electrophoresis on 1% agarose gels, and PCR products were purified using the MinElute kit (Qiagen, Mississauga, ON, Canada) and quantified using a Nanodrop 2000.

2.3.5.2 PCa tiling microarray study

In our PCa tiling array study, the digestion product was amplified using two rounds of adaptor mediated PCR to allow the production of enough target for hybridization onto high-density tiling arrays. In the first round, 25 µL of digested product was amplified in 100 µL final volume under the following conditions: 1× PCR buffer (Sigma-Aldrich, Oakville, ON, Canada), 3 mM MgCl₂ (Sigma-Aldrich), 2.3 µl dNTP/dUTP mix (10 mM dCTP/dATP/dGTP; 8 mM dTTP and 2 mM dUTP), 275 µm aminoallyl dNTP mix (Ambion, Austin, TX), 1.6 µM oligo 1 and 25 units Taq DNA polymerase (NEB). The PCR program for the first amplification round started with an initial extension at 72°C for 5 minutes followed by 10 cycles of denaturation at 95°C for 1 minutes, annealing at 94°C for 40 seconds and elongation at 72°C for 2.5 minutes, with a final extension step at 72°C for 5 minutes. Five µL of the product from the first amplification round was used as the template for the second amplification round in 100 µL final volume. The same PCR conditions were used as described above, but 15 amplification cycles were used, instead of 10. Amplification was verified by electrophoresis in agarose gels. PCR products were purified using the MinElute kit (Qiagen, Mississauga, ON, Canada) and quantified using a NanoDrop 2000. The products of 3 second PCR reactions per sample were pooled to obtain 9.5 µg target.
2.4 Microarray experiments and data analysis

2.4.1 Prostate cancer pilot study

2.4.1.1 Microarrays

For the PCa pilot study, we employed the use of HCGI12K microarrays (UHN, Toronto, ON, Canada), a two-channel microarray technology, which contain 12,192 GC-rich clones covering 5,347 unique sequences and 1,245 repetitive DNA elements—interrogating ~1% of the epigenome (Heislер et al., 2005). This technique is based on the competitive hybridization of differentially labeled test and reference samples. In this experiment, cirDNA from individual PCa or control samples were labeled with the fluorescent dye Cy3, and the DNA from our common reference pool was labeled with Cy5. The common reference pool acts as a filter for any differential modification signals that may originate from contaminating lymphocyte DNA, and also allows for normalization and comparisons between arrays.

The pooled test and reference samples are then hybridized to a single array containing thousands of probes. After removal of unbound or non-specifically bound fragments, the microarray is scanned in a laser scanner registering the fluorescence of the two channels separately: one for the test sample (Cy3) and one for the reference sample (Cy5). The two images are then merged and intensity ratios are calculated (test/reference, Cy3/Cy5) for each.
probe. The relative changes in each sample against the common reference pool can be assessed to allow comparisons of samples across arrays (Patterson et al., 2006).

In this study, DNA samples extracted from WBCs of 20 individuals (13 females and 7 males) unrelated to this project were pooled and used as our common reference. Participants were recruited at CAMH in Toronto, Canada and provided informed consent. DNA from these samples were extracted using phenol-chloroform isolation followed by ethanol precipitation. Further, isolated DNA was pooled and sheared by sonication to 200–500 bp fragments prior to undergoing the modification detection protocol.

Two technical replicates were conducted for each array in all patients and controls, with 1.5 μg of purified DNA labeled using Cy3 (GE Healthcare, Baie d’Urfe, QC, Canada) for cirDNA samples and Cy5 (GE Healthcare) for the reference pool; the DNA was then co-hybridized to the HCGI12K microarrays.

### 2.4.1.2 Microarray data analysis

#### 2.4.1.2.1 Pre-processing and statistical analysis

Hybridized microarrays were scanned using the Axon 4000B scanner and signals were processed using the GenePix Pro software (v6.1.0.4). The raw intensity data underwent extensive quality control, followed by pre-processing and statistical analysis as described previously (Ponzielli et al., 2008). Briefly, microarray data were pre-processed and normalized to correct for background noise using a modified version of the variance stabilizing normalization method – using the vsn package (Huber et al., 2002) (v3.20.0) – to yield a raw p-value based on a moderated t-statistic. Pre-processed data were then analyzed with spot-wise linear-model fitting followed by an empirical Bayes moderation of the standard error (Smyth, 2004). A false-discovery rate (FDR) adjustment for multiple-testing was used and q-values (FDR-adjusted p-values) were assigned (Storey and Tibshirani, 2003). Per-probe microarray signals relative to the reference pool (called M-values) were calculated by subtracting (in logspace) the normalized signal intensities corresponding to the target prepared using the cirDNA sample (Cy3 channel) and the target prepared from the blood reference pool (Cy5 channel). Coefficients were calculated as the ratio of modification between PCa and control individuals, with positive values signifying an increase in DNA modification and negative values indicating a decrease in DNA modification in PCa.
We used the limma package (v3.8.3) for the R statistical environment (2.13.2) (Smyth, 2004) for linear model fitting. To assess the effects of age differences between prostate cancer cases and controls, we fit a second linear model to identify loci whose modification signal is associated with age. Microarray raw data were deposited in NCBI’s Gene Expression Omnibus (GEO) database (accession number GSE36195). Hierarchical clustering was used to investigate global changes in DNA modifications in the loci exhibiting association with the disease (q<0.25) but not age (q>0.25). Ward’s clustering (Ward, 1963) on M-values, using Pearson’s correlation as the distance metric, was performed in the R statistical environment (v2.13.2), using the lattice (v0.19.26) and latticeExtra (v0.6.18) packages for visualization. The genomic distribution of the candidate probes was investigated using the BLAT tool in the Ensembl database (www.ensembl.org, Human reference assembly GRCh37.p3; Ensembl release 62; last accessed April 18, 2011). Available microarray probe sequences were aligned with the Homo sapiens LATESTGP database and the search sensitivity was set to ‘Near-exact matches’.

2.4.1.2.2 Multivariate classification analysis

Sample set 1 was used to train and test a classification model to differentiate between prostate cancer samples and control samples based on patterns of cirDNA modification. In each sample, the M-values from the two replicates were averaged per probe. A random forest of 100,000 trees was generated for classification in the R programming language (v2.13.1) using the randomForest package (v.4.6–2) with default parameterization (Liaw and Wiener, 2002). We used a standard doubly nested cross validation approach to ensure full separation of feature selection and classifier evaluation (Appendix, Figure S1). Briefly, an inner Leave One Out Cross Validation (LOOCV) loop was used for feature-size selection and an outer one for model evaluation. Thus, in the outer LOOCV, each sample was used for validation once using the optimal feature size as determined in the inner LOOCV. The feature sizes tested were 3, 10, 30, 50, 75 and 100. These features were chosen in the same manner as above to identify genes that are disease-dependent and age-independent (based on uncollapsed M-values) in the 37 training samples, using linear models for disease status and age. The feature size that led to the highest testing accuracy was then used in the outer LOOCV. A second round of linear models was performed using the 38 training samples from the outer LOOCV to determine the features to use in this classifier (i.e. feature selection was performed once for each sample in the outer LOOCV loop). A random forest fitted on the 38 training samples was then used to classify the held-out
validation sample. The entire process was repeated such that each sample was used as the validation sample in the outer LOOCV exactly once and thus a different random forest classifier was used to validate each sample, providing fully unbiased estimates of classifier accuracy.

2.4.1.3 Real-time PCR

To test the role of DNA copy number variants on chromosome 10, we used the adaptor-mediated amplification strategy described above, except no DNA digestion with modification sensitive restriction enzymes was performed: universal adaptors were ligated to 50 ng of cirDNA and amplified over 15 cycles using primers complementary to the adaptor sequences. This step enabled us to increase the cirDNA to the amount which is amenable to the real-time PCR-based estimation of copy number variation. PCR products were purified using the MinElute kit (Qiagen, Mississauga) and quantified using the Nanodrop 2000. 15ng of purified amplicon were subjected to real-time PCR (ABI StepOne plus device) using 1× ABI master mix containing Taq polymerase, dNTPs, SYBR green dye and ROX as passive dye (Life Technologies, Carlsbad, CA, USA) and 200 nM of specific primers (GAATGGAACGGCAACGAA and CGATGTCACTCCATTCTTCT). The PCR program started with a Taq polymerase activation step (10 min at 95°C) followed by 40 cycles at 95°C for 15 s, 60°C for 1 min and 95°C for 15 s. Amplification and melting curve analyses were performed using the Step One Software version 2.1 (Life Technologies).

2.4.1.4 Fine mapping of individual CpG locations

2.4.1.4.1 Principle of fine mapping

In order to perform fine mapping of the modification status of individual CpG sites, the DNA is first subjected to sodium bisulfite treatment. This process selectively changes the DNA sequence depending on the modification status of individual cytosine residues, resulting in single nucleotide changes after PCR. Unmodified cytosines are selectively deaminated to uracil after treatment with sodium bisulfite, while modified cytosines remain protected from this conversion. During PCR amplification, uracils are replaced with thymine, while the modified-cytosines remain cytosines, thereby allowing the detection of modifications at a single base-pair resolution. Seven loci were selected from the microarray experiment for further analysis and fine mapping of modified cytosines. These included discs large homolog 2 (DLG2), guanine nucleotide binding protein (G protein) gamma 7 (GNG7), heparanase 2 (HPSE2), KIAA1539 protein
(KIA1539), NudC domain containing 3 (NUDCD3), protocadherin beta 1 (PCDHB1), and ring finger protein 219 (RNF219).

2.4.1.4.2 Bisulfite treatment and whole genome amplification

The cirDNA, from the same samples subjected to microarray analyses, was bisulfite-treated using the Qiagen Epitect kit (Qiagen, Mississauga, ON, Canada). Further, to ensure the amplification of minimal amounts of template, bisulfite-treated cirDNA was amplified using the Epitect Whole Bisulfitome kit (Qiagen, Mississauga), according to the manufacturer’s instructions. This kit allows the whole genome amplification of bisulfite-treated DNA (whole bisulfitome amplification).

2.4.1.4.3 Nested PCR

We employed the use of a nested PCR technique for single locus amplification to increase the yield of our desired product, while minimizing the amount of non-specific contaminating amplification. This involves two successive runs of PCR, with two different sets of primers: one pair amplifying a larger region, and a second pair amplifying our desired target within the amplicons from the first reaction. Briefly, 10 ng of whole bisulfitome amplified bisulfite treated DNA was added to a reaction mixture of 1× Hotstart PCR buffer containing 1.5 mM MgCl₂ (Qiagen, Mississauga), 120 nM specific primers, 200 nM dNTPs and 0.65 units of Hotstar Taq polymerase (Qiagen, Mississauga). PCR started with a Taq polymerase activation step (15 min at 95°C) followed by either 10 (external fragment) or 40 (internal fragment) cycles of 95°C for 1 min, 55°C for 45s, 72°C for 1 min and a final extension step at 72°C for 10 min. The list of primers that were used for each gene can be found in (Table 2, Table 3).
<table>
<thead>
<tr>
<th>Locus</th>
<th>External Fragment Forward</th>
<th>External Fragment Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLG2</td>
<td>TGGGGGGTTTAAGTTTTTTTGAT</td>
<td>CCTCTTTAAACTCTCTTCTTCAAAAT</td>
</tr>
<tr>
<td>GNG7</td>
<td>TTGGTTGTTTTTGGAGTTGGGT</td>
<td>AAACCCCTACAAAAAAAAAAAATAACT</td>
</tr>
<tr>
<td>HPSE2</td>
<td>TAGTAGAGATAGGGTTTTTTATGT</td>
<td>TCTTTCACTAATTATCCTCCACA</td>
</tr>
<tr>
<td>KIAA1539</td>
<td>TTAAGGAGGAAGGAGGAGATA</td>
<td>AAACCCCTAAAATAACTAATTAAC</td>
</tr>
<tr>
<td>NUDCD3</td>
<td>TAGGGTTATTTTTTAGGTTAGGTA</td>
<td>TTTCTAAATTAAAACCCCTACAAAAC</td>
</tr>
<tr>
<td>PCDHB1</td>
<td>AAGTAGATGTGATAGTGAGGATTTTA</td>
<td>AACTCCTACCTCAAATAATCT</td>
</tr>
<tr>
<td>RNF219</td>
<td>GTTATATTTTTGGGAAGGTAAT</td>
<td>ACCCAAATAATCCATTAATCA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locus</th>
<th>Internal Primer Forward</th>
<th>Internal Primer Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLG2</td>
<td>GTTGTTGGGAATGTAGTTAAA</td>
<td>TCAAATTTCTTTCAACCTTCCCT</td>
</tr>
<tr>
<td>GNG7</td>
<td>GGGTTTTTTAGTTTGGAGTTTTAGT</td>
<td>TACCACCTCCTATATACCTTACCA</td>
</tr>
<tr>
<td>HPSE2</td>
<td>GTTGTTGGGATGTAGGTATGA</td>
<td>AACACTAAATTAAAACACTTAC</td>
</tr>
<tr>
<td>KIAA1539</td>
<td>AGGAAGGAGGAGATAAAGGTGAT</td>
<td>CCCCTCTAAACTTATCATAAC</td>
</tr>
<tr>
<td>NUDCD3</td>
<td>AGGGAGATAGGTGTTTAGTGGTT</td>
<td>ATAAAAATATAACCACCTCAAAC</td>
</tr>
<tr>
<td>PCDHB1</td>
<td>TTGTGTGTTTATATAATATGAAA</td>
<td>TAATCTCCACCTTAACCT</td>
</tr>
<tr>
<td>RNF219</td>
<td>GTGATTGTTGGATAGTTAAAAAT</td>
<td>ACTACCCCACTCTCCAAAA</td>
</tr>
</tbody>
</table>
2.4.1.5 **Bisulfite pyrosequencing**

The validation of the modification status of candidate genes was performed using bisulfite pyrosequencing. This technology was developed as a sensitive and quantitative method for assessing the modification status of CpGs. The technique involves the hybridization of single stranded biotinylated PCR product to sequencing primers in a reaction mixture containing DNA polymerase, ATP sulfurylase, luciferase and apyrase; and is based on the incorporation of new dNTPs by DNA polymerase which causes the release of pyrophosphate (PPi). This PPi byproduct is then converted into ATP by ATP sulfurylase which results in the production of light emissions by luciferase. The detection and quantification of this light is displayed as a peak on a Pyrogram, and corresponds to the number of nucleotides being incorporated. In the case of a modified cytosine, cytosines and thymines are dispensed sequentially and the ratio of light with each reaction can be interpreted as the percent modification at that locus. In our study, our DNA from the nested PCR was analyzed by pyrosequencing using a Qiagen PyroMark Q24 according to the manufacturer’s standard protocol. Modification values at single CpG positions were assessed using the PyroMark Q24 1.0.10 software (Qiagen, Mississauga). The primers used for pyrosequencing were the reverse primers used for amplification of the internal fragments in nested PCR (Table 3).

2.4.1.5.1 **Univariate classification analysis of ring finger protein 219**

Mean cirDNA modifications of all CpG positions included in the pyrosequencing PCR amplicon of ring finger protein 219 (RNF219) were calculated for each sample in both sample sets. To identify the cirDNA modification percentage which best differentiate controls from cases, various classifiers were built, each one corresponding to increasing cirDNA modification thresholds. The procedure was iterated using thresholds ranging from 1 to 100% in 0.1% increments. This analysis was performed comparing prostate cancer cases and unaffected controls in the training cohort (sample set 1): samples with cirDNA modification percentages above the threshold were classified as controls. The accuracy of each model was evaluated by using LOOCV on sample set 1. A receiver operating characteristic (ROC) curve was generated comparing the predicted and real classifications. The FDR and the true positive rate obtained using the optimal cirDNA modification threshold (i.e. the most accurate classifier) were noted. This threshold was then applied to independent sample set 2 to generate sensitivity and specificity values, as well as an ROC curve. This analysis was performed in the R statistical...
environment (v2.13.1). Area under the curves (AUCs) were calculated by summing the areas of the rectangles defined by each event (i.e. cancer versus not cancer).

2.4.1.6 Primary prostate cancer tissues analyses

2.4.1.6.1 Primary prostate cancer tissue DNA modification analysis

Previously published DNA modification data from primary prostate tumors and non-cancerous prostate tissue from prostate cancer patients were downloaded from GEO. The data of interest included 95 primary prostate tumors, 85 of which have associated disease-free survival information and 86 non-tumor prostate tissues (processed on Illumina's Infinium HumanMethylation27 BeadChip kits) (Kobayashi et al., 2011). ComBat (Johnson et al., 2006) normalized data were downloaded from GEO for each sample and was further normalized such that each intensity value was greater than zero (by adding the two smallest, positive intensities to each value) and then log_{10}-transformed. Fold-changes were calculated for each feature and their significance was assessed using two-sample, unpaired t-tests with Welch's adjustment for heteroscedasticity. Probes were mapped to Entrez Gene IDs within 1500 bp provided by an Illumina's probe description file. Cox proportional hazards regression was used to assess whether the DNA modification of each probe is associated with the disease-free survival. The Kaplan–Meier curves were generated using the lattice (v0.19–34) and latticeExtra (v0.6–18) libraries for the R statistical environment (v2.13.1).

2.4.1.6.2 Primary prostate tumor mRNA expression analysis

Publicly available steady-state mRNA data from primary prostate tumors were examined for evidence of disease involvement in the context of our cirDNA findings. Raw data (.CEL files) were downloaded for two large mRNA abundance microarray studies which provided patient survival information (Best et al., 2005; Taylor et al., 2010). In total 150 primary tumor samples were available for survival analysis, an additional study (Wang et al., 2010) was used to compare primary prostate tumors (n = 68) to prostate tissue from healthy controls (n = 45). Probes were re-annotated using updated Entrez Gene CDFs (Dai et al., 2005) (R packages huex10stv2hsentrezgcdf v14.0.0 and hgfocushsentrezgcdf v.14.0.0). Pre-processing used the robust multi-array algorithm (Bolstad et al., 2003). Using the Wang et al. data, gene-wise fold-changes and their statistical significance were calculated using two-sample, unpaired t-tests adjusted for heteroscedasticity. Survival analyses were performed by fitting the Cox proportional
hazards regression models for data-set-wise median-dichotomized per-gene data, followed by the Wald test using the survival (v2.36.10), lattice (v0.19–34) and latticeExtra (v0.6–18) libraries for the R statistical environment (v2.13.1).

2.4.1.6.3 Gene ranking using multiple independent data sets

To identify genes which show evidence of importance in prostate cancer across various data sets, we developed a novel scoring metric. To compare results from our cirDNA analysis to the public primary prostate cancer data sets, we mapped our probes to the nearest Entrez gene (average distance: 55.6 kbp). For each gene that was differentially modified in cirDNA in a disease-dependent ($q < 0.25$) and an age-independent ($q > 0.25$) manner, we assigned a score based on equal weighting of four binary questions: (i) is the gene differentially modified in primary prostate cancer relative to healthy prostate tissue ($q < 0.05$); (ii) is the gene's modification status associated with poor prognosis in primary prostate cancer ($p < 0.05$); (iii) is the gene differentially expressed in primary prostate cancer at the mRNA level ($q < 0.05$); (iv) is the gene's mRNA expression level associated with patient prognosis in primary prostate cancer ($p < 0.05$). These data were visualized in heatmaps using the lattice (v0.19.26) and latticeExtra (v0.6.18) packages for R (v2.13.2).

2.4.2 Prostate cancer tiling microarray study

2.4.2.1 Microarrays

Continuous improvements to microarray technology have allowed manufacturers to increase the probe density of slides from several thousand to several millions. This allows microarrays to cover not only specific genomic loci, but entire genomes. These high density, whole genome microarrays, are named tiling arrays for the way the probes “tile” across the whole genome and allow the interrogation of the genome in an unbiased manner. In contrast to earlier generation microarrays that are designed with probes in only promoter or coding regions, tiling arrays cover the entire genome, excluding repetitive elements.

In this study, we utilized Affymetrix GeneChip Human Tiling 2.0R microarrays, available as a whole-genome set of seven arrays – each containing over 6.5 million probes fully covering the non-repetitive regions of the genome. Affymetrix microarrays are a single channel system utilizing only one fluorophore. This design enables the hybridization of a single sample to each microarray, after which, signal intensities are compared across microarrays and between
different sample groups (Patterson et al., 2006). The probes are tiled at an average resolution of 35 bp, as measured from the central position of adjacent 25-mer oligos, leaving a gap of approximately 10 bp between probes. The enriched fraction for each sample was interrogated on Microarray B (Human Tiling 2.0 array set, Affymetrix, Santa Clara, CA), which contains full chromosomes 2, 9 and 19, representing approximately 15% of the non-repetitive human genome at high resolution. 9µg of purified PCR product was labeled and hybridized to the microarray, according to the manufacturer’s conditions. After 16 hours of hybridization, arrays were stained and washed according to standard manufacturer's protocols. The array was scanned using the GeneChip Scanner 3000 7G (Affymetrix) and signals were processed using the GCOS software producing a single .CEL file per array.

2.4.2.2 Microarray data analysis

2.4.2.2.1 Pre-processing and statistical analysis

Raw array data (CEL files) were loaded into the R statistical environment (v2.15.1) using the oligo package (v1.20.4) (Carvalho and Irizarry, 2010) from the BioConductor library (Gentleman et al., 2004). Data were investigated for distributional homogeneity and inter-array correlation: 17 samples were excluded from the analysis due to poor inter-array correlation using Pearson's correlation metric. Array data were pre-processed with the robust multi-array average (RMA) algorithm (Irizarry et al., 2003) as implemented in the oligo package (v1.20.4). Probe intensities were corrected for background noise and quantile normalized. Model-based analysis of tiling-arrays (MAT) (Johnson et al., 2006) was used to identify regions of differential cirDNA modification by combining adjacent probes showing significant differences between the groups. A sliding window of 500 bp was set, according to the average size of the fragments produced in the amplicon preparation step.

To identify probes with significant differential cirDNA modification between the PCa and control groups, we employed a general linear-model. For each probe, we modeled cirDNA modification levels across all samples as a function of group, microarray batch number and age. Coefficients were fit to terms representing each of these effects as previously described (Boutros et al., 2011). The standard errors of each coefficient were adjusted with an empirical Bayes moderation of the standard error (Smyth, 2004). To assess if the coefficients were significantly different from zero, model-based t-tests were employed; p-values for each term were then
adjusted for multiple testing using the False Discovery Rate (FDR) method (Storey and Tibshirani, 2003). Probes with FDR ≤ 0.01 were defined as having significant differential cirDNA modification.

2.4.2.2 Multivariate classification analysis

The sample set was randomly divided into 60% training (n = 110) and 40% testing (n = 73); a process which was repeated 100 times to prevent sampling bias (Figure 4). We then took the initial training dataset at each iteration and split it into training and testing dataset (70% training and 30% testing; n = 75 and 35, respectively); this process was also repeated 100 times. A random forest (version 4.6-7) (Breiman, 2001) with 100,000 trees was trained for classification in the R statistical environment with default parameterization (Liaw and Wiener, 2002). In addition, the naive Bayes (R package \e1071\, version 1.6-1) algorithm (Domingos and Pazzani, 1997) was used with default parameterization to compare to the random forest results. Within the inner loop, a series of feature sizes were tested (30, 100, 300, 500, 1000). Features were chosen based on risk-group-dependent fold-changes, using linear models for disease status and age as described above. The most optimal feature size was determined based on median accuracy for training in the outer loop. If a tie occurred during the selection process, the smaller feature size was chosen as the optimal feature size. In addition, within the inner loop, features were ranked based on their median Gini importance score (decrease in Gini impurity) from random-forest algorithm (Díaz-Uriarte and Alvarez de Andrés, 2006) and the highest scoring probes were selected for use in the outer loop. Classifier performance was evaluated only in the outer loop.

Further, within each iteration, the most optimal features selected based on a resampling approach were modeled together with PSA (as a continuous variable) using random forest (randomForest v4.6-10; default parameters but with 100,000 trees) on the training set; and risk groups were predicted on the hold-back/testing set. Sensitivity, specificity and accuracy were calculated for each iteration.
Figure 4. Building of an “epigenetic signature” for prostate cancer screening.
Machine learning algorithm scheme. Total dataset was randomly divided into training (60%) and testing (40%) datasets, with 100-times random sampling. Training dataset was further divided (70% and 30%) with 100-times random sampling. Machine learning approaches were used for building and testing the performance of panels with different number of features (n=30, 100, 300, 500 and 1000). Sensitivity, specificity and accuracy were calculated for each feature size and the feature size with the highest median accuracy was selected to train and test on the larger training and testing cohort.
2.4.2.2.3 Gene ontology and network analysis

As a next step, we listed the genes associated with regions displaying differential cirDNA modifications between the PCa and control groups. To explore the potential biological function of these genes, they were uploaded to DAVID 6.7 for functional annotation clustering analysis, selecting all available annotation categories (Huang et al., 2009a, 2009b). For each functional group (i.e. gene cluster or ontology term), the algorithm provides an “Enrichment Score”, which is calculated as the geometric means of the metrics obtained by a modified version of Fisher’s Exact test (EASE score, (Hosack et al., 2003)). A higher score for a group indicates that this particular group is over-represented (enriched) in the candidate list compared to the full list of annotated genes. The enrichment analysis was performed with the default human genome as background and a Benjamini-Hochberg FDR <0.05 as the significance threshold (Benjamini and Hochberg, 1995).

Since the UCSC_TFBS category displayed the highest enrichment score in the functional annotation clustering analysis, each of the 44 transcription factor terms passing our significance threshold of Benjamini-Hochberg FDR <0.05 were queried using MotifMap (Xie et al., 2009) to obtain their corresponding binding motif chromosomal locations. Using the Human (hg18 multiz28way_placental) species track, each transcription factor ID was queried at the default conservation and motif scores, with a default “distance closest to gene” at 1000bp and an FDR cutoff of 0.05. The resulting binding motifs were sorted by FDR significance and the top five binding motif’s chromosomal locations and closest associated genes were recorded (Appendix, Table S1). The DNA sequences for each of these binding motifs were then obtained using the UCSC Genome browser assembly NCBI36/h18, and all CG dinucleotide presence were also noted for potential epigenetic regulation (Appendix, Table S4).

In addition, GO terms were further separated by categories: GOTERM_BP, GOTERM_CC, and GOTERM_MF, and sorted by enrichment score at a significance cutoff of p<0.05 to investigate the functional role of the candidate genes. Further, DAVID’s gene functional classification analysis was performed using the list of differentially modified genes using the default human population background.
Chapter 3
Results

3 RESULTS

3.1 Prostate cancer pilot study

3.1.1 Differentially modified cirDNA in prostate cancer

In our PCa pilot study, we first studied cirDNA modification profiles in 19 prostate cancer patients, 20 BPH patients and 20 control individuals (sample set 1). The modified fraction of cirDNA was enriched using DNA modification sensitive restriction enzymes (HpaII, HinP1I and HpyCH4IV) and adaptor-mediated PCR. The enriched modified DNA fractions were interrogated on microarrays containing 12,192 GC-rich clones covering 5,347 unique sequences and 1,245 repetitive elements (Heisler et al., 2005). We compared the plasma cirDNA modification profiles of age-matched prostate cancer and BPH patients in sample set 1 (68.7 ± 6.4 years for BPH and 68.5 ± 9.2 years for cancer patients), but did not find any differentially modified loci between the two groups. This may be attributable to our comparison of a malignant PCa and premalignant BPH patient populations. In order to generate a larger biological contrast for candidate identification, we chose to compare the cirDNA profiles of prostate cancer against younger controls without any evidence of prostate disease (age 46.3 ± 6.5 years). We performed two separate analyses to identify disease- and age-dependent DNA modification changes (Figure 5, Figure 6). At FDR $q < 0.05$, we identified 117 differentially modified regions between prostate cancer and asymptomatic controls. Of these, 39 loci were age-independent, including 26 that were mapped to genomic regions. Of these, 18 correspond to non-repetitive sequences.

To verify our microarray results, we performed fine mapping of modified cytosines using bisulfite modification coupled with pyrosequencing on the initial discovery cohort (sample set 1), as well as on an independent patient cohort (sample set 2) consisting of 20 prostate cancer patients (68.7±6.8 years) and 18 BPH patients (69.1±7.2 years). For bisulfite sequencing, we selected (i) four genes which showed statistically significant differences ($q < 0.05$): NudC domain containing 3 (NUDCD3), protocadherin beta 1 (PCDHB1), KIAA1539 protein (KIA1539), ring finger protein 219 (RNF219), and (ii) three genes which showed marginal statistical significance (0.05 < $q < 0.08$): heparanase 2 (HPSE2), discs large homolog 2 (DLG2),
guanine nucleotide binding protein (G protein) gamma 7 (GNG7). These genes were selected from a list of the most significant differences between cancer patients and control individuals in the 12K CpG Island microarray experiment and were previously reported as relevant for carcinogenesis and progression in prostate and other cancers. All selected genes were more closely associated with disease than with age.
Figure 5. Differentially modified regions in cirDNA in prostate cancer and control individuals.
Volcano plot of microarray data in prostate cancer and control samples using q-value as statistics. The x-axis represents DNA modification differences between groups, with coefficients expressed in the log$_2$ scale. Samples with increased microarray signals in prostate cancer and control individuals had positive and negative coefficients, respectively. The y-axis represents -log$_{10}$-transformed q-values. The number of counts represented by each point in the plot is shown in a color gradient, from light gray to black (representing 1 and 600 probes, respectively). The horizontal red line depicts the cutoff value for the q-value (q < 0.05). The vertical red line depicts the zero value, i.e. no differences in DNA modification.
Figure 6. Differentially modified regions in cirDNA in prostate cancer and control individuals.
Heatmap of the 181 loci with age-independent, disease-dependent DNA modification in our microarray data at $q < 0.25$. Samples are on the x-axis and PCa and Ctrl prefixes refer to prostate cancer and control samples, respectively. Differentially modified loci are placed on the y-axis. Biological groups are differentiated by colors on the upper bar: green for control samples and red for prostate cancer samples. Normalized array signals are represented on a gradient scale ranging from red (-6) to blue (6). Color gradient represents the differences in probe signal intensity, ranging from red (higher intensity in CTRL group), over white (no differences) to violet (higher intensity in PCa group).
3.1.2 Microarray verification of candidate genes by pyrosequencing

The candidate genes chosen for fine mapping verification were mapped within the gene body or immediate upstream/downstream regions of the genes (Table 4). All pyrosequencing assays contained at least one restriction site for the enzymes used in the enrichment of modified DNA and contained different number of CpGs (Table 4). Table 5 summarizes the pyrosequencing results for the seven genes and the results of a statistical analysis using linear mixed-effect (LME) models. It is important to note that bisulfite pyrosequencing, although precise, is limited to relatively short DNA fragments (in our study pyrosequencing fragments were 106 ± 25 bp), which makes it difficult to test cumulative DNA modification effects at the restriction sites of the three enzymes across extended DNA fragments captured by long probes on HCGI12k microarrays. Nevertheless, in our independent validation cohort (sample set 2), pyrosequencing yielded a 71% replication of microarray results; with five of the seven genes demonstrating concordance of results obtained by both microarray and pyrosequencing experiments, by looking at the ratio of DNA modification in prostate cancer over controls for both methods. Furthermore, two loci (KIAA1539 and RNF219) showed significant disease-dependent ($P < 0.05$) and age-independent ($P > 0.05$) differences (Table 5).

In RNF219, differential modification between 19 prostate cancer cases and 20 controls was detected at the restriction site as well as at the neighboring CpG positions (Figure 7). In sample set 1, the mean RNF219 modification was significantly lower in prostate cancer cases than in controls ($35 \pm 18\%$ and $55 \pm 17\%$, respectively), (Mann Whitney test; p=7.5 x 10^{-4}) (Figure 8). Consistently with the first sample set, mean DNA modification in the sample 2 was lower in cases ($35 \pm 32\%$) than in controls ($44 \pm 25\%$), but the difference was not statistically significant (Mann Whitney test; p=0.56). (Figure 8). There were three outliers in the control group in sample set 2, with mean DNA modification 3.4%, 2.6% and 2.1% (Z-score >1.6). It is interesting to note that since the controls in sample set 2 were BPH patients, some of them must have had a higher risk to develop prostate cancer (Alcaraz et al., 2009; Hammarsten and Högstedt, 2002). In the follow-up studies we learned that the first individual had high preoperative PSA level (10ng/ml) but no further prostate disease has been reported yet. The second individual had elevated preoperative PSA level (6.4 ng/ml) and three months after prostate surgery was diagnosed with B-cell chronic lymphocytic leukemia. The third individual had low preoperative PSA level (1.2 ng/ml), but developed skin basal cell carcinoma 6 months
after prostate surgery. Removal of the outliers led to a higher difference in the mean DNA modification between prostate cancer cases and BPH controls (35 ± 32 % and 52 ± 18 %, respectively, Mann Whitney test; p=0.12). The two combined sets detected significantly lower density of modified cytosines in cases (35 ± 25 %, n=40) than in controls either when the outliers were included (50 ± 21 %, n=38) (Mann Whitney test; p=6.33 x10^{-3}) or excluded (54 ± 17 %, n=35) (Mann Whitney test; p=5.3 x 10^{-4}) (Figure 8).
Table 4. Loci selected for fine mapping of modified CpG positions.

<table>
<thead>
<tr>
<th>Candidate probe ID</th>
<th>Probe length</th>
<th>Associated gene</th>
<th>Position</th>
<th>Distance to TSS</th>
<th>CpG island</th>
<th># CG positions</th>
<th>Restriction sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>UHNhscpg0006294</td>
<td>67 bp</td>
<td>DLG2</td>
<td>Intron 2</td>
<td>143170</td>
<td>No</td>
<td>5</td>
<td>HpaII</td>
</tr>
<tr>
<td>UHNhscpg0002962</td>
<td>350 bp</td>
<td>GNG7</td>
<td>Intron 2</td>
<td>25281</td>
<td>No</td>
<td>3</td>
<td>HinP1, HpyCH4IV</td>
</tr>
<tr>
<td>UHNhscpg0003529</td>
<td>98 bp</td>
<td>HPSE2</td>
<td>Intron 3</td>
<td>175449</td>
<td>No</td>
<td>1</td>
<td>HinP1</td>
</tr>
<tr>
<td>UHNhscpg0001923</td>
<td>646 bp</td>
<td>KIAA1539</td>
<td>3' Intergenic</td>
<td>12202</td>
<td>Yes</td>
<td>5</td>
<td>HpaII, HpyCH4IV</td>
</tr>
<tr>
<td>UHNhscpg0006245</td>
<td>784 bp</td>
<td>NUDCD3</td>
<td>3' Intergenic</td>
<td>108995</td>
<td>No</td>
<td>3</td>
<td>HinP1</td>
</tr>
<tr>
<td>UHNhscpg0008390</td>
<td>257 bp</td>
<td>PCDHB1</td>
<td>3' Intergenic</td>
<td>8166</td>
<td>No</td>
<td>1</td>
<td>HpaII</td>
</tr>
<tr>
<td>UHNhscpg0001950</td>
<td>505 bp</td>
<td>RNF219</td>
<td>Intron 1</td>
<td>138</td>
<td>Yes</td>
<td>17</td>
<td>HinP1, HpaII</td>
</tr>
</tbody>
</table>
Table 5. Summary of pyrosequencing results

<table>
<thead>
<tr>
<th>Loci</th>
<th>Sample set 1</th>
<th></th>
<th></th>
<th></th>
<th>Sample set 2</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Mod. PCa(a)</td>
<td>% Mod. Ctrl(a)</td>
<td>Disease pLME(b)</td>
<td>Age pLME(b)</td>
<td>% Mod. PCa(a)</td>
<td>% Mod. Ctrl(a)</td>
<td>Disease pLME(b)</td>
<td>Age pLME(b)</td>
</tr>
<tr>
<td>DLG2</td>
<td>48.7 ± 22.0</td>
<td>45.6 ± 23.6</td>
<td>\textbf{0.02}</td>
<td>\textbf{0.02}</td>
<td>42.9 ± 14.7</td>
<td>41.9 ± 18.3</td>
<td>0.85</td>
<td>0.07</td>
</tr>
<tr>
<td>GNG7</td>
<td>38.4 ± 3.0</td>
<td>38.2 ± 2.4</td>
<td>0.74</td>
<td>0.40</td>
<td>37.9 ± 11.1</td>
<td>36.8 ± 7.8</td>
<td>0.93</td>
<td>0.21</td>
</tr>
<tr>
<td>HPSE2</td>
<td>22.7 ± 13.7</td>
<td>16.1 ± 8.3</td>
<td>0.13</td>
<td>\textbf{0.02}</td>
<td>40.6 ± 30.0</td>
<td>29.6 ± 27.1</td>
<td>0.32</td>
<td>0.82</td>
</tr>
<tr>
<td>KIAA1539</td>
<td>38.8 ± 5.8</td>
<td>37.1 ± 10.9</td>
<td>(1 \times 10^{-3})</td>
<td>(8 \times 10^{-4})</td>
<td>30.8 ± 10.0</td>
<td>36.2 ± 6.4</td>
<td>(1 \times 10^{-4})</td>
<td>0.05</td>
</tr>
<tr>
<td>NUDCD3</td>
<td>59.4 ± 6.1</td>
<td>63.6 ± 9.0</td>
<td>0.09</td>
<td>0.19</td>
<td>67.6 ± 11.6</td>
<td>69.7 ± 11.1</td>
<td>0.47</td>
<td>0.76</td>
</tr>
<tr>
<td>PCDHB1</td>
<td>66.6 ± 23.3</td>
<td>50.9 ± 36.2</td>
<td>0.51</td>
<td>0.69</td>
<td>58.4 ± 27.6</td>
<td>58.0 ± 25.0</td>
<td>0.97</td>
<td>0.77</td>
</tr>
<tr>
<td>RNF219</td>
<td>35.0 ± 18.1</td>
<td>55.0 ± 17.6</td>
<td>(3 \times 10^{-41})</td>
<td>(3 \times 10^{-9})</td>
<td>35.3 ± 31.8</td>
<td>46.2 ± 23.7</td>
<td>(1 \times 10^{-4})</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Statistically significant differences (\(p < 0.05\)) are in \textbf{bold}.

\(a\) Mean cirDNA modification per amplicon.

\(b\) LME model \(p\)-value.
Figure 7. Fine mapping of modified cytosines in the *RNF219* gene in prostate cancer and control individuals
DNA modification for each CpG position studied in the *RNF219* gene. Control samples had higher content of modified cytosines than prostate cancer samples in all CpG positions within the studied region. Y-axis shows the DNA modification percentage, while the studied CpG positions are depicted in the X-axis. CpG positions within a restriction site are marked by a red square (position 6 and 7 are located within *HinPI* sites, while position 14 is within a *HpaII* site). Solid and dashed lines represent the DNA modification percentage in prostate cancer and control samples, respectively.
Figure 8. DNA modification in the RNF219 gene in two independent sample sets.
The studied region showed statistically significant differences in the first sample set (yellow box plots) between prostate cancer (PCa1) and control (Ctrl1) individuals (Mann Whitney test; \( p = 7.5 \times 10^{-4} \)). In the second sample set (PCa2 and Ctrl2, red boxes) prostate cancer patients and controls did not show statistically significant differences when three outliers from the control group were removed, as detailed in the text (Mann Whitney test; \( p = 0.12 \)). When both sample sets were considered together (PCa 1-2 and Ctrl 1-2, orange boxes) differences between prostate cancer cases and controls were statistically significant (Mann Whitney test; \( p = 5.3 \times 10^{-4} \)).

Average DNA modification is presented in the Y-axis, while the studied groups are accommodated in the X-axis.
3.1.3 Predictive value of DNA modification of RNF219

Since RNF219 exhibited the largest and significant differences in both training and validation cohorts (Table 5), we evaluated its ability to distinguish whether a sample was derived from an individual with PCa, BPH, or from healthy controls, using receiver operating characteristic (ROC) and area under the curve (AUC) analysis. While sample set 1 consisted of prostate cancer cases and control individuals without history of prostate disease, sample set 2 consisted of prostate cancer cases and BPH patients. The mean cirDNA modification values at RNF219 in sample sets 1 and 2 were used as training and testing cohorts, respectively. The mean cirDNA modification between 49.9% and 53.4% exhibited 89% sensitivity and 71% specificity in the training cohort (AUC = 0.79). Applying this analysis to the testing cohort resulted in 61% sensitivity and 71% specificity (AUC = 0.56, Figure 9).
Figure 9. Prediction accuracy of the cirDNA modification level in the RNF219 locus. Mean DNA modification percentages in RNF219 from sample set 1 (blue) and sample set 2 (red) were used to train classifiers. The optimal cutoff value (49.9% cirDNA modification difference between cases and controls) showed false-positive rate (FPR) = 0.29 (dashed vertical line) and true positive rate (TPR) = 0.89 in training cohort. Extrapolating the FPR value at the cutoff from the training to the testing cohort yielded TPR = 0.611 (dashed horizontal line). The AUC was 0.79 and 0.56 for training and testing cohorts, respectively.
3.1.4 Analysis of pericentromeric DNA repeats in chromosome 10

The number of microarray probes with disease-dependent, age-independent differences was overrepresented in chromosome 10 (Fisher’s exact test; $p = 7.8 \times 10^{-5}$, OR= 8.42, 95% CI = 2.97–21.06). Remarkably, all these probes contained repetitive elements, mainly low-complexity tandem repeats clustering at the pericentromeric region (10q11.21), and showed lower signals in prostate cancer patients compared with controls ($M$-values 20.51 ± 0.07 for cancer and 0.32 ± 0.18 for controls; $q = 1.8 \times 10^{-6}$). The number of individual samples with averaged $M$-values lower than zero were significantly overrepresented in prostate cancer (14 of 19 samples, 74%) compared with control samples (5 of 20 samples, 25%) (Fisher’s exact test; $p = 0.004$, OR = 7.87, 95% CI = 1.66–45.43; Figure 10).

These pericentromeric differences may have occurred due to differential DNA modification and/or DNA copy number variation. Since mapping DNA modifications in repetitive elements is challenging, we focused on verifying DNA copy number variation in cancer patients and controls. We analyzed pericentromeric repetitive elements on chromosome 10 in 12 prostate cancer cases and 11 controls of our independent validation cohort (sample set 2). Fifteen cycles of adaptor-mediated amplification in cirDNA were performed in order to generate more DNA material, which was followed by real-time PCR using the repeat- specific primers. The amount of repetitive DNA template at the pericentromeric region of chromosome 10 in prostate cancer samples was similar to the control samples ($C_t$ = 28.3 ± 6.4 and 26.8 ± 3.6, respectively; $p = 0.50$). However, we observed heterogeneity in Ct values across samples. Stage III prostate cancer samples showed a significant loss of pericentromeric DNA compared with the stage II and control samples ($C_t$ means were 26.8, 25.2 and 37.9 for controls, stage II and stage III samples, respectively; $p = 0.017$, Kruskal–Wallis test). Post hoc tests showed significant differences for stage III samples compared with stage II samples (mean ranks of stages II and III were 6.54 and 16.45, respectively; $p = 0.009$, effect size = 0.75, $U$-test) and control samples (mean ranks of controls and stage III were 6 and 17, respectively; $p = 0.005$, effect size = 0.68, $U$-test), suggesting that the pericentromeric region of chromosome 10 exhibits copy number loss in stage III prostate cancer.
Figure 10. Clusters of differential microarray signals identified in pericentromeric regions.
Average microarray differences (M-values) for probes on chromosome 10 in prostate cancer (PCa) and control (Ctrl) samples. Green and red peaks correspond to the average M-values in prostate cancer and control samples, respectively. Probes are accommodated longitudinally on x-axis. Dashed vertical line represents the position of the centromere. M-values.
3.1.5 Epigenetic signature of PCa using machine-learning analyses

Next, we sought to determine if we could combine multiple loci into a single epigenetic signature that would discriminate blood samples derived from prostate cancer patients from those derived from healthy controls. We applied a nested leave-one-out cross-validation (LOOCV) approach combined with a machine-learning method (Random Forests) to estimate the error rate of classification in the microarray data set (Supplementary Material, Fig. S1) and applied two linear models to find disease-dependent loci \( (q < 0.25) \) and age-independent loci \( (q > 0.25) \). The resulting signature had an accuracy of 72\% (28 of 39 correct classifications; inner LOOCV loop accuracy ranging between 68 and 79\%; Figure 11A). The percentage of trees that voted to classify each sample as prostate cancer was assessed as a measure of intra- and inter-group heterogeneity (Figure 11B) and did not indicate a significant effect of tumor stage in the classification accuracy (Fisher’s exact test; \( P = 1 \)).

3.1.6 Differentially modified and differentially expressed genes in prostate tumor

To show how differences in cirDNA modification reflect those in primary tissue, we evaluated DNA modification and steady-state mRNA levels of candidate genes in primary prostate tumors. For this analysis, we relaxed statistical significance (i.e. increased our false-positive rate to reduce our false-negative rate) and identified 181 loci (representing 132 genes) showing disease-specific \( (q < 0.25) \) and age-independent \( (q > 0.25) \) cirDNA modifications. We evaluated these genes in a series of mRNA abundance (Best et al., 2005; Taylor et al., 2010; Wang et al., 2010) and DNA modification data sets (Kobayashi et al., 2011) from primary prostate tumors and healthy prostatic tissues (Figure 12A). We included all 132 genes regardless of the distance to the differentially modified loci, since it is known that functional regulatory elements can be extremely distal to the gene (Heintzman and Ren, 2009; Nobrega et al., 2003).
Figure 11. Machine-learning analysis of microarray-based cirDNA modification profiles.

(A) Prostate cancer votes (n= 100 000 trees) from our Random Forest classifier. True negatives (TN) refer to accurately classified control samples; false negatives (FN) refer to prostate cancer samples classified as controls; false positives (FP) refer to control samples classified as prostate cancer samples; and true positives (TP) refer to accurately classified prostate cancer samples. (B) Percentage of prostate cancer votes per sample. For each sample, the percentage of trees (n = 100 000) that voted to classify the sample as a prostate cancer sample is shown. Dark colored bars signify accurate classifications (true positives and true negatives in blue and red, respectively), whereas light color bars signify samples that were misclassified (false negatives and false positives in pink and light blue, respectively).
Despite the use of distinct patient cohorts, we identified genes exhibiting both cirDNA modification differences (in sample set 1) and disease-specific differences in the primary tumor (in a published data set) (Kobayashi et al., 2011). For example, 25 genes exhibited both cirDNA modification differences and significant changes in DNA modification in primary prostate tumors ($q < 0.05$ and fold-change in the upper quartile). Another 25 genes exhibited both cirDNA modification differences and significant ($q < 0.05$ and fold-change in the upper quartile) change in the steady-state mRNA level in primary prostate tumors (Supplementary Material, Table S2). Finally, we looked for evidence that the significant cirDNA genes are associated with patient prognosis in primary prostate tumors: of 919 genes in which DNA modification changes in primary prostate tumors was associated with disease-free survival, six also showed tumor-normal cirDNA differences ($ARHGAP6$, $CDK6$, $XPR1$, $PRKAG2$, $PIAS2$ and $C22orf15$). Similarly, mRNA abundances of five cirDNA genes ($ARHGAP6$, $CDK6$, $MEF2D$, $INSIG2$ and $RASEF$) were associated with patient survival ($q < 0.05$).

These findings provide independent lines of evidence that the loci showing differential cirDNA modification may have potential roles in disease initiation and progression. The 132 genes identified as differentially modified in cirDNA were ranked by comparison with the four public mRNA and DNA modification analyses (Figure 12B and C). Two genes, Rho GTPase-activating protein 6 isoform 4 ($ARHGAP6$) and cyclin-dependent kinase 6 ($CDK6$), showed both cirDNA differences and three other clinical-molecular associations (Figure 13).
Figure 12. Overlap between cirDNA analysis and public primary prostate tissue analyses.

(A) Venn diagram comparing genes with a significant fold-change in modification from cirDNA (Blood T/N_{modif.}), significant fold-change in DNA modification from primary prostate tumors (Prostate T/N_{modif.}) and significant fold change in mRNA abundance in primary prostate tumors (Prostate T/N_{mRNA}). The total number of genes identified as differentially modified in cirDNA did not overlap significantly with any individual public data set (hypergeometric distribution; \( p > 0.05 \)).

(B) Heatmap showing for each of the 132 genes identified in the cirDNA analysis statistical significance (yellow) in differential DNA modification between tumor and control samples (T/N_{modif.}), prognostic relevance based on DNA modification differences (HR_{modif.}), differential mRNA abundance between tumor and control samples (T/N_{mRNA}) and prognostic relevance on mRNA abundance (HR_{mRNA}). Genes not present in a data set are shown in gray. The ranking score is the ratio of significant data sets in which the given gene is deemed significant compared with the number of available data sets for that gene.

(C) Detailed view of the top 25 genes identified (ranking score \( \geq 0.5 \)).
Figure 13. Differential DNA modification and mRNA expression of ARHGAP6 and CDK6 in primary patient tissue.

The Kaplan–Meier curves showing the disease-free survival advantage from a decrease \[ARHGAP6 \text{ locus} \text{ (A)}\] and an increase \[CDK6 \text{ locus} \text{ (B)}\] in DNA modification in primary tissue samples. Violin plots showing the increased steady-state \textit{ARHGAP6} (C) and \textit{CDK6} (D) mRNA abundances between primary prostate tumors and normal prostate tissue. The Kaplan–Meier curves showing increased patient survival with increased \textit{ARHGAP6} (E) and \textit{CDK6} (F) mRNA abundance in primary tissue samples.
3.2 CirDNA analysis in prostate cancer using high density tiling microarrays

3.2.1 Microarray based epigenetic profiling of cirDNA in prostate cancer patients and controls

Since we obtained promising results from our pilot study on 12K CpG Island microarrays, we aimed to improve our results by increasing the genome coverage (from 1% to 15%) using Affymetrix GeneChip® human tiling arrays covering chromosomes 2, 9, and 19; and by increasing the number of samples from n=97 to n=200. The original sample set consisted of cirDNA samples from 100 men diagnosed with PCa (PCa group) and 100 control samples (individuals with two negative biopsies; CTRL group). The PCa group was further divided in patients with Gleason score (GS) = 6 (low-GS group, n=50) and patients with GS>6 (high-GS group, n=50). Table 1 describes the clinical and demographic characteristics of each group. The modified fraction of cirDNA was enriched by a strategy combining digestion with DNA modification sensitive restriction enzymes and adaptor mediated PCR. The enriched template was interrogated on a tiling microarray covering full human chromosomes 2, 9 and 19, as detailed earlier in the materials and methods.

Following data preprocessing and outlier removal, we identified individual probes displaying differential cirDNA modifications between the PCa and CTRL groups. We employed a multivariate linear regression to model the effects of cirDNA modification changes as a function of PCa risk groups, microarray batches, and patient age. Utilizing a FDR threshold of 1%, we detected 978,259 probes showing significant cirDNA modification differences between the PCa and CTRL groups. Of these, 618,373 and 359,886 exhibited increased and decreased DNA modification in PCa group compared to CTRL group, respectively; with probes exhibiting up to four-fold differences in modification relative to controls (Figure 14A). When Gleason score was considered to stratify the PCa patient population, we identified 375,542 probes with differential cirDNA modification between the CTRL group and the low-GS group (FDR p-value < 0.01) (Figure 14B); and 331,584 probes between the CTRL and high-GS groups (FDR-corrected p-value < 0.01) (Figure 14C). We detected a large overlap of probes differentially modified in the two groups, with 247,380 shared probes and 459,746 distinct probes between the low-GS and high-GS groups compared to CTRL group (Figure 14D). Such overlap was significantly higher than that expected by chance (p=2.2 x10^{-16}, OR=45.3, 95% CI=44.81-45.71, Fisher's Exact Test for Count Data).
Figure 14. cirDNA modification profiling in PCa and control individuals.
A) Volcano plot of microarray data for differential cirDNA modification between PCa patients and controls. The x-axis represents the degree of differential cirDNA modification expressed as the log2-transformed ratio of the probe intensities between the two groups. The y-axis represents the significance of such difference expressed as FDR-adjusted p-value. Probe counts are depicted as a gray-scale gradient. B) Volcano plot for microarray data for High-GS and CTRL groups. X- and y-axes are depicted as in panel A. C) Volcano plot between Low-GS and CTRL groups. X- and y-axes are depicted as in panel A. D) Venn diagram showing the overlap between the differentially modified probes. Red and blue circles represent the differentially modified probes between High-GS and CTRL groups (n=622,922) and between the Low-GS and CTRL group (n=331,584), respectively. The overlapped zone between the two circles contain the differentially modified probes in both comparisons (n=247,380).
Next, we applied a sliding-window strategy (as described earlier in the Materials and Methods) to combine probes in the microarray that mapped adjacently in the genome and may define extended regions of differential cirDNA modification between the PCa and CTRL groups. We defined 1,081 candidate regions, with 646 and 435 regions showing a statistically significant increase in cirDNA modification in the PCa and CTRL groups, respectively (FDR p-value <0.001). Unsupervised clustering showed that most of the PCa samples separated from the CTRL samples based on the cirDNA modification status of the differentially modified regions (Figure 15A). Regions with increased cirDNA modification were (mean length; 95% CI): 486 bp; (364-584 bp) and 532 bp (433-601 bp) in PCa and CTRL groups, respectively (p=5.35 x 10^{-4}, Wilcoxon rank sum test; Figure 15B). Moreover, regions with increased cirDNA modification in the PCa group were more often intergenic (n=260 transcript associated regions and n=386 intergenic regions), whereas regions with increased cirDNA modification in the CTRL group were more frequently associated to protein coding genes (n=265 transcript associated regions and n=170 intergenic regions) (p=2.73 x 10^{-11}, Fisher’s Exact test).
Figure 15. cirDNA modification profiling in PCa and control individuals (continued).  
A) Heatmap of 1,081 regions showing significant cirDNA modification differences between the PCa and the CTRL groups. Differentially modified probes and samples are accommodated in columns and rows, respectively. Color gradient represents the differences in probe signal intensity, ranging from red (MAT score=−5, higher intensity in CTRL group), over white (MAT score=0, no differences) to violet (MAT score=5, higher intensity on PCa. Colored bar to the right indicates the group for the sample in each row (Green=PCa, Orange= CTRL) 
B) Density plot of the length of the candidate region in each group. X-axis depicts the length of the region in bp. Green and orange lines correspond to PCa and CTRL groups, respectively.
3.2.2 Functional Analysis

In order to understand the functional role of genes associated to regions of differential cirDNA modification, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID). Functional annotation clustering was performed using the list of candidate genes, and the enrichment score for each cluster was registered. The most enriched annotation cluster (enrichment score: 5.7) was related to the UCSC_TFBS (Transcription factor binding sites annotated at the UCSC browser). Of the 45 transcription factors listed in the cluster, 44 showed significant enrichment after multiple testing correction (Benjamini-Hochberg FDR p-value <0.05). The second highest enriched annotation cluster was involved in alternative splicing and splice variants. Using the MotifMap transcription factor database, we identified binding motifs for each of the transcription factors (Supplementary Material Table S1). Out of the 44 candidate transcription factor binding sites, 26 were listed in the database, with 13 of the entries containing DNA sequences that were amenable to possible epigenetic regulation (i.e. CG dinucleotide in at least 1 of the top 5 binding motifs). Moreover, 16 transcription factors were shown to play a role in cancer cell proliferation, metastasis, and cancer promotion (Supplementary Material, Table S1). Among them, four factors exhibited specific roles in PCa induction and progression: e.g. FOXO3 in cell proliferation and apoptosis in PCa cells (Chatterjee et al., 2013; Li et al., 2007, 2008b; Lynch et al., 2005; Yang et al., 2005a).

In addition, DAVID analysis enabled the identification of Gene Ontology (GO) terms significantly enriched (FDR p < 0.05) in the genes associated to regions of differential cirDNA modification. Within the Biological Processes category (GOTERM_BP), the terms showing the highest fold enrichment was related to gamma-aminobutyric acid (GABA) secretion (42.7 fold change). A role in cell proliferation and migration have been reported for GABA and GABA receptors during tumorigenesis (Zhang et al., 2014). Moreover, within both the Cellular Compartment (GOTERM_CC) and Molecular Function (GOTERM_MF) categories, the highest fold enrichment terms exhibited roles in cation channel complexes (11.9 fold change) and activity (11.5 fold change), including calcium, potassium, sodium, and proton transport. Cation channels play a major role in cell cycle progression and cell proliferation and division, which can be altered in cancer cells (Capiod, 2013; Choi et al., 2011; Misra et al., 2011).
Lastly, to classify the list of candidate genes into functionally related gene networks, we analyzed the 325 differentially modified genes between the PCa and CTRL group using the DAVID gene functional classification tool. The GO terms enriched by these 325 genes (169 displaying increased and 156 displaying decreased DNA modification when comparing PCa and CTRL groups, respectively) were assigned to 15 groups ranked by their biological significance. Seven of the thirteen genes included in the top cluster (enrichment score: 3.17) are involved in carcinogenesis, e.g. TSSC1—a tumor suppressing gene associated with lung, ovarian, and breast cancer (Hu et al., 1997; Wang et al., 2013). In the cluster with the second highest enrichment score of 2.86, twelve of the nineteen genes had roles in various cancer types, with four of the twelve genes (e.g. PKN3, ERBB4, PRKCE, and TYK2) related to PCa metastasis (Benavides et al., 2011; Leenders et al., 2004; Wang et al., 2014; Williams et al., 2003; Wu et al., 2002a).

3.2.3 Performance of an epigenetic signature as a diagnostic biomarker for PCa

We applied an iterative nested machine learning approach to build groups of microarray probes displaying differential cirDNA modification between the PCa and CTRL groups, towards the identification of an “epigenetic signature” that can be used for PCa screening using a blood-based test (Figure 4). The original dataset (n=183 samples) was divided into unequal sets: 110 (60%) and 73 (40%) samples were randomly assigned to the training and testing sets, respectively. Next, we further separated the training set into new training (n=75) and testing (n=35) subsets and evaluated the performance of the signature using random forest with Gini ranked features (Menze et al., 2009), as detailed previously in the materials and methods. When applied to differentiate cases and controls in the testing set, we registered separation between the groups at 85% sensitivity, 78% specificity and 82% accuracy (Figure 16A). To evaluate performance of epigenetic signature, we compared its performance as biomarker against a null-distribution for the dataset. The classifiers outperformed 90% of 1000 randomly selected probes in sensitivity, 71% of 1000 randomly selected features in specificity, and 98% of 1000 randomly selected features in accuracy (Figure 16B). When we analyzed the performance of the biomarker signature while incorporating accompanying PSA levels (Figure 16C), we further observed similar levels of accuracy at 81% – with a marginal increase in specificity to 80% at the cost of sensitivity at 83%.
3.2.4 Performance of an epigenetic signature as prognostic biomarker for PCa

To evaluate the performance of cirDNA modification profiles for the discrimination between high- and low-risk PCa patients, we built an epigenetic signature based on the two groups of PCa patients in the study: low-GS group (n=50) and high-GS group (n=50) (Table 1). Similarly to the algorithm for the diagnostic biomarker (Figure 4), we applied iterative nested machine learning for building prognostic epigenetic signatures based either on microarray probes (Figure 17A) or candidate regions (Figure 17B) that displayed cirDNA modification differences between the low- and high-GS groups. For both signatures, models were tested by 5-fold cross-validation and repeated 20 times.

The performance as prognostic biomarkers for both signatures was moderate. The probe-wise epigenetic signature trained on Random Forest (Figure 17A) showed 68% balanced accuracy rate (BAC), 63% area under the curve, 72% sensitivity and 67% specificity. Similarly, the epigenetic signature based on candidate regions (Figure 17B) showed 68% balanced accuracy rate, 62% area under the curve, 69% sensitivity and 69% specificity. While the area under the curve represents the performance of the classifier, the balanced accuracy rate is calculated as an average of the sensitivity and specificity (Li et al., 2008a).
Figure 16. Epigenetic signature as diagnostic biomarker for PCa.
A) Box plots indicating the distribution of sensitivity (blue box, median=85%), specificity (green box, median=78%), accuracy (red box, median=82%) values for the epigenetic signature defined by the machine learning algorithm (Figure 4) after correction by Gini coefficient. B) Epigenetic signature outperformed randomly selected probes in sensitivity (upper panel), specificity (middle panel) and accuracy (lower panel). In each panel, the first box at the left represents the distribution of values for the epigenetic signature, whereas the subsequent boxes depict the randomly selected probes (n=30, 100, 300, 500, 1000 probes, respectively). The red numbers on the bottom of each performance measure represent the percentage of randomly selected features that achieved a performance (in sensitivity, specificity or accuracy) greater than or equal to the median performance of the epigenetic signature. C) Box plots indicating the sensitivity, specificity and accuracy values for the epigenetic signature when combined with the assessment of PSA levels in plasma.
Figure 17. Epigenetic signature as prognostic biomarker in PCa.

Box plots indicating the performance of epigenetic signatures built based on probes (A) or regions (B) showing differential cirDNA modification between Low- and High-GS groups. Each box plot depicts the distribution of BAC, AUC, Sensitivity and Specificity parameters after 5-fold cross-validation and 20 repetitions.
Chapter 4
Discussion

4 DISCUSSION

4.1 Prostate cancer pilot study of DNA modification differences in plasma cirDNA

Many efforts have been directed towards the discovery of biomarkers for the diagnosis, prognosis and treatment-planning of cancer (Leman and Getzenberg, 2009). The biomarker potential for differentially modified loci in cirDNA isolated from plasma or other bodily fluids has been reported previously based on limited sample sizes and candidate genes (An et al., 2002; Anglim et al., 2008a; Bastian et al., 2005, 2008; Bearzatto et al., 2002; deVos et al., 2009; Ellinger et al., 2008b; Esteller et al., 1999; Gonzalgo et al., 2004; Payne et al., 2009; Usadel et al., 2002). The differential modification coupled with hybridization method (Huang et al., 1999) was extensively applied for DNA methylome profiling using tissue samples and cell lines (Huang et al., 2010; Lee et al., 2010; Lewin et al., 2007; Lu et al., 2010; Schmidt et al., 2010; van Straten et al., 2010; Yan et al., 2002), but genome-wide DNA modification profiling had not been performed in plasma and other bodily fluids at the time of our pilot PCa study. We adapted differential modification plus a microarray hybridization approach for the genome-wide cirDNA modification profiling of plasma from PCa patients and control individuals. Our findings can be categorized into three groups: i) differential DNA modification at individual genes and loci; ii) epigenetic and/or genetic differences at the pericentromeric repeat region in chromosome 10, and iii) an epigenetic signature of PCa generated using machine learning algorithms.

4.1.1.1 DNA modification profiling

We first attempted to discriminate PCa patients from BPH patients, but our microarray analysis did not reveal statistically significant cirDNA modification differences between the two groups. This is likely caused by the low sensitivity of our microarray technology and the fact that the two diseases may share etiopathogenic factors (Alcaraz et al., 2009), with some studies (Briganti et al., 2007) indicating that BPH might constitute a precursor for the development of PCa (Bostwick et al., 1992; Hammarsten and Högstedt, 2002). Given the lack of differences with BPH in the studied sample set, we compared cirDNA modification profiles of PCa patients with those from males without evident prostate pathology. Since BPH is a common condition in older men, occurring in about one quarter of men in their 50s and one-third of men in their 60s...
McVary, 2006), we selected a group of younger, asymptomatic controls. To account for age-dependent epigenetic changes (Calvanese et al., 2009; Issa, 2003), we used linear models to identify differences attributable to disease alone, and not to age.

By comparing the microarray profiles of PCa cases and controls, 117 loci out of 12,192 exhibited differential DNA modification ($q < 0.05$). Further, among these 117 loci, 5 and 112 loci showed higher and lower microarray signals in PCa compared to controls, reflecting a higher and lower density of modified cytosines in PCa patients, respectively. 73 of the 117 differentially modified loci were able to be mapped to genomic positions. The remainder were unable to be mapped, as the probes for the microarrays were generated from a CpG island library, where the remapping of probes is not fully complete (Heisler et al., 2005). 57 mapped probes, which revealed statistically significant differences between PCa and controls, contained repetitive elements – of which 2 and 55 showed increased and decreased microarray signals in PCa, respectively. A number of earlier studies detected that tumor epigenomes present with reduced global content of modified cytosines (Ehrlich, 2009; Gama-Sosa et al., 1983), which is thought to originate primarily from decreased modification of repetitive DNA elements (Ehrlich, 2009). Therefore, it is possible that the 55 statistically significant DNA modification differences reflect cancer-specific loss of modified cytosines in repetitive DNA elements.

From the 73 mapped loci exhibiting differential modification, 4 and 69 of them showed increased and decreased microarray signals in PCa compared to controls. In principle, this finding appears inconsistent with the general understanding that promoter-associated CpG islands generally exhibit a gain of cytosine modifications in cancer (Esteller, 2008b; Rauch et al., 2008). More recent and precise studies, however, reported similar degrees of modification changes in CpG island shores in colon cancer, with 44% and 56% showing decreased and increased DNA modification, respectively, in colorectal cancer samples compared to normal colon mucosa (Irizarry et al., 2009). In addition to this, a decrease of DNA modification of oncogenes has also been reported to be highly relevant in tumor formation and progression (Bhusari et al., 2011; Holm et al., 2005; Ogishima et al., 2005; Suzuki et al., 2006; Wolff et al., 2010). As a result, our candidate regions showing a decrease in DNA modification may represent genes with uncharacterized roles in oncogenesis. Another explanation for the loss of modified cytosines is that malignant cells shedding DNA into circulation represent an epigenetically distinct population that did not survive the evolutionary ‘bottlenecks’ of tumor growth. These
two interpretations are not mutually exclusive. To further address this issue, epigenetic profiles in cirDNA can be compared to the ones in tumor DNA. Distinct disease-specific epigenetic changes would argue that malignant cells forming tumors and malignant cells, which are prone to apoptosis and necrosis, represent different populations of tumor cells.

Of the 117 regions showing differential modifications between PCa patients and controls, 39 loci were age-independent, including 26 that were mapped to genomic regions. We identified 18 non-repetitive genomic regions associated with disease but not age. We validated seven of these using much more sensitive bisulfite sequencing analyses and exploited a second independent validation cohort (sample set 2) to confirm their disease association. We pioneered the use of a complex statistical modeling technique to epigenomic data: we applied LME models on the data generated by bisulfite-pyrosequencing enabling the detection of disease- and age-dependent differences on the candidate genes. Genes exhibiting age-dependent epigenetic changes, in fact, could be of primary interest because age is the major risk factor in prostate cancer.

4.1.1.2 Validation of microarray results by pyrosequencing

In our validation efforts, we employed the use of a different technology, bisulfite pyrosequencing, to assess the levels of modification at individual CpG sites in each of the seven non-repetitive genomic regions displaying differential DNA modification that were associated with disease but not age. We found that five of the seven loci (71%) examined by pyrosequencing showed concordant results with the microarray data. Further, two particular loci (KIAA1539 and RNF219) displayed significant disease-dependent ($P < 0.05$) and age-independent ($P > 0.05$) differences, specifically a loss of cytosine modification, between PCa samples and healthy controls in both the microarray and bisulfite pyrosequencing fine mapping experiments (Table 5).

The replication rate of the microarray data of 71% by fine DNA modification mapping using bisulfite pyrosequencing can be at least partially explained by the differences between the two technologies employed. In this experiment, the CpG island microarray probes interrogated large DNA regions that could span over several kilobases, although the interrogated CpGs were limited to the restriction sites. In contrast to this, bisulfite pyrosequencing allows for the assessment of modification levels at all CpG sites in much shorter DNA fragments. This
discrepancy between the probe size of the microarray (average length = 345 ± 237 bp) and the bisulfite pyrosequencing fragment size (~100-150bp) could result in differences in modification level results found between PCa and control samples by the two different methods. In this regard, microarray technologies that contain shorter probes, e.g. tiling arrays which contain 25 nucleotide long probes, have the potential to increase the resolution of cirDNA modification profiles by evaluating fewer CpGs per probe. Moreover, next generation sequencing based approaches, when combined with bisulfite modification, would be even more accurate, allowing interrogation at a single nucleotide resolution.

4.1.2 Predictive value of DNA modification of RNF219

From the results of our fine mapping effort, RNF219 exhibited the largest differences between PCa patients and controls. This gene is located in chromosome 13q31.1 and encodes the ring finger protein 219. Although its function is not yet completely understood, this intracellular protein binds a zinc ion and has been reported as post-translationally phosphorylated upon DNA damage (Matsuoka et al., 2007), likely participating in signal transduction cascades. RNF219 is a member of the RING-type zinc fingers protein family; which consists of 235 members – amongst them the breast cancer 1, early onset gene (BRCA1) and the BMI1 polycomb ring finger oncogene (BMI1). BRCA1 is a well-established breast cancer susceptibility gene and its germline mutations are found in 40–50% of hereditary breast cancers (Miki et al., 1994). In turn, BMI1 has been reported as a crucial regulator of self-renewal in adult prostate cells and plays important roles in PCa initiation and progression (Lukacs et al., 2010).

With significant DNA modification differences in RNF219 detected between cancer and healthy controls using two different technologies in sample set 1, we aimed to verify these results in an independent patient cohort (sample set 2) that was matched for ethnicity, age, and tumor type. However, while the control individuals in the first sample set were disease-free, the control individuals in sample set 2 were diagnosed with BPH, a benign condition that has been linked with an increased risk for PCa (Alcaraz et al., 2009; Hammarsten and Högstedt, 2002). Additionally, while the results obtained from the study of sample set 2 were consistent with those found in sample set 1 – with PCa samples exhibiting decreased DNA modification compared to controls – the differences were not statistically significant. Interestingly, in sample set 2, we identified 3 individuals with RNF219 modification patterns that were more similar to the PCa
patients than the remaining controls. One individual had high PSA level before BPH surgery and two others developed non-prostate malignant disorders after BPH treatment; which may help to explain the lack of statistical significance in differential modification at RNF219 in sample set 2. By combining the bisulfite-pyrosequencing results from both sample sets to increase the sample size, we were able to observe significantly lower density of modified cytosines in PCa cases than in controls.

Further, we wanted to determine the predictive accuracy of cirDNA modification at RNF219 by using the first and second sample sets as training and testing cohorts respectively. As a single marker, we found that the optimal cutoff of 49.9% cirDNA modification difference between cases and controls exhibited 71% specificity and 61% sensitivity. This predictive performance is comparable with existing biomarkers for PCa (Wu et al., 2011). While our specificity is much higher than that of PSA (20%) (Catalona, 1994), it is also slightly lower than that for differential modification in GSTP1 (89%) found in plasma (Wu et al., 2011). In turn, sensitivity of 61% is lower than that of PSA levels (80%) (Catalona, 1994), but higher than differential modification in GSTP1 (52%) (Wu et al., 2011).

The ability to achieve moderate predictive value using RNF219 as a single marker is promising considering the limited sample size and coverage of the CpG island arrays, which interrogated only 1% of the epigenome. It further suggests that studies performed on larger sample sets with more informative technologies, including higher resolution tiling microarrays, have the potential to uncover other novel markers with even higher performing diagnostic value. While further validation is needed to determine the utility of RNF219 as a diagnostic biomarker for PCa; nevertheless, our data spurred us to combine multiple loci displaying alterations in their modification profiles into a single diagnostic signature that may provide higher specificity and sensitivity than any individual markers alone (Anglim et al., 2008b; Szyf, 2012).

4.1.3 Improved accuracy with biomarker panels

Since cancer presents as a heterogeneous disease, it is unlikely that a single molecular alteration would be present in every cancer case. While epigenetic changes, such as DNA modification, at a single gene may demonstrate predictive value, it is likely that it would only allow the detection of a subset of cancers. As a result, the assembly of a panel of differentially modified genes and non-coding DNA regions would therefore increase predictive performance (Anglim et al., 2008a;
Tsou et al., 2007). In particular, the development of diagnostic platforms that allow parallel investigation of multiple markers in a reproducible and rapid manner, such as the microarray, presents opportunity for biomarker signature sets where a minimal subset of molecular quantities can be used together to produce maximal predictive performance. A widely accepted method for identifying such signatures involves the measurement of large sets of molecular changes from large sample sets of biological specimens, and following that with data-analysis approaches that allow for the selection of the most informative set of features.

In this regard, we constructed a PCa signature using a machine learning technique known as Random Forests (Breiman, 2001). Many genes in the molecular signature have been shown to be associated with tumorigenesis and tumor progression in prostate and other cancer types. For example, expression of pituitary tumor-transforming protein 1 (PTTG1) was detected in a high percentage of PCa tissues (34/41, 82.9%), but to a much lesser extent in non-malignant tissues (5/14, 35.7%) (Zhu et al., 2006). In addition, the promoter region of the C20orf15, also known as Low in Lung Cancer 1 (LLC1), gene was modified in seven out of nine lung cancer cell lines (Hong et al., 2007). Through this method, we were able to distinguish cirDNA derived from PCa patients from that derived from controls with 72% accuracy. We anticipate that this performance would be further improved by considering larger data sets and more homogeneous patient populations (i.e. similar tumor stage and Gleason scores).

Machine learning algorithms have previously been employed to build molecular signatures for cancer management, e.g. a six gene panel was sufficient for accurately classifying 100% of basal cell skin carcinomas from control specimens (O’Driscoll et al., 2006). Similarly, a panel of 35 nucleosides and ribosylated metabolites distinguished breast cancer patients from controls with 83% specificity and 90% sensitivity (Henneges et al., 2009). However, while these earlier studies applied machine-learning techniques to study malignant tissue samples, this is the first time, to our knowledge, that machine-learning techniques have been applied in the study of cirDNA modification.

4.1.4 Changes in repetitive DNA elements in cancer

In the second group of analyses, we detected significantly weaker microarray signals in repetitive pericentromeric elements at 10q11.21. Interpretation of this finding is not straightforward, since the observed pericentromeric differences might result from epigenetic changes and/or copy
number variation. Molecular alterations in pericentromeric regions leading to aneuploidy and
defective centromere/kinetochore assembly have been known to occur in carcinogenesis (Pezer
and Ugarković, 2008). In particular, genetic alterations in chromosome 10 have been widely
studied in PCa, with reported deletions (Matsuyama et al., 2003), allelic imbalances (Lu and
Hano, 2008), loss of heterozygosity (Phillips et al., 1994) and single-nucleotide polymorphisms
in non-coding regions of the genome associated with increased PCa risk (Yeager et al., 2009).
However, only allelic imbalances in a few tumor suppressor genes were reported in cirDNA of
PCa patients (Schwarzenbach et al., 2011b).

Quantitative PCR targeted at the pericentromeric repeats in chromosome 10 revealed that
the number of DNA repeats were lower, although non-significantly, in PCa patients compared to
controls. The difference, however, was very significant when stage 3 PCa patients were
compared to the remaining samples ($p = 0.017$, Kruskal–Wallis test). This finding indicates that,
in addition to epigenomic studies of cirDNA, the search for structural differences may also be a
productive line of research. In order to differentiate epigenetic- and DNA sequence-effects, the
same cirDNA sample should be tested twice: one experiment would interrogate the enriched
modified or unmodified DNA fraction, while the second one would be a standard analysis of
DNA copy number variation. Presence of differences in the latter would immediately suggest
structural DNA aberrations, while the former would estimate the contribution of combined
structural and epigenetic effects. At the time of our study, this represented the first report of copy
number variation in repetitive elements in cirDNA of PCa patients.

4.1.5 Primary PCa tumor analysis

To help prioritize candidate genes from our analysis, we applied systems biology techniques and
integrated cirDNA modification profiles with public data sets evaluating DNA modification and
mRNA data in primary prostate cancers. Despite major variations in study design and analysis,
this analysis identified a subgroup of genes with significant differences in PCa compared with
normal controls in various molecular contexts, including $ARHGAP6$ and $CDK6$. These genes
have the potential to serve as prognostic markers for PCa, although deeper investigations are
required to define whether differential expression or epigenetic modification of these genes is
involved in PCa carcinogenesis. Interestingly, a lower density of modification in $ARHGAP6$
appears to provide a protective advantage by preventing disease recurrence (Figure 13). On the
other hand, we do not see this in *CDK6* where an increase in DNA modification is associated with better patient prognosis (Figure 13).

4.2 PCa tiling microarray study

4.2.1 DNA modification profiling on high resolution tiling arrays

In our pilot PCa study, we developed a method for large-scale cirDNA modification profiling, which can be applied for the discovery of novel biomarkers. In this study, we improved our results by using a similar approach for enriching the modified fraction of the genome, followed by hybridization to Affymetrix tiling microarrays covering the non-repetitive regions of chromosomes 2, 9 and 19 (approximately 15% of the genome). While the CpG island microarrays from the pilot study contained only ~12,000 probes and interrogated 1% of the epigenome, we were still able to detect 117 regions exhibiting statistically significant differences between affected PCa patients (n=20) and healthy controls without any known prostate disease (n=20) (*q* < 0.05). As a result, we hypothesized that by employing the use of technologies that have higher resolution and exhibit greater coverage of the genome, we could discover novel and more informative biomarkers that may be combined into an epigenetic signature for early detection of PCa.

We increased the sample size by two-fold with 100 PCa cases and 100 controls. In the PCa group, half of the samples were individuals with aggressive PCa (Gleason score >6 at biopsy or radical prostatectomy), and the other half were individuals with indolent PCa (Gleason score of 6 at biopsy or radical prostatectomy). This increase in sample size allowed us to have more statistical power when discovering candidate regions, and the stratification of patients presented us with an opportunity to develop potential epigenetic signatures for prognosis of PCa using this low-invasive blood-based test.

The improvement in genome coverage and increase in sample size resulted in a higher number of significantly differentially modified loci detected between PCa and control groups (1,081 candidate regions compared to the 117 detected in our pilot study). These regions were generated using a sliding window strategy: combining neighboring probes that mapped adjacent in the genome and exhibited similar differential cirDNA modification between PCa and the control groups. This allowed us to move from a probe-level analysis to a region-level analysis that may reveal long-range epigenetic effects.
Of our 1,081 candidate regions, 646 regions showed increased cirDNA modification and 435 regions showed decreased cirDNA modification in PCa compared to controls. This predominant increase in DNA modification is in contrast to the pattern observed in our pilot study, where the majority of the loci exhibiting differential modification showed a decrease in microarray signal in PCa compared with control. This can be explained partly by the fact that repetitive elements have been removed from the Affymetrix GeneChip tiling arrays, and all interrogated loci corresponded to unique sequences.

In addition to the increased resolution and coverage of our tiling arrays, another advantage is the ability to detect cancer-specific cirDNA modification differences outside of CpG islands that have not been widely studied, but have potential for serving as markers for cancer diagnosis and prognosis. In this study, we found that regions with increased cirDNA modification in tumors were more frequently intergenic (n=260 transcript associated regions and n=386 intergenic regions). While DNA modifications in intergenic regions have not been as well characterized as similar effects in CpG islands, recent research suggests that aberrant modification patterns in these regions could have functional consequences. One study showed that modification of intergenic regions can play a role in long range regulation of genes, or may even be involved in regulating nearby transcripts that are not yet known or have yet to be annotated (Yegnasubramanian et al., 2011). For example, differential modification was observed at an intergenic region close to MEG3 (maternally expressed 3), a gene that encodes a noncoding RNA involved with the suppression of cell growth pituitary tumors. This group found that increased DNA modification at this intergenic region resulted in a loss of MEG3 expression and has been linked with pituitary adenomas (Gejman et al., 2008). Furthermore, comparing our intergenic regions with chromatin states defined in PCa cells can be conducted to obtain a better understanding of their functional roles in gene regulation.

4.2.2 Predictive performance of epigenetic signature

As with our pilot PCa study, we aimed to improve the accuracy of detecting PCa by creating an epigenetic signature using the differentially modified probes from our tiling microarrays. Employing the use of machine learning algorithms, we achieved a better predictive performance of 85% sensitivity, 78% specificity and 82% accuracy, as compared with the 72% accuracy achieved by our low-resolution microarrays in our pilot study. Therefore, it can be expected that
extending the coverage to the full genome, we will be able to design a biomarker panel with even higher levels of accuracy.

Further, we examined the effect of combining our epigenetic signature with the patient’s associated PSA levels. While we observed a nominal improvement in specificity increasing to 80%, sensitivity and accuracy exhibited a 2% and 1% decrease respectively. Since our test displays a higher specificity than PSA (20%) (Catalona, 1994), it may be of greater value to combine the strengths of both tests by using the two sequentially rather than simultaneously. As a result, the PSA test could be used to screen out potential patients, with our epigenetic signature assay given only to those patients that exhibited elevated PSA levels. By increasing the specificity of the test above that of PSA alone, the number of unnecessary biopsies could be reduced considerably. This type of serial testing is widely used clinically for embolisms and diarrhea (Fekety, 1997; Wells et al., 2001).

Since there are no biomarkers or prognostic tools that can accurately predict tumor progression at the time of diagnosis, distinction between indolent and aggressive PCa remains a major challenge for the management of the disease. Currently, PCa staging is based almost exclusively on routine clinical pathological parameters such as serum PSA, clinical stage, and Gleason score (Kattan et al., 2003; Steyerberg et al., 2007). While several nomograms exist, none have been accepted as the gold standard for staging, and as a result, many individuals diagnosed with PCa are misclassified, which can result in suboptimal treatment decisions (Ploussard et al., 2011). Furthermore, available clinical nomograms have not typically included molecular parameters despite mounting evidence for the prognostic value of such biomarkers (Shariat et al., 2008). In our study, patient samples were stratified by Gleason score staging and the utility of an epigenetic signature of cirDNA for identifying indolent PCa patients from aggressive PCa patients was evaluated. Several studies have shown correlations of DNA modification with clinicopathological markers such as Gleason Score (GS) and tumor stage (Cottrell et al., 2007; Kron et al., 2012, 2010; Liu et al., 2011a; Moritz et al., 2013). As a result, we hypothesized that classification based on modification profiles could be useful for staging PCa and guide treatment decisions (Li et al., 2005).

From our tiling microarrays, we identified a higher number of significant differentially modified probes between our high-GS and control groups (n=622,922) than between the low-GS
and control groups (n=331,584), which aligns with the expectation that low-grade cancer cells with Gleason scores of 6 would be more similar to healthy cells than higher-grade, poorly differentiated neoplasia. To test whether these cirDNA modification profiles could discriminate between high and low-risk PCa patients, we constructed an epigenetic signature based on the two groups of PCa patients in the study. Using a machine-learning algorithm based on microarray probes or candidate regions generated using a sliding window strategy, we discovered a moderate predictive performance (72% sensitivity and 67% specificity) of our epigenetic signature as a prognostic marker. One reason for such modest values may include the limited number of chromosomes (2, 9 and 19) studied from our single Affymetrix tiling array. It is possible that using the full set of tiling arrays, covering the entire non-repetitive genome, as well as increasing our sample size may improve the predictive performance of our epigenetic signatures. However, even on a limited number of chromosomes and a small number of samples, we were able to detect cirDNA modification differences between cancer groups that could differentiate between the indolent and aggressive cancer types to a moderate degree. This provides the basis for further efforts in the detection of prognostic markers in free floating DNA which may present as a non-invasive approach to PCa management, an area of research that continues to require much attention.

4.2.3 Gene ontology analysis

From the list of 1,081 candidate regions differentially modified between PCa patients and controls, we identified 329 associated genes – with 110 of these involved in cancer development. Bridging Integrator 1 (BIN1), for example, encodes an adaptor protein that interacts with c-MYC to regulate cell-cycle control, acting much like a tumor-suppressor by mediating apoptosis (Sakamuro et al., 1996). This gene, while completely unmodified in normal tissues, was found to be modified in 20% and 10% of breast and prostate tumors, respectively (Kuznetsova et al., 2007).

NAD-dependent protein deacetylase sirtuin-2 (SIRT2) is another gene that has been found to be involved with prostate tumor occurrence and progression (Hou et al., 2012). SIRT2 acts by changing the acetylation status of a cytoplasmic substrate known as cortactin, and studies have reported that this altered status can result in an increase of cell mobility, and consequently lead to the accelerated invasion and metastasis in tumor progression (Dryden et al., 2003). An increased
expression of SIRT2 has also been found in 38% of PCa samples as compared with benign prostate tissue, with a pattern of increased expression correlating with increasing Gleason scores (Hou et al., 2012). Further to this, upregulation of TCF3, a gene encoding Transcription Factor E2A, has also been detected in PCa and linked with promoting cell proliferation and conferring resistance to apoptosis (Patel and Chaudhary, 2012).

We also performed gene ontology (GO) analysis using Database for Annotation, Visualization and Integrated Discovery (DAVID). Using the functional annotation clustering on our list of candidate genes, we found that the differentially modified genes had a significant enrichment of conserved transcription factor binding sites. It is known that DNA modifications can regulate transcription through the mediation of transcription factor binding (Watt and Molloy, 1988). This was confirmed by our analysis of the transcription factor binding site DNA sequences, of which half (13/26) of the sites listed in the database overlapped with CG dinucleotides that were amenable to possible epigenetic regulation. Of these genes, 44 were significant after correction for multiple testing, and of those 16 were implicated in PCa cell proliferation, metastasis, and progression. For example, oncoprotein growth factor independent-1 (GFI1) has been linked with PCa development, as it functions as a repressor of tumor-suppressor genes (Dwivedi et al., 2005a, 2007); and decreased FOXO3 expression observed in PCa cells have been found to contribute to roles in apoptosis and cell proliferation in the malignancy (Chatterjee et al., 2013).

Our DAVID analysis also identified a number of GO terms that were significantly enriched ($q < 0.05$) in the genes associated to regions of differential cirDNA modification. In the biological processes category (GOTERM_BP) the highest fold enrichment term was related to gamma-aminobutyric acid (GABA) secretion. In PCa specifically, GABA has been shown to increase cellular proliferation and cellular invasion via the receptor subtypes GABAa and GABAb respectively (Abdul et al., 2008). We also observed enrichment of the terms related to cation channel complexes; cation channels are actively involved in cell cycle progression, proliferation, and division, which are often altered in cancer cells (Capiod, 2013; Choi et al., 2011; Misra et al., 2011). Moreover, specific Ca$^{2+}$ permeable cation channels have been linked to roles in apoptosis in human PCa cell lines (Gutiérrez et al., 1999). For example, increased expression of the transient receptor potential cation channel, subfamily V, member 6 (TRPV6) was found in 67% of prostate tumors as compared with benign prostate tissues (including BPH).
where it is not expressed. These findings have found an association of TRPV6 expression with PCa progression, and have suggested potential for its use as a prognostic marker and a therapeutic target (Fixemer et al., 2003).

4.3 Limitations and recommendations for future studies

This thesis provides a basis for future efforts in the discovery of biomarkers in cirDNA. In our studies, we used microarray platforms to interrogate the cirDNA fraction of the genome that have been enriched by combining DNA modification sensitive restriction enzymes and adaptor mediated PCR; and we were able to detect differences in DNA modification in an unbiased, genome-wide approach to identify novel candidate loci that exhibited promise for future studies. At the time of our studies, this represented the first report of its kind to comprehensively analyze the cirDNA modification profiles in cancer. However, since our approach is based on restriction enzyme digestion, the data is limited by the presence of enzyme recognition sites. As a result, genomic regions without the restriction site cannot be analyzed, leaving a number of CpG sites uninterrogated. This may be of importance if the phenotypic outcome, such as disease presence and/or development, depends on the DNA modification changes at these isolated cytosines outside of the enzyme’s recognition sites.

Another limitation of current enzymatic approaches is the inability to distinguish between 5-mC and 5-hmC (Nestor et al., 2010); in our study, the modification-sensitive restriction enzymes’ actions are blocked by both CpG modifications. As a result, the detected differences in our study may be a combination of 5-mC and 5-hmC modifications. Future cirDNA analyses will also need to consider distinguishing 5-mC and 5-hmC profiles. Recently several methods have been developed for the detection of 5-hmC using DNA glycosylation and modification-insensitive restriction enzymes (Davis and Vaisvila, 2011; Song et al., 2011). The combination of these techniques with our current method will provide an integral strategy for DNA modification profiling of cirDNA. Beyond 5-mC and 5-hmC, it can be expected that the presence of other covalently modified nucleotides in DNA could distinguish between cases and controls. For example, another DNA modification, 5-hydroxymethyl-2′-deoxyuridine, was reported as significantly increased in peripheral blood cells of breast cancer patients and has been suggested as a candidate biomarker for breast cancer diagnosis (Djuric et al., 1996). The development of
precise methods for the detection and quantification of modified nucleotides will enable us to test the biomarker potential of these modifications.

The use of microarrays for DNA modification profiling is a commonly used method to identify broad differences between groups of samples, as it represents a less time-consuming, less labor intensive, and less costly method compared with sequencing. In addition to this, microarrays allow for the simultaneous analysis of a larger number of samples with a wider coverage. However, while microarray platforms have been extensively used for years, their efficiency has been questioned given technical limitations that include low sensitivity, probe coverage, hybridization bias, and saturation. In this thesis, we addressed issues of low sensitivity and probe coverage by scaling up from CpG island microarrays to Affymetrix tiling microarrays. While the performance of the diagnostic markers found in our PCA tiling array study are acceptable and already display higher specificity values than PSA, it is also important to keep in mind that our discovery study interrogated only chromosomes 2, 9 and 19. It is possible that there are many more significant loci associated with PCA in the remaining genome, and requires further investigations. While the use of tiling arrays addressed probe and resolution limitations, it is important to note that it still excludes potentially interesting regions of the genome. These include repetitive elements and pericentromeric regions that are not included on the arrays due to their difficulty to map back to the genome.

The findings of our studies provide many possible directions for future research. Initially, the major results of our studies will require replication in larger cohorts of patients. These experiments should be performed using significantly larger sample sizes of affected individuals and controls to achieve higher statistical power in the detection of individual markers, and more precise molecular signatures for training and testing steps in machine learning approaches. Further the use of more informative platforms such as next generation sequencing (NGS) would more accurately assess individual CpG positions. Next generation sequencing platforms provide an unprecedented opportunity for fine and high-throughput mapping of epigenetic signals, including genome-wide mapping of modified cytosines at single base resolution (Lister and Ecker, 2009; Lister et al., 2009b). Currently, the cost of NGS-based population studies is still prohibitive for large scale population studies. Due to the mosaicism of DNA modification patterns across cells, deep-sequencing (higher than 30X read coverage per cytosine) per sample is required – considerably elevating the cost per reaction. As the cost of sequencing decreases,
further studies using NGS platforms for interrogating cirDNA modification profiles in plasma are warranted. Moreover, our pilot study, although initially designed for the analysis of DNA modification differences, also points at possible changes of structural DNA sequence differences, which need to be investigated in parallel with their epigenetic counterparts.

4.4 **Summary of Findings**

To conclude, in this thesis, we applied a microarray-based technology for large-scale epigenomic profiling of plasma cirDNA in a population of PCa patients and controls. Overall, the results indicate that our approach represents a promising way to discover novel diagnostic or prognostic biomarkers in free-floating DNA. In our pilot proof-of-principle study, we demonstrated the feasibility of this new approach in investigating epigenetic changes in cirDNA of PCa patients. Further, we found moderate predictive values despite a limited number of samples and a microarray platform that interrogated less than 1% of the epigenome. Our pilot study represented a significant improvement over existing research not only by employing an unbiased approach in discovering candidate markers, but by assessing cirDNA – which may have translational potential as a non-invasive marker of disease presence. Our results represented the first comprehensive analysis of epigenetic patterns in plasma cirDNA in cancer.

Spurred by the results of our pilot study, we sought to expand the scope of our research by assessing the utility of our technique in the discovery of novel diagnostic and prognostic markers using more informative platforms, such as the *Affymetrix* tiling microarray. Further, a relatively large independent sample set of PCa-affected individuals and matched controls were included in this study, with patients stratified according to tumor aggressiveness as indicated by Gleason score. By applying machine learning algorithms, we built an epigenetic signature that differentiated PCa cases and controls at relatively high specificity, sensitivity and accuracy. Moreover, we detected that a panel of markers could be used for cancer prognosis – an avenue of research that could improve patient survival and quality of life through better treatment selection. Through gene ontology analysis, we identified a number of genes related to PCa carcinogenesis and progression. The concordance between our data and that from literature, with respect to the identity of differentially modified genes and their function, strengthens the validity of our results.

This thesis shows that cirDNA modification profiling of human neoplasia using DNA microarrays can identify clinically relevant diagnostic and prognostic markers, defined upon the
combined modification levels of dozens or hundreds of genes. Validation studies on larger series of samples are now required to confirm these data, as well as studies investigating whether discriminator genes are, or are not, functionally relevant to oncogenesis and disease progression. By delineating these genes, understanding cancer development and progression will be improved, and new diagnostic and/or prognostic markers might soon be developed, leading to improvements in cancer management. We believe that the application of our method has the potential for the discovery of important biomarkers in other types of cancers, even expanding beyond diagnostic and prognostic applications, to include predictive markers that can help stratify patients for drug response and improve patient management in clinical settings.
References


Supplementary Figure S1. Flowchart of machine learning analysis.

A nested leave-one-out-cross-validation (LOOCV) approach was coupled to machine learning with random forests to build a classifier capable of distinguishing prostate cancer (PCa) from controls (Ctrl) based on plasma cirDNA modifications. Sample set 1, composed of 19 prostate cancer and 20 control samples, was used for testing and training. In short, one sample \(i\) is chosen as the validation sample and the remaining samples are used to determine the optimal number of methylation loci to use for training. Thus, after sample \(i\) is chosen, the LOOCV A loop is iterated 38 times such that each of these samples is used as the test sample, \(j\), exactly once. One random forest classifier is built per feature size (i.e. 3, 10, 30, 50, 75 and 100); the 'top features' are the specified number of loci which are most differentially modified between cases and controls using the 38 training samples. After the 38th iteration, the optimal feature size is selected based on the performance of each of the random forest classifiers and is used in the LOOCV B loop. In LOOCV A loop we determined the optimal number of features to build a random forest classifier (n=100,000 trees) for each sample in the training set. In the LOOCV B loop, the original 38 training samples are trained using another random forest with the specified number of features which were selected with a second t-test using the 38 training samples. Each
of the 39 samples is used as the validation sample \(i\) once, requiring 39 iterations of the LOOCV B loop. Finally, the final accuracy is calculated by determining the proportion of samples which were accurately classified in LOOCV B. The number of trees “voting” for classifying each validation sample as prostate cancer or control was computed. The category supported by the highest number of trees (>50%) was assigned to each sample, assessing the number of true and false calls.
### Supplementary Table S1: Binding motifs for transcription factors associated to regions of differential cirDNA modification.

<table>
<thead>
<tr>
<th>TF</th>
<th>Location</th>
<th>Sequence</th>
<th>Closest Gene</th>
<th>Reports for TF in cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>POU6F1</td>
<td>chr7:18504109-18501426</td>
<td>GAAACCACATTATGAGC</td>
<td>HDAC9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr1:163779827-163779844</td>
<td>GGCACACCTTATTGTGATG</td>
<td>LRRC92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr9:88752308-88752325</td>
<td>TGGGCCCCCTATTATGGCAT</td>
<td>GAS1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr8:92330642-92330659</td>
<td>AACTACCTTATTACCATT</td>
<td>SLC26A7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr9:92330642-92330659</td>
<td>GGAACCTAATTAAGGCT</td>
<td>SRGAP1</td>
<td></td>
</tr>
<tr>
<td>CHX10</td>
<td>chr16:71650529-71650543</td>
<td>AAGGGGCTATTAGCC</td>
<td>ZFHX3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr9:14712916-14712930</td>
<td>AGGGGGTATTAGCT</td>
<td>CER1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr7:13995355-13995369</td>
<td>TAGCTAATTACCCTC</td>
<td>ETV1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr13:45324741-45324755</td>
<td>CATAGCTAATTACC</td>
<td>SIAH3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr7:13995353-13995367</td>
<td>GGAACCTAATTAAGGCT</td>
<td>ETV1</td>
<td></td>
</tr>
<tr>
<td>SRY</td>
<td>chr10:74323253-74323265</td>
<td>AGGAAAAACATAGG</td>
<td>OIT3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr3:116273149-116273161</td>
<td>AGGAAAAACATAGG</td>
<td>ZBTB20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr1:31310852-31310864</td>
<td>TGTAAACATAGG</td>
<td>PUM1</td>
<td></td>
</tr>
<tr>
<td>HFH3</td>
<td>chr10:50177718-501777181</td>
<td>AGGCTGTGGTTTTA</td>
<td>C10orf71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chrX:144715781-144715794</td>
<td>TTAACAAACAGCC</td>
<td>CXorf1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr3:115825871-115825884</td>
<td>ATAAATAAACACAC</td>
<td>ZBTB20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:219750805-219750818</td>
<td>TGTTGTGTTTTA</td>
<td>FAM134A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr5:58608674-58608687</td>
<td>TTGGTTGTATTATA</td>
<td>PDE4D</td>
<td></td>
</tr>
<tr>
<td>LHX3</td>
<td>chr10:21826875-21826892</td>
<td>CAGTCATTAAATATACCC</td>
<td>C10orf114</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr10:21826876-21826893</td>
<td>AGTCAATTAAATATACCA</td>
<td>C10orf114</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr6:33698092-33698109</td>
<td>CAAACATTAAATTATCTT</td>
<td>ITPR3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr8:131483878-131483895</td>
<td>GGGATTTAAATATACCAT</td>
<td>ASAP1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr7:130070032-130070049</td>
<td>TAGCCATTAAATATAGC</td>
<td>KLF14</td>
<td></td>
</tr>
<tr>
<td>HLF</td>
<td>chr1:148868767-148868777</td>
<td>AATGACAGTAC</td>
<td>ENSA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:27788033-27788043</td>
<td>GTTACATCAT</td>
<td>C2orf31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr3:101077740-101077750</td>
<td>AGTCACTATCCG</td>
<td>FILIPI1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr3:101078245-101078255</td>
<td>GCCTCATATAACG</td>
<td>FILIPI1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr12:115803695-115803705</td>
<td>GGTACATTAGG</td>
<td>HRK</td>
<td></td>
</tr>
<tr>
<td>E4BP4</td>
<td>chr9:135389963-135389975</td>
<td>CCGTTACATAAAGC</td>
<td>ADAMTSL2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr20:4614755-4614767</td>
<td>CCGTTACATAAAGC</td>
<td>PRNP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:166713928-166713940</td>
<td>CGGTATTGTAAC</td>
<td>SCN1A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:202810997-202811009</td>
<td>CCGGTATGTAAC</td>
<td>SUMO1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:64224299-64224311</td>
<td>CCGGTATGTAAC</td>
<td>PEL1</td>
<td></td>
</tr>
<tr>
<td>SOX9</td>
<td>chr12:111979541-111979555</td>
<td>TCCCCATTGGTCTCA</td>
<td>DTX1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr11:61315972-61315986</td>
<td>TCCCCATTGGTCTGA</td>
<td>MIR611</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr13:25727493-25727507</td>
<td>CATAGAACAATGGG</td>
<td>CDK8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:191819016-191819030</td>
<td>GGGAGAACAATGGG</td>
<td>MOY1B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr17:35017990-35018004</td>
<td>AGGCGGAACAATGGG</td>
<td>NEUROD2</td>
<td></td>
</tr>
<tr>
<td>HNF3B</td>
<td>chr15:53997288-53997303</td>
<td>CTAGGCCCCCATCTTCTCA</td>
<td>NEDD4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr2:10008936-100089401</td>
<td>TGGAAAGTGGTCTT</td>
<td>AFF3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr9:38059922-38059937</td>
<td>TCGGAAGTGGTCTT</td>
<td>SHB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr12:83953473-83953488</td>
<td>GTGAATATTTTCTT</td>
<td>TSPAN19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr7:113841745-113841760</td>
<td>ATGCAATTGTTACTT</td>
<td>FOXP2</td>
<td></td>
</tr>
<tr>
<td>RSRFC4</td>
<td>chr3:135853336-135853352</td>
<td>CTTCTATTTTTAGGCC</td>
<td>KY</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr13:108079600-108079616</td>
<td>CTGGCATTATTTAAGCCT</td>
<td>MYO16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr12:47679884-47679900</td>
<td>GGGCGTATTATTAGCC</td>
<td>DDN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr14:44648956-44648972</td>
<td>CGGTAAAAATAGAAT</td>
<td>SNORD127</td>
<td></td>
</tr>
<tr>
<td></td>
<td>chr10:69094928-69094944</td>
<td>AGTGCATTTTTAATAAC</td>
<td>CTNNA3</td>
<td></td>
</tr>
<tr>
<td>Chr</td>
<td>Region</td>
<td>Motif</td>
<td>Gene</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----------------</td>
<td>---------------------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>chr5</td>
<td>158459865-158459877</td>
<td><strong>ACCAATCA</strong>&lt;sup&gt;CGCG&lt;/sup&gt;</td>
<td><strong>HIST1H1E</strong></td>
<td></td>
</tr>
<tr>
<td>chr6</td>
<td>116488683-116488698</td>
<td><strong>TTGCTATGATAAACC</strong></td>
<td><strong>FRK</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16472501-16472516</td>
<td><strong>AGGTAAATTATTGCCA</strong></td>
<td><strong>SOSTDC1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>147191470-147191485</td>
<td><strong>TGGTTATTGAGTACT</strong></td>
<td><strong>SPINK1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>242417676-242417691</td>
<td><strong>TGGTTACTCTTTAA</strong></td>
<td><strong>HNF4A</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4638930-4638945</td>
<td><strong>AGGTTAATTATTAATC</strong></td>
<td><strong>GLTPD2</strong></td>
<td></td>
</tr>
<tr>
<td>chr6</td>
<td>116488700-116488702</td>
<td><strong>TTGCTAATGAATAACC</strong></td>
<td><strong>DUOXA2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>116488704-116488706</td>
<td><strong>AGGTAAATTATTGCCA</strong></td>
<td><strong>MYH1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>116488710-116488712</td>
<td><strong>TGGTTATTGAGTACT</strong></td>
<td><strong>MMP20</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>147191470-147191485</td>
<td><strong>TGGTTACTCTTTAA</strong></td>
<td><strong>PNLIP</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>242417676-242417691</td>
<td><strong>AGGTTAATTATTAATC</strong></td>
<td><strong>GLIS1</strong></td>
<td></td>
</tr>
<tr>
<td>chr11</td>
<td>118741050-118741062</td>
<td><strong>TGCTATTTTAG</strong></td>
<td><strong>DKFZp566F</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>118741064-118741066</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>0947</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>118741068-118741070</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>C12orf39</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>118741072-118741074</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>FLJ41200</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>118741076-118741078</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>MIR614</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>118741080-118741082</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>MIR3666</strong></td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>188866532-188866551</td>
<td><strong>TGGCTATTTTAG</strong></td>
<td><strong>USP2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>28806287-28806306</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>RGS3</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>371715096-371715115</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>CPT1B</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72676896-72676915</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>KCNN1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72676917-72676929</td>
<td><strong>GCTAAAAATAGCC</strong></td>
<td><strong>SPEG</strong></td>
<td></td>
</tr>
<tr>
<td>chr11</td>
<td>27450934-27450953</td>
<td><strong>CTGGCTATATTC</strong></td>
<td><strong>USP2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>28806287-28806306</td>
<td><strong>CTTCAAATTTGCAATTG</strong></td>
<td><strong>LG4</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>371715096-371715115</td>
<td><strong>CTTCAAATTTGCAATTG</strong></td>
<td><strong>HORMAD2</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72676896-72676915</td>
<td><strong>CTTCAAATTTGCAATTG</strong></td>
<td><strong>FOX1</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>72676917-72676929</td>
<td><strong>CTTCAAATTTGCAATTG</strong></td>
<td><strong>MLXIPL</strong></td>
<td></td>
</tr>
</tbody>
</table>
**Supplementary Table S2:** Genes exhibiting cirDNA modification differences and steady-state mRNA and DNA modification differences in primary tumours.

<table>
<thead>
<tr>
<th>EntrezID</th>
<th>Gene Symbol</th>
<th>EntrezID</th>
<th>Gene Symbol</th>
<th>EntrezID</th>
<th>Gene Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>2346</td>
<td>FOLH1</td>
<td>32</td>
<td>ACACB</td>
<td>1271</td>
<td>CNTFR</td>
</tr>
<tr>
<td>4232</td>
<td>MEST</td>
<td>771</td>
<td>CA12</td>
<td>2346</td>
<td>FOLH1</td>
</tr>
<tr>
<td>4233</td>
<td>MET</td>
<td>1756</td>
<td>DMD</td>
<td>2626</td>
<td>GATA4</td>
</tr>
<tr>
<td>91851</td>
<td>CHRDL1</td>
<td>2346</td>
<td>FOLH1</td>
<td>2895</td>
<td>GRID2</td>
</tr>
<tr>
<td>126123</td>
<td>SCRL</td>
<td>2628</td>
<td>GATM</td>
<td>3205</td>
<td>HOXA9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3725</td>
<td>JUN</td>
<td>3394</td>
<td>IRF8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4212</td>
<td>MEIS2</td>
<td>4232</td>
<td>MEST</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4232</td>
<td>MEST</td>
<td>4233</td>
<td>MET</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4233</td>
<td>MET</td>
<td>5524</td>
<td>PPP2R4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4330</td>
<td>MN1</td>
<td>5582</td>
<td>PRKCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5638</td>
<td>PRRG1</td>
<td>6506</td>
<td>SLC1A2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9117</td>
<td>SEC22L3</td>
<td>7539</td>
<td>ZFP37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9515</td>
<td>STXBP5L</td>
<td>9354</td>
<td>UBE4A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9781</td>
<td>RNF144</td>
<td>25828</td>
<td>TXN2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10735</td>
<td>STAG2</td>
<td>29984</td>
<td>RHOD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>54433</td>
<td>NOLA1</td>
<td>51642</td>
<td>MRPL48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>55172</td>
<td>C14orf104</td>
<td>55615</td>
<td>PRR5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64326</td>
<td>RFWD2</td>
<td>60675</td>
<td>PROK2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80059</td>
<td>LRRTM4</td>
<td>90134</td>
<td>KCNH7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>90525</td>
<td>SHF</td>
<td>91851</td>
<td>CHRDL1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>91851</td>
<td>CHRDL1</td>
<td>126123</td>
<td>SCRL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>126123</td>
<td>SCRL</td>
<td>130367</td>
<td>SGPP2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>158158</td>
<td>RASEF</td>
<td>150248</td>
<td>C22orf15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>388228</td>
<td>SBK1</td>
<td>266553</td>
<td>OFCC1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>399694</td>
<td>SHC4</td>
<td>375567</td>
<td>UNQ739</td>
</tr>
</tbody>
</table>