Epigenetic markers of prostate cancer in plasma circulating DNA

Rene Cortese1, Andrew Kwan1, Emilie Lalonde2,3, Olga Bryzgunova4, Anna Bondar4, Ying Wu2, Juozas Gordevicius1,5, Mina Park1, Gabriel Oh1, Zachary Kaminsky1, Justina Tverkuviene6, Arvydas Laurinavicius7, Feliksas Jankevicius8, Dorota H.S. Sendorek2, Syed Haider2,9, Sun-Chong Wang10, Sonata Jarmalaite6, Pavel Laktionov4, Paul C. Boutros2,3,* and Arturas Petronis1,*

1The Krembil Family Epigenetics Laboratory, Centre for Addiction and Mental Health, 250 College St, Toronto, ON, Canada M5T 1R8, 2Informatics and Biocomputing Platform, Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 800, Toronto, ON, Canada M5G 0A3, 3Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada M5G 0A3 4Institute of Chemical Biology and Fundamental Medicine, SD RAS, Lavrentiev Ave. 8, Novosibirsk, Russia 630090, 5Vilnius University Institute of Mathematics and Informatics, Akademijos str. 4, LT-08663, Vilnius, Lithuania, 6Division of Human Genome Research Centre, Faculty of Natural Sciences, Vilnius University, Ciurlionio 21, Vilnius, Lithuania LT-08663, 7Vilnius University Faculty of Medicine and National Center of Pathology Affiliate of Vilnius University Hospital Santariskiu Clinics, P. Baublio 5, Vilnius LT-08406, Lithuania, 8Vilnius University Faculty of Medicine and Vilnius University Hospital Santariskiu Clinics, Santariskiu St 2, Vilnius 08661, Lithuania, 9Computer Laboratory, University of Cambridge, Cambridge, UK CB3 0FD and 10Systems Biology and Bioinformatics Institute, National Central University, Jhongli city, Taiwan.

Received March 13, 2012; Revised April 25, 2012; Accepted May 16, 2012

Epigenetic differences are a common feature of many diseases, including cancer, and disease-associated changes have even been detected in bodily fluids. DNA modification studies in circulating DNA (cirDNA) may lead to the development of specific non-invasive biomarkers. To test this hypothesis, we investigated cirDNA modifications in prostate cancer patients with locally confined disease (n = 19), patients with benign prostate hyperplasias (n = 20) and in men without any known prostate disease (n = 20). This initial discovery screen identified 39 disease-associated changes in cirDNA modification, and seven of these were validated using the sodium bisulfite-based mapping of modified cytosines in both the discovery cohort and an independent 38-patient validation cohort. In particular, we showed that the DNA modification of regions adjacent to the gene encoding ring finger protein 219 distinguished prostate cancer from benign hyperplasias with good sensitivity (61%) and specificity (71%). We also showed that repetitive sequences detected in this study were meaningful, as they indicated a highly statistically significant loss of DNA at the pericentromeric region of chromosome 10 in prostate cancer patients (p = 1.8 × 10^-6). Based on these strong univariate results, we applied machine-learning techniques to develop a multi-locus biomarker that correctly distinguished prostate cancer samples from unaffected controls with 72% accuracy. Lastly, we used systems biology techniques to integrate our data with publicly available DNA modification and transcriptomic data from primary prostate tumors, thereby prioritizing genes for further studies. These data suggest that cirDNA epigenomics are promising source for non-invasive biomarkers.
INTRODUCTION

Early diagnosis represents the most effective way to improve cancer prognosis (1). Accurate tests for early diagnosis and screening of asymptomatic populations result in better patient management and decreased mortality (2). Prostate cancer is the second-most common malignancy in men, with ~218,000 new cases and 32,000 thousand deaths in the USA in 2010 (3). Elevation of prostate-specific antigen (PSA) levels is the most common marker for the diagnosis of prostate cancer, but its clinical value has been intensively questioned (4–6). Increased PSA levels are also detected in benign conditions such as prostatitis and benign prostatic hyperplasia (BPH), leading to unnecessary prostate biopsies, which are expensive and have non-trivial morbidities (7). There is, therefore, a critical need for specific, sensitive and non-invasive biomarkers for prostate cancer diagnosis that could be used routinely in clinical laboratories.

New opportunities for cancer diagnosis and screening may arise from the identification of cancer-specific epigenetic DNA alterations. In mammalian cells, biochemical modifications occur at position 5 of cytosine residues: methylation (5-mC) and hydroxymethylation (5-hmC) (8). Recently, two new types of cytosine modifications, 5-formylcytosine and 5-carboxycytosine, were discovered but are very rare (9). In differentiated mammalian cells, modified cytosines are detected predominantly in cytosine-guanine (CpG) dinucleotides, the majority of which are clustered in CpG-rich areas typically related to transcription start sites. These CpG-rich regions are referred to as ‘CpG islands’ and are mainly unmethylated, while the remainder of the genome is mainly methylated (10,11). Conventional methods, such as bisulfite-sequencing and methylation-sensitive enzyme restriction, do not differentiate 5-mC from 5-hmC (12). We will use the term ‘DNA methylation’ as per primary publications; however, the more correct term ‘DNA modification’ will be applied in the description of our own experiments.

Epigenetic changes are among the most common molecular alterations in human neoplasias (13), making them good candidate biomarkers (14). Tumor-related changes in DNA methylation can even be detected in plasma, serum and other bodily fluids (15). The presence of circulating nucleic acids in plasma was first reported in 1948 (16), and it is now generally assumed that circulating DNA (cirDNA) can originate from both normal and tumor cells (17,18). Several mechanisms explain DNA shedding, including apoptosis, necrosis and release from macrophages/scavengers after the absorption of necrotic cells (18,19). In addition, DNA can be released via intracellular fusion between plasma membranes and multi-vesicular bodies, forming exosomes (20).

Studying cirDNA for epigenetic changes is a promising source of highly specific biomarkers. cirDNA can be extracted in a non-invasive manner and is enriched for tumor-derived DNA. Disease-specific epigenetic modifications would thus provide a secondary layer of sensitivity and specificity. A few previous studies have considered this approach, although only with small numbers of candidate genes. For example, hypermethylation at the gene encoding glutathione S-transferase p1 (GSTP1) has been detected in tissue samples and cirDNA of prostate cancer patients compared with controls (summarized in Supplementary Material, Table S1). A recent meta-analysis of 22 articles indicated good specificity (89%) but modest sensitivity (52%) (21) of GSTP1 differential methylation for prostate cancer screening. The analysis of cirDNA methylation has also been performed in other cancer types. For example, the hypermethylation of the gene encoding septin 9 (SEPT9) in cirDNA identified colorectal cancer patients from control individuals with 68% sensitivity and 89% specificity (22). Three independent studies detected the hypermethylation of the gene for cyclin-dependent kinase inhibitor 2A (CDKN2A, also known as p16INK4) in cirDNA from non-small cell lung cancer patients (23–25). Similar studies identified differential methylation in cirDNA from plasma or serum of patients with breast, ovarian and gastrointestinal cancers (26–28). Essentially, all previous studies have examined individual genes: to our knowledge, the largest studies of cirDNA methylation are 56-gene analyses of pancreatic (29), breast (30) and ovarian (31) cancers.

Here, we provide the first global analysis of cirDNA methylation in any tumor type. We focus on prostate cancer and compare cirDNA modifications in prostate cancer patients and two control groups using the differential methylation hybridization method (32). We replicate our microarray findings on the original sample set as well as on independent samples from prostate cancer patients and BPH control patients using bisulfite sequencing. We also used machine learning to identify an epigenetic signature for prostate cancer in cirDNA. Finally, we apply system biology techniques to integrate our results with DNA modification and mRNA abundance data from primary prostate tumors.

RESULTS

We first studied cirDNA methylomes 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 polymerase chain reaction (PCR; Supplementary Material, Fig. S1). Because tumor cirDNA is expected to be degraded and shorter than DNA fragments released from white blood cells (WBCs) ruptured during sample storage and processing (33), we selected PCR conditions which favor the amplification of shorter (<1.5 kbp) cirDNA fragments (Supplementary Material, Fig. S2). Enriched DNA fractions were interrogated on microarrays containing 12 192 GC-rich clones covering 5347 unique sequences and 1245 repetitive elements (34). Quality assessments indicated that all arrays were robust: before normalization, signals in both channels showed similar distributions for all arrays and the distribution of the inter-replicate variability was unimodal with a peak at zero, indicating that no systematic bias was introduced by technical variation. The distribution of the variation attributed to batch effects is unimodal and symmetrical around the origin, indicating no systematic bias (Supplementary Material, Fig. S3). All arrays were thus used in subsequent analyses.

Differentially modified genes in cirDNA of prostate cancer

We compared the cirDNA modification profiles of prostate cancer and BPH patients in sample set 1, but did not find
any differentially modified loci between these two groups (Supplementary Material, Table S2). While these two groups were age-matched (68.7 ± 6.4 years for BPH and 68.5 ± 9.2 years for cancer patients), they represent a complex comparison of malignant versus pre-malignant patient populations. We therefore chose to compare the cirDNA profiles of prostate cancer against younger controls without any evidence for prostate disease (age 46.3 ± 6.5 years) to generate a larger biological contrast for candidate identification. We performed two separate analyses to identify disease- and age-dependent DNA modification changes (Fig. 1). At \( q < 0.05 \), we identified 117 differentially modified regions between prostate cancer and asymptomatic controls. Of these, 39 loci were age-independent (Supplementary Material, Table S3), 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 (KIAAI539, PCDH1, RNF219 and NUCCD3) which showed statistically significant differences \( (q < 0.05) \) and (ii) three genes (HPS2, GNG7 and DLG) which showed marginal statistical significance \( (0.05 < q < 0.08) \) and were previously reported to be associated with cancer (Supplementary Material, Table S4). All selected genes were more closely associated with disease than with age (Supplementary Material, Table S2).

All bisulfite-treated DNA pyrosequencing assays contained at least one restriction site for the enzymes used in the enrichment of modified DNA and contained different numbers of CpGs. Table 1 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), two loci (KIAAI539 and RNF219) showed the significant disease-dependent \( (P < 0.05) \) and age-independent \( (P > 0.05) \) differences. Since RNF219 exhibited the largest and significant differences in both training and validation cohorts (Table 1), we evaluated its ability to predict whether a sample was derived from an individual with prostate cancer from controls with no history of prostate disease and BPH patients using receiving operating curve (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 the 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 \( (AUC = 0.56, \text{Fig. } 2) \).

![Figure 1. Differentially modified regions in cirDNA in prostate cancer and control individuals. (A) 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 log2 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. (B) 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 (-7) to blue (7).

### 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,
### Table 1. Summary of pyrosequencing results

<table>
<thead>
<tr>
<th>Loci</th>
<th>Sample set 1</th>
<th>Disease pLMe&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Age pLMe&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Sample set 2</th>
<th>Disease pLMe&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Age pLMe&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Mod. Pca&lt;sup&gt;a&lt;/sup&gt;</td>
<td>% Mod. Ctrl&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td>% Mod. Pca&lt;sup&gt;a&lt;/sup&gt;</td>
<td>% Mod. Ctrl&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td><strong>DILG2</strong></td>
<td>48.7 ± 22.0</td>
<td>45.6 ± 23.6</td>
<td>0.02</td>
<td>42.9 ± 14.7</td>
<td>41.9 ± 18.3</td>
<td>0.85</td>
</tr>
<tr>
<td><strong>GNG7</strong></td>
<td>38.4 ± 3.0</td>
<td>38.2 ± 2.4</td>
<td>0.74</td>
<td>37.9 ± 11.1</td>
<td>36.8 ± 7.8</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>HPSE2</strong></td>
<td>22.7 ± 13.7</td>
<td>16.1 ± 8.3</td>
<td>0.13</td>
<td>40.6 ± 30.0</td>
<td>29.6 ± 27.1</td>
<td>0.32</td>
</tr>
<tr>
<td><strong>KIAA1539</strong></td>
<td>38.8 ± 5.8</td>
<td>37.1 ± 10.9</td>
<td>1 × 10&lt;sup&gt;-3&lt;/sup&gt;</td>
<td>30.8 ± 10.0</td>
<td>36.2 ± 6.4</td>
<td>1 × 10&lt;sup&gt;-4&lt;/sup&gt;</td>
</tr>
<tr>
<td><strong>NUCD3</strong></td>
<td>59.4 ± 6.1</td>
<td>63.6 ± 9.0</td>
<td>0.09</td>
<td>67.6 ± 11.6</td>
<td>69.7 ± 11.1</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>PCDHB1</strong></td>
<td>66.6 ± 23.3</td>
<td>50.9 ± 36.2</td>
<td>0.51</td>
<td>58.4 ± 27.6</td>
<td>58.0 ± 25.0</td>
<td>0.97</td>
</tr>
<tr>
<td><strong>RNF219</strong></td>
<td>35.0 ± 18.1</td>
<td>55.0 ± 17.6</td>
<td>3 × 10&lt;sup&gt;-4&lt;/sup&gt;</td>
<td>35.3 ± 31.8</td>
<td>46.2 ± 23.7</td>
<td>1 × 10&lt;sup&gt;-4&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Statistically significant differences (p < 0.05) are in bold.

<sup>a</sup>Mean cirDNA modification per amplicon.

<sup>b</sup>LME model p-value.

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). The amount of repetitive DNA template at the pericentromeric region of chromosome 10 in prostate cancer samples was similar to the control samples (Ct = 28.3 ± 6.4 and 26.8 ± 3.6, respectively; two-tailed homoscedastic Student’s t-test, p = 0.50). However, we observed heterogeneity in Ct values across samples. Stage III prostate cancer samples showed a significant loss of pericentrometric DNA compared with the stage II and control samples (Ct 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.

**Machine-learning analysis of microarray-based cirDNA modification profiles**

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. S4) 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%; Fig. 4A). The percentage of trees that voted to classify each sample as prostate cancer was assessed as a measure of intra- and inter-group
heterogeneity (Fig. 4B) and did not indicate a significant effect of tumor stage in the classification accuracy (Fisher's exact test; \( P = 1 \)).

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 (Supplementary Material, Table S6). We evaluated these genes in a series of mRNA abundance (35–37) and DNA modification data sets (38) from primary prostate tumors and healthy prostatic tissues (Fig. 5A). 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 (39,40).

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) (38). 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 S7). 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 cirDNA modification are associated with

Figure 3. 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, as detailed in Supplementary Material, Table S2, are shown in the y-axis.
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 (Fig. 5B 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 (Fig. 6).

DISCUSSION

Many efforts have been directed toward the discovery of biomarkers for the diagnosis, prognosis and treatment-planning of cancer (41). The biomarker potential for differentially modified loci in cirDNA isolated from plasma or other bodily fluids has been reported previously (42,43) based on limited sample sizes and pre-specified candidate genes (22–25,44–50). To our knowledge, this study comprises the first epigenome-wide DNA modification profiling of plasma or other bodily fluids from cancer patients.

We first attempted to discriminate prostate cancer patients from BPH controls, 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: our bisulfite sequencing analyses effectively differentiated prostate cancer and BPH. In addition, the two diseases may share etiopathogenic factors (51), and some studies (but not all) (52) showed evidence, indicating that BPH might constitute a precursor for the development of prostate cancer (53,54). Given the lack of differences with BPH in the studied sample set, we compared cirDNA modification profiles of prostate cancer 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 (55), we selected a group of younger, asymptomatic controls. To account for age-dependent epigenetic changes (56,57), we used linear models to identify differences attributable to disease alone, and not to age.

We identified 18 non-repetitive genomic regions associated with disease but not age. We validated six of these using much more sensitive bisulfite sequencing analyses and exploited a second independent validation cohort (sample set 2) to confirm their disease association. To validate these data, 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. Analysis of differentially modified regions by bisulfite-pyrosequencing exhibited the largest differences between prostate cancer patients and controls at RNF219 on chromosome 13q31.1. The predictive accuracy for plasma cirDNA modification at this locus (71% specificity and 61% sensitivity) is comparable with the existing biomarkers for prostate cancer (21). Specificity was much higher than that in PSA (20%) (58), but slightly lower than that for differential modification in GSTP1 (89%) (21). In turn, sensitivity was lower than that in PSA levels (80%) (58), but higher than differential modification in GSTP1 (52%) (21). Further validation is needed to determine the utility of RNF219 as a diagnostic biomarker for prostate cancer; nevertheless, these data spurred us to combine multiple epigenomically altered loci into a single diagnostic signature. Using a machine-learning technique called Random Forests, we were able to distinguish cirDNA derived from prostate cancer patients from that derived from controls with 72% accuracy. We anticipate that this performance would be further improved by consideration of larger data sets and more homogeneous patient populations (i.e. similar tumor stage and Gleason scores). Although some earlier studies applied machine-learning techniques to study malignant tissue samples (59–62), this is the

Figure 4. 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).
Figure 5. 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/Nmodif.), significant fold-change in DNA modification from primary prostate tumors (Prostate T/Nmodif.) and significant fold change in mRNA abundance in primary prostate tumors (Prostate T/NmRNA). The total number of genes identified as differentially modified in cirDNA did not overlap significantly with any individual public data set (hypergeometric distribution; \( p > 0.05 \); Supplementary Material, Table S8). (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/Nmodif.), prognostic relevance based on DNA modification differences (HRmodif.), differential mRNA abundance between tumor and control samples (T/NmRNA) and prognostic relevance on mRNA abundance (HRmRNA). 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 \( \geq 0.5 \)).
first time, to our knowledge, that machine-learning techniques have been applied in a study of cirDNA modification. In the second group of analyses, we detected significantly weaker microarray signals in repetitive pericentromeric elements at 10q11.21 that could be attributed to cancer-related DNA copy number variations. Genetic alterations in chromosome 10 have been widely studied in prostate cancer, with reported deletions (63), allelic imbalances (64), loss of heterozygosity (65) and non-coding single-nucleotide polymorphisms associated with increased prostate cancer risk (66). However, only allelic imbalances in a few tumor suppressor genes were reported in cirDNA of prostate cancer patients (67). To our knowledge, this is the first report of copy number variation in repetitive elements in cirDNA of prostate cancer patients.

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 prostate cancer compared with normal controls in various molecular contexts. These genes have the potential to serve as prognostic markers for prostate cancer, although deeper investigations are required to define whether differential expression or epigenetic modification of these genes is involved in prostate cancer tumorigenesis. Interestingly, a decrease in the DNA modification of ARHGAP6 appears to provide a protective advantage by preventing patient recurrence and, in an independent patient cohort, an increase in mRNA abundance is associated with greater overall patient survival (Fig. 6). On the other hand, we do not see this inverse relationship between DNA modification and mRNA abundance in CDK6 where an increase in both mRNA and DNA modification is associated with better patient prognosis (Fig. 6).

Future cirDNA analyses will also need to consider distinguishing 5-mC and 5-hmC profiles. Most of the earlier cancer epigenomic studies used techniques such as DNA modification sensitive restriction enzymes and bisulfite sequencing (12), which do not discern between the two types of DNA modification: little is known concerning the distribution of 5-hmC in cancer cells. Recently, several methods have been developed for the detection of 5-hmC using DNA glycosylation and modification-insensitive restriction enzymes (68,69). The combination of these methods with the one presented here will provide a comprehensive 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 (70).

In summary, our study of cirDNA in prostate cancer provides a basis for further efforts in the detection of diagnostic markers in free floating DNA. The comparison of different strata of prostate cancer patients (i.e. indolent and aggressive cancer) will enable the identification of prognostic biomarkers to predict tumor aggressiveness. Such experiments should be performed using larger samples and more informative platforms, such as tiling microarrays and next-generation sequencing, which would more accurately assess DNA modification profiles. Our 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.

MATERIALS AND METHODS

Samples

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). An 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). Blood samples were collected and the plasma fraction separated by centrifugation and frozen at −20°C until DNA isolation. cirDNA was isolated from 1 mL of plasma using glass columns (sample set 1; Vivantis, Selangor, Malaysia) and QiAamp DNA Blood Mini Kit (sample set 2; Qiagen, Hilden, Germany), according to the manufacturers’ instructions.

**Blood reference pool**

DNA samples extracted from WBCs of 20 individuals unrelated to this project were pooled and used as reference. Participants were recruited at CAMH and provided informed consent. Isolated DNA was pooled and sheared by sonication to 200–500 bp fragments.

**Enrichment of the modified cirDNA fraction**

DNA modification profiles of plasma cirDNA were produced using the method shown in Supplementary Material, Figure S1. Briefly, 50 ng of cirDNA and sonicated DNA from pooled WBC genomic DNA (reference sample) were blunted using T4 DNA polymerase (NEB, Pickering, ON, Canada) and universal adaptors were ligated by overnight incubation at 4°C with T4 ligase (NEB). Blunt universal adaptors were prepared by annealing two oligonucleotides (oligo 1: GCGGTGACCCGGGAGATCTGAATTC and oligo 2: GAATTCCAGATC). Next, adaptor-ligated DNA was digested using a cocktail of DNA modification-sensitive restriction enzymes (HpaII, HinP1II and HpyCH4IV; NEB), which do not cut when the corresponding restriction sites contain modified nucleotides, 5-meC or 5-hmC. Therefore, the subsequent PCR, with primers complementary to the adaptors, amplifies only the fragments bearing such modifications. The digestion product was amplified in 25 μL of final volume under the following conditions: 1× PCR buffer (Sigma-Aldrich, Oakville, ON, Canada), 2.875 mM MgCl2 (Sigma-Aldrich), 275 μM amionoylly 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. Amplification was checked by electrophoresis in agarose gels. PCR products were purified using the MinElute kit (Qiagen, Mississauga, ON, Canada) and quantified using a Nanodrop 2000.

**Microarray hybridization and data analysis**

For each sample, 1.5 μg of purified enriched modified cirDNA was labeled using Cy3 (GE Healthcare, Baie d’Urfé, QC, Canada) and blood DNA reference pool was labeled using Cy5 (GE Healthcare) and co-hybridized to the HCG112k microarrays (UHN, Toronto, ON, Canada) (34) containing 12 192 CpG rich clones, as described previously (71). For all patients and controls, two technical replicates were tested.

Hybridized microarrays were scanned using the Axon 4000B scanner and signals were processed using the GenePix Pro software (v6.1.0.4). Raw data underwent extensive quality control, followed by pre-processing and statistical analysis as described previously (72). Briefly, microarray data were pre-processed using a modified version of the variance stabilizing normalization method, using the vsn package (v3.20.0) (73). Pre-processed data were then analyzed with spot-wise linear-model fitting followed by an empirical Bayes moderation of the standard error (74). A false-discovery rate (FDR) adjustment for multiple-testing was used and q-values (FDR-adjusted p-values) were assigned (75). Per-probe microarray signals relative to the reference pool (called M-values) were calculated by subtracting (in log-space) 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). We used the limma package (v3.8.3) for the R statistical environment (2.13.2) (74) 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 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, where aligned with the Homo sapiens LATESTGP database and the search sensitivity was set to ‘Near-exact matches’.

**Real-time PCR analysis of chromosome 10 pericentromeric repeats region**

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 methylation 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.
controls were assessed using an
The statistical significance of differences between cases and
random factor, adjusting for the variance across individuals.
within each amplicon. Patients and controls were set as the
constituted the clustering variable, adjusting for variance
fixed effects. cirDNA modification values per CpG position
seven loci was analyzed separately, with age and disease as
in the candidate loci, we used an LME model (76). Each of the
Table S9. To evaluate age- and disease-specific modifications
Mississauga). Sequences of primers used for nested PCR amp-
turer. DNA modification values at single CpG positions were
applied to the real-time PCR-based estimation of
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. Fifteen nanograms of purified amplicons
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 Tech-
nologies, Carlsbad, CA, USA) and 200 nM of specific primers
(GAATGGAACGGCAACGAA and CGATGTCACTCC-
TACCTTCTTCT). The PCR program started with a Taq poly-
merase 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. Amplifica-
tion and melting curve analyses were performed using the Step
One Software version 2.1 (Life Technologies).

Fine mapping of modified cytosines
cirDNA samples were bisulfite-treated using the Epitect kit
(Qiagen, Mississauga, ON, Canada) and amplified using the
whole bisulfirome amplification kit (Qiagen, Mississauga),
according to the manufacturer’s instructions. The loci of inter-
est were further amplified using nested PCR. Bisulfite treated
and amplified cirDNA (10 ng) was subjected to locus-specific
amplification using 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 poly-
merase (Qiagen, Mississauga). PCR started with a Taq poly-
merase 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 45 s, 72°C for 1 min and a final ex-
tension step at 72°C for 10 min. PCR products were analyzed
by pyrosequencing using a PyroMark Q24 (Qiagen, Missis-
sauga) following the protocols recommended by the manufac-
turer. DNA modification values at single CpG positions were
assessed using the PyroMark Q24 1.0.10 software (Qiagen,
Mississauga). Sequences of primers used for nested PCR am-
plification and pyrosequencing are in Supplementary Material,
Table S9. To evaluate age- and disease-specific modifications
in the candidate loci, we used an LME model (76). Each of the
seven loci was analyzed separately, with age and disease as
fixed effects. cirDNA modification values per CpG position
constituted the clustering variable, adjusting for variance
within each amplicon. Patients and controls were set as the
random factor, adjusting for the variance across individuals.
The statistical significance of differences between cases and
controls were assessed using an F-test.

Univariate classification analysis of ring finger protein 219
Mean cirDNA modifications of all CG 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. An
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). AUCs
were calculated by summing the areas of the rectangles
defined by each event (i.e. cancer versus not cancer).

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 param-
eterization (77,78). We used a standard doubly nested cross-
validation approach to ensure full separation of feature selection
and classifier evaluation (Supplementary Material, Fig. S4).
Briefly, an inner LOOCV loop was used for feature-size selec-
tion and an outer one for model evaluation. Thus, in the inner
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 deter-
mine 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 ac-
curacy.

Primary prostate cancer tissue DNA modification analysis
Previously published DNA modification data from primary
prostate tumors and non-cancerous prostate tissue from pro-
tate 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) (38). ComBat (79) nor-
malized 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 inten-
sities 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).

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 (35,36). In total 150 primary tumor samples were available for survival analysis, an additional study (37) 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 (80) (R packages huex10stv2hsentrezgcf v14.0.0 and hgfocushsentrezgcf v14.0.0). Pre-processing used the robust multi-array algorithm (81). 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).

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).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

Conflict of Interest statement. The authors report no conflict of interest.

FUNDING

This work was supported, in part, by the Krembil Foundation grant to A.P., by the Ontario Institute for Cancer Research to P.C.B. through funding provided by the Government of Ontario, by a CIHR Doctoral Fellowship to E.L. and by an Oncology Research and Methods Training Program Fellowship to Y.W. A.P. is the Tapscott Chair, University of Toronto.

REFERENCES


