Loci-specific differences in blood DNA methylation for early detection of hepatocellular carcinoma in populations at risk

Katarzyna Lubecka¹, Kirsty Flower², Megan Beetch¹, Lucinda Kurzava¹, Hannah Buvala¹, Samer Gawrieh³, Suthat Liangpunsakul³, Tracy Gonzalez⁴, George McCabe⁴, Naga Chalasani³, James M Flanagan², and Barbara Stefanska¹,5*

¹ Department of Nutrition Science, Purdue University, West Lafayette, IN, USA
² Epigenetics Unit, Department of Surgery and Cancer, Imperial College London, London, UK
³ Division of Gastroenterology and Hepatology, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA
⁴ Department of Statistics, Purdue University, West Lafayette, IN, USA
⁵ Purdue University Center for Cancer Research, Purdue University, West Lafayette, IN, USA

Running title: Biomarkers for HCC early detection

Keywords: DNA methylation, HCC, biomarker, early detection

The word count: 5807

The Authors declare no conflict of interest.

*To whom correspondence should be addressed:

Dr. Barbara Stefanska
Assistant Professor
Department of Nutrition Science
Purdue University
201 S. University Street, Hansen Bldg, Room 109
West Lafayette, IN, 47907
Phone: 765 494 4401, Fax: 765 494 0906
bstefanska@purdue.edu
www.hhs.purdue.edu/nutr

Translational Relevance:
Poor prognosis and high mortality rate in hepatocellular carcinoma (HCC), the most prevalent primary liver cancer type, result from late diagnosis due to late onset of symptoms and lack of reliable screening tests. It is therefore crucial to identify tools that will improve HCC early detection and surveillance for patients at risk. Our study demonstrates a workflow for the discovery of early detection
HCC biomarkers. We determined for the first time a set of specific alterations in DNA methylation that could indicate by a simple blood test the presence of HCC before the onset of symptoms. Using blood as a minimally invasive material and pyrosequencing as a straightforward quantitative method, the established candidate biomarker panel has high potential to be developed into a routine clinical test after validation in larger cohorts. Such a tool will accelerate progress aiming to decrease HCC mortality because HCC is treatable when diagnosed early.

Authors’ contributions:

BS designed the entire study, supervised various aspects of the project, wrote and corrected the manuscript. KL, MB, LK, and HB performed experiments and analyzed the data. KL, KF, and JMF performed bioinformatics on the genome-wide data. KL, TG, and GM performed statistical analyses on the biomarker panel. SG, SL, and NC oversaw clinical aspects of the study, sample selection and matching. All the authors had full access to the final version of the report and agreed to the submission.
ABSTRACT

Purpose: Late onset of clinical symptoms in hepatocellular carcinoma (HCC) results in late diagnosis and poor disease outcome. Approximately 85% of individuals with HCC have underlying liver cirrhosis. However, not all cirrhotic patients develop cancer. Reliable early detection biomarkers that can distinguish cirrhotic patients who will develop cancer from those who will not are urgently needed.

Experimental Design: We used Illumina-450K microarray to test whether white blood cell DNA, an easily accessible source of DNA, exhibits site-specific changes in DNA methylation in prospectively collected blood specimens of HCC cases who were cancer-free at blood collection (pre-diagnostic). Top 11 differentially methylated loci were selected for validation by pyrosequencing. Diagnostic potential of the 11 pyrosequenced probes was tested in blood samples from a prospective cirrhotic cohort.

Results: We established 5 probes that distinguished cirrhotic patients who subsequently developed HCC from those who stayed cancer free (cirrhotic controls), demonstrating potential as biomarkers of early detection in populations at risk. These probes were identified within regulatory regions of BARD1, FXYD6, BRUNOL5, MAGEB3, and TET1. Using additive logistic regression model, we further showed that 9 CpG loci within those probes constitute a panel with high diagnostic accuracy (AUC=0.89; 95% CI:0.80-0.98). The panel was able to distinguish pre-diagnostic cases from cirrhotic controls free of cancer with 88% sensitivity at 70% specificity.

Conclusion: Using blood as a minimally invasive material and pyrosequencing as a straightforward quantitative method, the biomarker panel has high potential to be developed into a routine clinical test after validation in larger cohorts.
INTRODUCTION

Hepatocellular carcinoma (HCC), one of the most prevalent primary liver cancers, is the 4th most common cancer in the world and the 2nd leading cause of cancer death in the US (Globocan 2012) (1). In contrast to the stable or declining trends for most cancers, incidence rates for HCC increased more than any other cancer in the last two decades (1). Much of this increase is likely due to the rising incidence of chronic hepatitis C and B infections (HCV and HBV), obesity and type 2 diabetes, all being major risk factors for cirrhosis and subsequent HCC (2). Due to late onset of clinical symptoms and lack of reliable screening tests, HCC is diagnosed late and is associated with poor prognosis and high mortality rate (2).

The current recommended method for HCC surveillance is a combination of serum alpha-fetoprotein (AFP) and a liver ultrasound every 6 to 12 months (3). However, using AFP as a screening test can miss up to 40% of patients since a significant proportion of HCC patients do not have elevated AFP (4). In addition, elevated AFP levels are only detectable after HCC has already developed, which excludes AFP as an early detection marker. Using ultrasound to distinguish malignant from benign lesions is also problematic due to the nodularity of cirrhotic livers (4). It is therefore critical to identify new effective tools that will improve the early detection of liver cancer and the surveillance for patients at risk. It is estimated that early HCC detection would increase the cure rate from 5% to 80%, which further strengthens the importance of identifying reliable HCC biomarkers (2).

Etiological factors for HCC include numerous environmental exposures, such as chronic infection with HBV and HCV, alcohol abuse, consumption of aflatoxin contaminated food, and a diet deficient in methyl donors (5,6). These risk factors clearly suggest the involvement of epigenetic components in HCC pathogenesis, which has been demonstrated in several studies by us and others (5,7-9). Epigenetics refers to alterations in gene expression without changes in the underlying DNA sequence and consists of three main components: DNA methylation, histone modifications, and non-coding RNA mechanisms. DNA methylation is considered to be the gate-keeper of gene expression providing stable long-term regulation (10). Aberrations in DNA methylation patterns, including increase (hypermethylation) in certain regions and decrease (hypomethylation) in others, have been linked to cancer initiation, progression and metastasis (8,10). Hypermethylation of tumor suppressor genes linked to transcriptional silencing, and recently reported promoter hypomethylation linked to activation of oncogenes and pro-metastatic genes have been described in tumor tissues from nearly all types of
cancer including HCC (8,11-13). Ethanol intake, for instance, decreases levels of a ubiquitous methyl donor S-adenosyl-methionine (SAM) in hepatic cells, which subsequently results in global DNA hypomethylation of the liver tissue, and leads to liver cirrhosis and HCC (5,7). On the other hand, supplementation with SAM can prevent HCC in rodents (14). These studies are re-enforced by our previous report where, using genome-wide high throughput technology, we discovered thousands of gene promoters differentially methylated in HCC tumors compared with normal adjacent tissue (8). Interestingly, recent discoveries suggest that aberrations in DNA methylation occur not only in target tissues where tumors develop but also in other tissues such as blood (15,16). An excellent example is exposure to aflatoxin B1, one of the etiological factors of HCC, that is associated with changes in DNA methylation patterns of white blood cells (15).

Epigenome-wide association studies provide increasing evidence that changes in DNA methylation are primary events in cancer and therefore may be an excellent tool for identifying reliable early disease biomarkers (8,9). Clinically relevant early detection biomarkers must be detected by minimally invasive methods in readily accessible tissues such as blood. There is precedent in support of this approach using blood in other human cancers, including colorectal, ovarian, and breast cancers (17-19), where serum was tested. In these studies, DNA methylation detected using DNA isolated from blood serum corresponds to circulating tumor DNA. While testing circulating tumor DNA sounds logical, low levels of such DNA in serum, especially at early phases of the disease (20), decrease the relevance of serum-based tests for early detection in cancer. Interestingly, recent studies in colorectal, breast and bladder cancers show that DNA methylation patterns in white blood cells (WBC) might carry reliable biomarkers of early detection (21-24). In HCC, studies on blood biomarkers have been limited so far to candidate genes without genome-wide investigations and to blood serum looking at circulating tumor DNA (19,25). Only one study addressed methylation of one candidate gene in WBC (26). Most importantly, all these studies were performed on samples from already diagnosed patients, often at advanced stages of HCC, which impedes translation of these markers into clinical use for early detection.

To establish clinically relevant biomarkers for HCC early detection and elaborate on their diagnostic accuracy in populations at risk, specifically in individuals with liver cirrhosis, we used blood DNA collected from individuals with HCC at the time when they were clinically considered cancer free. Upon delineating genome-wide DNA methylation landscape in blood DNA from HCC cases prior to
conventional diagnosis (pre-diagnostic cases) vs. matched healthy controls, we selected several CpG loci for validation of the difference in DNA methylation using pyrosequencing (Figure 1). Validated probes were verified in a cirrhotic population and probes whose methylation level discriminated pre-diagnostic cases from cirrhotic controls were established as candidate biomarkers. Using logistic regression, we built a predictive panel consisting of 3-9 CpG sites, located within the candidate biomarkers, which had high diagnostic accuracy (AUC > 0.872). Our present study establishes for the first time differences in DNA methylation at specific CpG sites that are detectable in blood of individuals prior to conventional diagnosis with HCC as compared with healthy or cirrhotic individuals free of cancer. Such epigenetic biomarkers could be a breakthrough in managing liver disease in populations at risk after validation in a larger cohort.

MATERIALS AND METHODS

Patients and blood DNA samples

Blood samples from hepatocellular carcinoma (HCC) patients and matched controls were collected by the Indiana Biobank and by Dr. Naga Chalasani group (cirrhotic Study Group, liver clinic at Indiana University). Please see Supplementary Materials and Methods for details on sample processing. Samples were obtained from three study groups: Study Group #1) blood samples collected from 21 HCC cases up to 4 years prior to conventional diagnosis and 21 matched healthy controls (Supplementary Table S1); Study Group #2) blood samples collected from 13 HCC cases with underlying cirrhosis up to 2 years prior to conventional diagnosis and matched cirrhotic controls without HCC (Supplementary Table S2); Study Group #3) blood samples collected from 24 HCC cases after conventional diagnosis and 24 matched healthy controls (Supplementary Table S3). Cases were matched with controls on gender, age, ethnicity, hepatitis C infection, and diabetes. The presence or absence of HCC in our study was determined based on the AASLD criteria (3). All patients provided written informed consent, and the Ethics Committee from concerned institutions approved the study.

Illumina Infinium Human Methylation 450K BeadChip microarray

Genomic DNA from blood samples in Study Group #1 was processed for genome-wide DNA methylation analysis using Infinium HumanMethylation 450K BeadChip as described previously (please see Supplementary Materials and Methods for details) (27). The methylation score for each
CpG was represented as a beta value according to the fluorescent intensity ratio with any values between 0 (unmethylated) and 1 (completely methylated). Raw microarray data and processed data are available from Gene Expression Omnibus (accession number pending).

**Bisulfite treatment of DNA and pyrosequencing**
For validation of candidate genes selected based on the microarray data, DNA bisulfite conversion was performed as previously described (please see Supplementary Materials and Methods for details) (8,27). Specific bisulfite converted gene sequences were amplified with HotStar Taq DNA polymerase (Qiagen) using biotinylated primers listed in Supplementary Table S4. Please see Supplementary Materials and Methods for details. The biotinylated DNA strands were pyrosequenced in the PyroMarkTMQ24 instrument (Biotage, Qiagen) as previously described (27). Data were analyzed using PyroMarkTMQ24 software.

**Statistical analysis**
**Microarray data**
DNA was isolated from frozen whole blood (WBC DNA and potentially circulating tumor DNA) and bioinformatics analyses excluded marginal changes (less than 1%) that could account for circulating tumor DNA. Hence, the presented differences correspond to changes in DNA methylation of WBC population. For convenience, we refer to detected changes as “differences in blood DNA methylation”. Raw methylation data from Human Methylation 450K microarrays were pre-processed using GenomeStudio and IMA (Illumina Methylation Analyzer for 450K, R/Bioconductor). White blood cell type adjustment was performed using previously published data (28-30). Please see Supplementary Materials and Methods for details on the analyses. The median beta value across all cases or all controls was calculated for each CpG and a paired Wilcoxon test was used to assess differences between cases and controls (delta beta=differential methylation=median case–median control). A methylation difference with \( P<0.05 \) was considered statistically significant. Although false discovery rate (FDR) correction of the \( P \) values was not successful due to limited sample sizes, the selected differences were successfully validated by an independent quantitative method, namely pyrosequencing, and verified in independent Study Groups. Among statistically significant differences, we took into account only those that are variable in the population yet stable over time as identified by intraclass correlation coefficient (ICC)>0.50 based on previously published data (please see Supplementary Materials and Methods for details) (31).
Pyrosequencing and biomarker panel

The statistical analyses were performed using SAS 9.4. Non-parametric Mann–Whitney U-test was used to compare the difference in DNA methylation at every tested CpG site in the selected genes (probes) determined by pyrosequencing between the case group and the control group in all the Study Groups. The results were considered statistically significant when \( P < 0.05 \). Receiver Operating Characteristic (ROC) curve analysis was performed using the Proc Logistic function in SAS 9.4 and verified using the ROC Curve analysis function in SPSS 23. Area under the ROC curve (AUC) was used to evaluate the diagnostic performance of the selected biomarker panel in our Study Groups. Please see Supplementary Materials and Methods for details.

RESULTS

The landscape of DNA methylation in blood samples collected from HCC patients prior to conventional diagnosis as compared to matched healthy controls

Previous data generated in blood provide increasing evidence that white blood cells (WBC) have different DNA methylation patterns specific to cancer patients compared with healthy individuals without cancer (21-24,26). To investigate whether these differential patterns are present at very early stages prior to diagnosis with conventional tools (pre-diagnostic stage) and thus could serve as early detection biomarkers, we delineated, using Illumina 450K microarray, genome-wide profiles of DNA methylation in blood collected from patients (cases) at the time when they were clinically considered cancer free (Figure 2, Study Group #1). Twenty one cases, who provided blood samples up to 4 years prior to diagnosis, were matched with 21 healthy individuals (control group), enrolled by the Indiana Biobank, as described in Materials and Methods. Please see Supplementary Materials and Methods for microarray characterization.

We identified 971 differentially methylated CpG sites in pre-diagnostic HCC cases as compared with healthy controls \( (P < 0.05, \text{paired Wilcoxon test, ICC} \geq 0.5) \) (Figure 2A). Difference in methylation expressed as median case–median control (delta beta=differential methylation) was higher than 0.1 in 20% of identified changes (Figure 2B). Chromosomal views of these differences were plotted using the IGV (Integrative Genomics Viewer) visualization tool (Figure 2C). Each vertical bar corresponds to a single differentially methylated CpG site, with blue indicating hypomethylation and red indicating
hypermethylation in case vs. control. Nearly 76% of differentially methylated CpG sites showed lower levels of methylation in cases vs. controls ($P = 2.973 \times 10^{-11}$, Wilcoxon test) (Figure 2A-C). This could suggest stronger regulatory role of hypomethylation in gene transcription in blood cells in individuals developing cancer which is consistent with previous findings (32). Indeed, classification of the CpG sites according to their location relative to CpG islands and transcription start site revealed that those hypomethylated loci are located in regulatory regions important for gene transcription such as CpG island shores, promoters, and 5’UTR at higher frequency than hypermethylated sites (Figure 2D). Furthermore, the bioinformatics analysis using GO, KEGG and DAVID knowledgebase indicate that hypomethylated CpG sites are located in genes associated with functions that are essential for gene transcription, cell adhesion, cell migration, and regulation of signal transduction pathways, including MAPK, WNT, and TLR (Figure 2E). It suggests that these hypomethylated genes might be important for immunological response of white blood cells to cancer at very early stages of the disease. Addressing this intriguing hypothesis requires future extensive studies.

Among 735 CpG sites hypomethylated in cases vs. controls, 482 sites were assigned to gene coding regions whereas 236 hypermethylated sites corresponded to 160 genes. The most robust changes with delta beta (diff.) higher than 0.30 were identified within 5’UTR of ACAN, CpG island of ATHL1, and CpG island shelf of RPTOR (Figure 2F). These differences were located within probes that show a larger variation between individuals than within individuals as indicated by ICC ≥ 0.93 based on previously published data (31). Indeed, they were consistent across the majority of matched pairs case-control in our Study Group.

**Identification of candidate differentially methylated regions in HCC for biomarker discovery using pyrosequencing**

The next step in our workflow of biomarker discovery was to identify CpG loci among differentially methylated sites found on the microarray which would have high potential to predict HCC in the studied population (Figure 1). We therefore applied several criteria to increase the probability that the CpG loci are ubiquitously differentially methylated in blood of HCC cases prior to diagnosis vs. controls. First, the magnitude of the difference in DNA methylation was ≥0.1 with $P<0.05$. Second, DNA methylation at the CpG loci showed a larger variation between individuals than within individuals with ICC threshold ≥0.70 (31). Third, the difference within a given CpG locus was
consistent across the majority of matched pairs case-control in our Study Group (≥17/21 pairs). Fourth, the position of the CpG locus corresponded to a gene regulatory region.

Taking into account those criteria, we selected 11 CpG loci corresponding to 10 genes for further validation. Relative methylation levels (beta values) in cases (median of n=21) and matched controls (median of n=21) for those loci are shown in a heatmap in Figure 2F. Two CpG sites were hypermethylated in cases vs. controls and corresponded to CHRDL1 and RPTOR. The remaining CpG sites, located within ACAN, ATHL1, PSG4, BRUNOL5, BARD1, FXYD6, TET1, or MAGEB3, demonstrated lower methylation levels in cases vs. controls based on the array data. All selected loci are located in regulatory gene regions including CpG islands (CpGI), CpGI shores (< 2kb flanking CpGI) and shelves (< 2kb flanking outwards from a CpGI shore), 5'UTR, or promoters (TSS1500) (Figure 2F). The genes encompassing the selected loci are associated with a wide range of functions. RPTOR and CHRDL1 regulate respectively mTOR and TGF-beta signaling (33,34). PSG4 is involved in regulation of innate immune system (35) whereas BRUNOL5 regulates gene transcription through pre-mRNA alternative splicing and mRNA editing (36). BARD1 is implicated in transcriptional regulation, control of DNA damage repair and cell cycle (37), whereas ACAN and FXYD6 are important for cell adhesion and cell-cell communication (38,39). Another interesting finding is that one of the genes containing hypomethylated CpG sites in blood DNA of cases is TET1. TET1 is implicated in gene-specific hypomethylation (40). Thus, TET1 could be a likely candidate responsible for the profound loss of methylation at gene-specific loci that we detect in blood DNA of cases. This interesting hypothesis will be mechanistically tested in future studies.

The values obtained from the microarray are semi-quantitative and require further validation. For validation by pyrosequencing, PCR primers were designed to encompass the CpG site differentially methylated on the microarray (CpG of interest, marked in square in box plots of Figure 3) and additional neighboring CpG sites (please see gene maps in the left panel of Figure 3). Thus, we could obtain a pattern of methylation within a broader DNA region. We call the fragments validated by pyrosequencing “probes” and take into account DNA methylation levels at every CpG within the tested probe. Pyrosequencing quantitatively measures the percentage of methylation at a single CpG site resolution and requires only PCR of bisulfite converted genomic DNA. Such a simple method would be easily translatable to a clinical setting.
DNA methylation levels within selected probes quantified by pyrosequencing distinguish blood of HCC cases prior to diagnosis from blood of healthy controls

Technical validation by pyrosequencing of 11 probes indicates statistically significant difference in methylation level at majority of CpG sites of interest in pre-diagnostic blood samples from HCC cases vs. healthy controls ($P < 0.05$, Mann–Whitney U-test) (Figure 3, right panel). The results are shown in box plots created from dots corresponding to patient samples and representing the percentage of methylation at a given CpG site. Among CpG sites of interest (covered on the array), the highest differences were detected for CpG#2 of $ATHL1$ (42.5% difference in median), CpG#1 of $ACAN$ (26% difference in median), CpG#3 of $CHRD1$ (17% difference in median), and CpG#4 of $BARD1$ (13.3% difference in median). Similarly to pyrosequencing results, $ATHL1$ and $ACAN$ show the highest differences on the microarray (Figure 2F). $BRUNOL5$ is an interesting example where the influence of the neighboring CpG on the intensity detectable by the array is clearly visible. Specifically, CpG covered on the array (CpG#3) was significantly hypomethylated in cases vs. controls based on the array data. However, this CpG was hypomethylated in 16 cases vs. matched controls according to quantitative pyrosequencing readings. The remaining 5 samples were hypomethylated at neighboring CpG#2. Significant hypomethylation ($P<0.05$, Mann–Whitney U-test) was detected only for combined analysis of CpG#2 and CpG#3. Similarly, pyrosequencing data indicate significant hypomethylation at CpG#1 within $TET1$ instead of CpG#2 covered on the array.

Differential methylation within $BARD1$, $FXYD6$, $BRUNOL5$, $MAGEB3$, and $TET1$ as quantified by pyrosequencing distinguishes blood of HCC cases prior to diagnosis from blood of cirrhotic controls free of cancer in the independent cirrhotic Study Group

Liver cirrhosis is one of the main risk factors of HCC, but not all cirrhotic patients develop cancer. The main challenge that clinicians face today in liver disease management is to identify individuals with liver cirrhosis who are at higher risk of HCC development. This challenge may be overcome by identifying early detection biomarkers that distinguish not only HCC subjects from healthy individuals but also cirrhotic patients that will develop HCC from those that will not. We therefore tested using pyrosequencing whether 11 differentially methylated probes established as potential biomarkers in samples derived from HCC patients prior to diagnosis and matched healthy controls (Study Group #1), could serve as biomarkers for HCC early detection and surveillance in populations at risk (i.e., individuals with cirrhosis, independent Study Group #2).
To address this question, we tested blood samples from a prospective cohort of cirrhotic patients without HCC at the time of blood collection. Thirteen of these patients developed HCC within 2 years of follow-up. These 13 cases were matched with 23 cirrhotic controls that stayed cancer free during the same follow-up period (Study Group #2) (Supplementary Table S2). Each of 10 cases were matched with 2 controls and the remaining 3 cases were each matched with 1 control. DNA methylation was quantified by pyrosequencing in all 11 probes. Methylation levels at CpG sites of interest (covered on the array, marked in square in boxplots of Figure 3 and 4) and neighboring CpG sites within BARD1, FXYD6, and MAGEB3 distinguished pre-diagnostic HCC cases from matched cirrhotic controls (Figure 4A). The highest difference was detected at CpG#4 and CpG#6 within BARD1 (15% difference in median). Similarly to Study Group #1, significant hypomethylation within BRUNOL5 was detected only for combined analysis of CpG#2 and CpG#3 (Figure 4A). For TET1, the average methylation across first three CpG sites in the tested fragment separated cirrhotic individuals who developed HCC within 2 years from those who stayed cancer free (Figure 4A). Cirrhotic controls showed methylation levels similar to cases for the remaining probes, namely ATHL1 (probe 1 and 2), ACAN, PSG4, CHRDL1, and RPTOR (Supplementary Figure S1A). Although ATHL1 shows significant differences at CpG#2 and CpG#12, small magnitude of those changes and lack of variable methylation levels between cases and controls at other neighboring CpG sites led to excluding ATHL1 probes from further analysis as potential markers in the cirrhotic population (Supplementary Figure S1A). Altogether, 5 out of 11 probes show a promise as candidate biomarkers of HCC early detection in populations at risk.

Performance of the established candidate biomarkers can be diminished due to the small size of case samples in Study Group #2. For this reason, we performed a combined analysis including pre-diagnostic HCC cases from Study Group #1 (n = 21). Fifteen out of 21 cases had registered underlying cirrhosis and thus would be included into populations at risk along with 13 cases from Study Group #2. Although the remaining 6 samples did not have registered cirrhosis, the age and HCC etiology would strongly indicate liver cirrhosis in those individuals. Hence, we included all 21 case samples in the combined pyrosequencing analysis (Figure 4B). We found that 5 out of 11 probes consistently distinguish cases from cirrhotic controls which confirms the results in matched samples in Study Group #2 (Figure 4B, Supplementary Figure S1B).

**Differential methylation within candidate biomarkers for HCC early detection in populations at risk is maintained throughout the disease progression**
We established 5 probes that separate HCC cases prior to diagnosis from cirrhotic controls (Figure 4). Biomarkers at pre-diagnostic stages of a disease will be accurate only if they persist throughout the disease progression. We therefore tested whether those 5 probes can also distinguish HCC samples collected after diagnosis from cirrhotic controls or healthy controls.

We tested 24 blood samples collected from HCC cases after conventional diagnosis (post-diagnostic cases) and 24 blood samples from matched healthy controls (independent Study Group #3, post-diagnostic) (Supplementary Table S3). Pyrosequencing of the 5 probes, BARD1, FXYD6, BRUNOL5, MAGEB3, and TET1, indicates statistically significant difference in methylation levels at nearly all CpG sites of interest (i.e., covered on the array) and several neighboring sites (P < 0.05, Mann–Whitney U-test) (Supplementary Figure S2). The changes in DNA methylation levels between cases and controls within the 5 probes were also maintained when we compared post-diagnostic cases with cirrhotic controls as indicated in Figure 4C.

**Establishing the predictive panel of CpG sites within candidate biomarkers for HCC early detection in populations at risk**

Five probes that we determined to separate cases from controls in all tested Study Groups constitute candidate biomarkers for HCC early detection. To test their performance in a panel for detection of HCC at the time when diagnosis with conventional tools is impossible, we selected 9 out of 19 CpG sites within those probes which demonstrate consistent difference in DNA methylation between cases and controls in all three Study Groups (Table 1). Receiver Operating Characteristic (ROC) analysis was performed as described in Materials and Methods. Area under the curve (AUC) values from ROC analyses of CpG loci in pre-diagnostic cases vs. healthy controls or pre-diagnostic cases vs. cirrhotic controls are shown in Table 1. The CpG sites best at discriminating between pre-diagnostic cases and cirrhotic controls were detected to be CpG#3 and CpG#6 of BARD1, CpG#4 of FXYD6, and CpGs #1-3 of TET1 (asymptotic significance <0.05, Table 1). However, none of the CpGs alone reached AUC > 0.80. Therefore, we explored the discriminant ability of a panel containing these 9 CpG sites (Table 1). To test the panel, the individual CpG sites were combined together and ran as an additive logistic regression model. The logistic model determines the probability of having cancer for each individual subject in the dataset, given the DNA methylation measurement for the CpGs in the panel. ROC curves for the panel of 9 CpG sites demonstrate high accuracy in discriminating pre-diagnostic cases from healthy controls (AUC = 0.914; 95% CI: 0.82-1.00), and pre-diagnostic cases from cirrhotic
controls (AUC = 0.887; 95% CI: 0.80-0.98) (Table 1, Figure 5A-B). In Study Group #1 (pre-diagnostic cases vs. healthy controls), where the probes were discovered, the panel was able to correctly identify 86% of the total number of subjects in the data at a specificity of 86% (Figure 5D). In our independent cirrhotic population, the panel was able to distinguish pre-diagnostic cases from cirrhotic controls free of cancer with 88% sensitivity at a specificity of 70% (Figure 5E). High accuracy of the panel was further confirmed in discriminating post-diagnostic cases from cirrhotic controls (AUC = 0.835; 95% CI: 0.72-0.95), resulting in 80% sensitivity at a specificity of 70%. Restricting the panel to three CpG sites, including CpG#6 of BARD1, CpG#3 of TET1, and CpG #1 of MAGEB3, that had the most influence in calculating the respective probability in the model based on the model estimates, resulted in similar diagnostic accuracy in the cirrhotic population as the panel of 9 CpGs (AUC = 0.872; 95% CI: 0.78-0.97) (Figure 5C and 5F). However, the panel of 3 CpGs is not as accurate as the panel of 9 CpGs in discriminating post-diagnostic cases from cirrhotic controls (AUC = 0.64; 95% CI: 0.48-0.81), suggesting the advantage of the panel consisting of 9 CpGs.

Among 21 pre-diagnostic cases in Study Group #1 and 13 cases in Study Group #2, 12 cases were diagnosed with HCC within 6 months of blood collection. Such a short time to diagnosis may boost the performance of the established biomarker panel in HCC early detection. Therefore, we ran the model after excluding those 12 samples. Interestingly, we reached 86% sensitivity at a specificity of 80% in distinguishing pre-diagnostic cases with blood draw more than 6 months of diagnosis from cirrhotic controls (AUC = 0.915; 95% CI: 0.84-0.99). This confirms that the model has a high potential to serve as an early diagnosis tool.

DISCUSSION

Lack of reliable screening tests accounts for high mortality rate in HCC (1,2). Hence, identifying tools that will improve HCC early detection and surveillance in patients at risk is of high interest. As epigenetic alterations are implicated in HCC development and progression (5,7-9), first attempts in biomarker research aimed at establishing epigenetic changes in tissue biopsies that could discriminate cancer from normal samples (8,9,19). Commonly studied hypermethylated genes include p16, SOCS1, GSTP1, and CDH1, whose differential methylation was proposed as a prognostic marker of HCC associated with clinicopathological data (19). We further proposed three genes, GPM6B, MAGEA12 and FCRL1, as candidate HCC biomarkers whose DNA methylation levels discriminate tumors from normal adjacent tissue and tumors from cirrhotic tissue (9). Although promising, those markers have
low likelihood to be applied in the clinic because biopsies are not appropriate for early diagnostic tests. The next challenge in the field was to find an approach where those biomarkers can be detectable by minimally invasive methods. Numerous studies, including HCC (19,25), colorectal (17), and ovarian cancers (18), showed that blood serum may be a promising material carrying the same DNA methylation signature of markers as tumor tissue because it contains circulating tumor DNA. However, there is an obstacle in using that approach in early cancer detection mainly due to low levels of circulating tumor DNA in serum at early phases of the disease (20). Since cancer is a systemic disease, we and others hypothesize that white blood cells (WBC) will have different DNA methylation patterns specific to patients developing HCC compared to individuals without cancer (19,26). This pattern of alterations can constitute an early diagnostic marker of HCC.

Recent studies show that DNA methylation patterns in white blood cells (WBC) might carry reliable biomarkers. Global methylation of genomic DNA in WBC differentiated cancer patients from healthy individuals and was suggested as a useful epigenetic biomarker for cancer risk estimation and early detection (21,22). Hypomethylation within BLCA-4 repeats in blood cells was associated with higher risk of bladder cancer (23). Methylation in \( ATM \) intragenic loci in DNA from WBC was proposed as a potential marker of breast cancer risk detectable on average 3 years before diagnosis (24). Furthermore, higher methylation of exportin 4 (\( XPO4 \)) in WBC DNA was shown to be associated with HCC and higher levels of AFP (26).

Most of the studies on HCC biomarkers, either using plasma or WBC DNA as described above (19,25,26), were performed on samples from already diagnosed patients, often at advanced HCC stages, which impedes the clinical use of those proposed biomarkers for early detection. In our present investigation, we use blood samples collected in prospective cohorts from HCC cases at the time when they were clinically considered cancer free (pre-diagnostic stage); thereby our study is unique and highly relevant to the field of early detection biomarker discovery. Despite the limited sample size at the selection stage using a genome-wide approach (Figure 1), we were able to verify the established biomarkers in independent Study Groups #2 and #3 using pyrosequencing, a quantitative method. This provides a strong support for successful implementation of the biomarkers in larger cohorts.

Using the design strategy shown in Figure 1, we delineated genome-wide DNA methylation profiles in blood samples from HCC cases prior to conventional diagnosis and from matched healthy controls.
We found that blood DNA has different DNA methylation patterns specific to patients developing HCC compared to individuals without cancer (Figure 2). The majority of differentially methylated CpG sites were hypomethylated in cases vs. controls (Figure 2A-C). These hypomethylated CpG sites were predominantly located in regulatory regions of genes that are involved in regulation of transcription and signal transduction (Figure 2D-E). Hypomethylation could be potentially linked to up-regulation of those genes suggesting that they might be functionally involved in immunological response of white blood cells to cancer. Furthermore, we found among hypomethylated genes several candidates implicated in immunomodulation, including CSF2, IL9, and IFITM5. Indeed, differential methylation of growth factor genes such as FGF12 and IGF1, and immune-related genes such as CCL28, TLR2 and TGFBI, was previously reported in blood DNA upon exposure to aflatoxin B1, a risk factor of HCC (15).

Among differentially methylated sites, we have identified 11 CpG loci that met several criteria for successful biomarker discovery. The selected loci were also located in regulatory regions of genes (Figure 2F) with functions relevant to cellular processes including gene transcription, signal transduction, and cell motility, adhesion, and communication (33-39). This resembles the functions and pathways that we found in our previous studies to be enriched in genes differentially methylated in HCC tumor tissues (8). Strikingly, further investigation of those genes showed certain similarities of the differential DNA methylation pattern between blood DNA and tumor tissues. Similar concordance between DNA methylation patterns in WBC and tumor tissue was previously reported (16). Supplementary Figure S3 presents ACAN as an example. Using publicly available tumor microarray data in TCGA database (methylation), we discovered that ACAN, hypomethylated in blood DNA in our Study Groups, is also hypomethylated in HCC tumor tissues (Supplementary Figure S3A). Our previous report indicated 4-fold overexpression of ACAN in HCC tumors vs. normal adjacent tissue (8). Overexpression in many other types of cancer was further confirmed using the Oncomine database (Supplementary Figure S3B). Importantly, DNA methylation levels determined in our genome-wide investigation reflect patterns found in WBC not blood serum, thus excluding circulating tumor DNA. We are intrigued by the observation that some of the genes found to be hypomethylated in WBC DNA are also hypomethylated in liver tumor tissue. This observation requires further future investigation.

Differential methylation registered on the microarray within those 11 CpG loci was validated by an independent quantitative method, namely pyrosequencing. The validated probes are candidate
biomarkers for HCC early detection as they discriminate pre-diagnostic HCC cases from healthy individuals. However, these probes will be relevant in the clinic only if they can be used in populations at risk. Using samples from an independent Study Group, we found that 5 of the validated probes located within regulatory regions of BARD1, FXYD6, BRUNOL5, MAGEB3, and TET1 can distinguish cirrhotic patients who will develop HCC within next 4 years from those who will stay cancer free (Figure 4, Supplementary Figure S1). They consistently separated pre-diagnostic HCC cases in Study Groups #1 and #2 from cirrhotic controls (Figure 4A-B) as well as post-diagnostic HCC cases in Study Group #3 from cirrhotic controls (Figure 4C). Hence, the discriminative power of those probes persist when the disease progresses which ensures lack of false negative results where a patient with developed HCC could be assessed as a control (Figure 4, Supplementary Figure S2). Using an additive logistic regression model, we further showed that 9 CpG sites within those probes (Table 1), which demonstrate consistent difference in DNA methylation between cases and controls in all three Study Groups, constitute a panel with high diagnostic accuracy in populations at risk (Figure 5). Its performance, once validated in larger cohorts, would strongly exceed the diagnostic accuracy of AFP measurements that are currently recommended as HCC surveillance method (3). In Study Group #1, AFP values were available for 15 patients. None of those patients had AFP values higher than 400-500 ng/ml that would be considered diagnostic for HCC (Supplementary Table S1). Based on AFP values, we would be able to diagnose HCC only in 2 out of 13 patients in Study Group #2 (Supplementary Table S2) and 2 out of 23 patients in Study Group #3 (Supplementary Table S3). Using our biomarker panel, we could tremendously increase the diagnostic potential by reaching sensitivity between 80-90% (Figure 5).

Our present study establishes for the first time a panel of specific CpG sites whose methylation in blood DNA can distinguish HCC cases from cirrhotic individuals at the time when cases are clinically considered free of cancer. Pre-diagnostic samples used in our investigation constitute highly valuable material that is the most suitable for discovery of early diagnosis markers. Using blood as a minimally invasive material and pyrosequencing as a straightforward quantitative method, the identified probes have high potential to be developed into a routine clinical test after validation in larger studies. Such a set of specific alterations in DNA that could indicate by a simple blood test the presence of HCC before the onset of symptoms will accelerate progress to decrease HCC mortality by diagnosing cancer early, and thus achieving high cure rate.
FUNDING

This research was supported by the American Cancer Society Institutional Research Grant (IRG-14-190-56); Showalter Trust Award; and the Purdue University Center for Cancer Research Undergraduate Research Program, granted to BS. JMF and KF were funded by Breast Cancer Now and acknowledge funding support from Cancer Research UK (A13086) and the National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London.

The Authors declare no conflict of interest.

REFERENCES


FIGURE LEGEND

Figure 1. Overview of the design strategy.

Figure 2. The landscape of DNA methylation delineated using Illumina Infinium Human Methylation 450K BeadChip microarray platform in blood DNA collected prior to conventional diagnosis with hepatocellular carcinoma (HCC). (A) Pie chart of the 971 CpG sites differentially methylated between HCC cases and matched healthy controls. Diff. refers to differential methylation (delta beta = case beta value - control beta value). (B) Distribution of differences in DNA methylation levels (delta beta values = differential methylation, Y axis) across all 971 differentially methylated loci (X axis). Seventy six percent of probes are hypomethylated in case versus control ($P = 2.973 \times 10^{-11}$, Wilcoxon test, 0.22% methylation across the whole array). (C) Chromosomal view of 971 differentially methylated CpG sites. Regions differentially methylated in HCC cases versus matched healthy controls are shown in the bar track of the chromosomal views, with blue indicating hypomethylation and red indicating hypermethylation (case minus control). (D) Classification of differentially methylated CpG sites according to their location relative to CpG islands (CpGI) and transcription start site (TSS). Hypomethylated loci are located in CpGI shores (left panel), and within promoters and 5’UTR regions (right panel) at higher frequency than hypermethylated sites. (E) Functional analyses using GO, KEGG and DAVID knowledgebase indicate biological functions and pathways associated with genes corresponding to CpG sites hypomethylated in cases versus controls. (F) A heatmap showing relative methylation levels (beta values) in cases (average of n=21) and matched controls (average of n=21) for the 11 CpG loci corresponding to 10 genes selected for further validation.

Figure 3. Quantitative analysis of methylation states of differentially methylated loci corresponding to ATHL1, BRUNOL5, ACAN, BARD1, TET1, FXYD6, MAGEB3, PSG4, RPTOR, and CHRD1 in pre-diagnostic hepatocellular carcinoma (HCC) cases versus matched healthy controls. Using Illumina 450K microarray, DNA methylation was determined in blood DNA collected from individuals up to 4 years prior to conventional diagnosis with HCC. Based on the microarray data, 11 CpG sites were chosen for validation by pyrosequencing. The difference in DNA methylation,
statistical significance, consistency of the difference, location of the CpG site, and the function of a corresponding gene were taken into account in the selection. Right panel depicts average methylation state of CpG sites in HCC cases (grey, n=21) and healthy controls (white, n=21) in 11 regions (probes) corresponding to 10 genes. Each probe encompasses a differentially methylated CpG site covered on Illumina 450K microarray (marked in square). The exact position of validated CpG sites relative to transcription start site (TSS) is shown in gene maps in the left panel. The tested region is shaded, and pyrosequenced CpGs are circled and numbered. The putative transcription factor binding sites are indicated as predicted by TransFac. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

**Figure 4.** Methylation levels of BARD1, FXYD6, BRUNOL5, MAGEB3, and TET1 in pre-diagnostic and post-diagnostic hepatocellular carcinoma (HCC) cases versus cirrhotic controls free of cancer: potential for biomarkers in populations at risk. (A-C) Average methylation state of CpG sites in: A) pre-diagnostic HCC cases with registered cirrhosis (n=13) and matched cirrhotic controls (n=23); B) all pre-diagnostic HCC cases (n=21 + n=13) and cirrhotic controls (n=23); C) post-diagnostic HCC cases (n=24) and cirrhotic controls (n=23), as determined by pyrosequencing in BARD1, FXYD6, BRUNOL5, MAGEB3, and TET1. Hypomethylated CpG site covered on Illumina 450K microarray is marked in square. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$

**Figure 5.** Predictive properties of a panel of CpG sites within 5 probes that were identified and validated as significantly differentially methylated between pre-diagnostic cases and healthy or cirrhotic controls. Receiver Operating Characteristic (ROC) curve analysis was performed to examine the discriminative power of the panel of CpG sites using the corresponding DNA methylation measurements at each site. The analysis was performed using the Proc Logistic function in SAS 9.4 and the ROC Curve analysis function in SPSS 23. (A-C) ROC curves for the ability of the panel of 9 CpG sites (A and B) or 3 CpG sites (C) to differentiate pre-diagnostic cases from healthy controls (A, Study Group #1), and pre-diagnostic cases from cirrhotic controls (B and C, pre-diagnostic cases from Study Groups #1 and #2). The individual CpG sites were combined together to form a panel and ran as an additive logistic regression model. Area under the curve (AUC) is depicted for the panel of biomarkers in healthy and cirrhotic populations. The logistic model determines the probability of having cancer for each individual subject in the dataset, given the measurement for the CpGs in the panel. Sensitivity and specificity values for the panel are depicted. (D-F) Performance of the biomarker panel of 9 CpGs (D and E) or 3 CpG sites (F) in pre-diagnostic cases vs. healthy controls (D, Study
Group #1), and in pre-diagnostic cases vs. cirrhotic controls (E and F, pre-diagnostic cases from Study Groups #1 and #2), as depicted in heatmaps. Blue color corresponds to a sample predicted by the model to be a case whereas light grey to a sample predicted to be a control.

Table 1. Area under the curve (AUC) values from Receiver Operating Characteristic (ROC) analyses of CpG loci in pre-diagnostic cases vs. healthy controls or cirrhotic controls. ROC curve analysis was performed using the Proc Logistic function in SAS 9.4 to and the ROC Curve analysis function in SPSS 23 to examine the discriminative power of each individual methylation CpG site. Each of the CpG methylation sites were explored individually to gain insight on how well each site could help correctly identify the controls and cancer subjects in the data.
**SELECTION**

**Study Group #1:** Pre-diagnostic HCC vs. healthy controls

Genome-wide DNA methylation (Illumina 450K): Blood from HCC cases prior to conventional diagnosis (n=21) vs. matched healthy controls (n=21)

*Threshold:* differential methylation > 0.1, P < 0.05, ICC ≥ 0.7, consistency of the difference in matched pairs case-control (≥17/21 pairs)

11 CpG loci corresponding to 10 genes as candidate biomarkers distinguishing cases from controls

Validation of 11 CpG loci using pyrosequencing: establishing 11 probes

---

**VERIFICATION**

**Study Group #2:** Pre-diagnostic HCC vs. cirrhotic controls

Verification of 11 probes in population at risk: HCC cases prior to conventional diagnosis (n=13) vs. matched cirrhotic controls free of cancer (n=23)

5 probes established as candidate biomarkers for HCC early detection in cirrhotic populations (cases: n=13+21, controls: n=23)

**Study Group #3:** Post-diagnostic HCC vs. healthy controls

Confirmation of differential methylation within 5 established probes in HCC cases after conventional diagnosis (n=24) vs. matched healthy controls (n=24)

---

**BIOMARKER PANEL**

Logistic regression to establish the prediction panel of CpG sites within 5 established probes

Test the accuracy of the panel in diagnosis of HCC in cirrhotic population
Table 1. Area under the curve (AUC) values from Receiver Operating Characteristic (ROC) analyses of CpG loci in pre-diagnostic cases vs. healthy controls or cirrhotic controls. ROC curve analysis was performed using the Proc Logistic function in SAS 9.4 and the ROC Curve analysis function in SPSS to examine the discriminative power of each individual methylation CpG site. Each of the CpG methylation sites were explored individually in order to gain insight on how well each site could help correctly identify the controls and cancer subjects in the data.

<table>
<thead>
<tr>
<th>CpG Site</th>
<th>Pre-diagnostic Cases Study Group #1 (n=21) vs. Healthy Controls (n=21)</th>
<th>Pre-diagnostic Cases Study Groups #1 and #2 (n=13+21) vs. Cirrhotic Controls (n=23)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>Asymptotic Significance</td>
</tr>
<tr>
<td>BARD1 CpG 3</td>
<td>0.711 (0.55-0.87)</td>
<td>0.0180</td>
</tr>
<tr>
<td>BARD1 CpG 6</td>
<td>0.779 (0.64-0.92)</td>
<td>0.0011</td>
</tr>
<tr>
<td>BRUNOL5 CpG 2</td>
<td>0.578 (0.40-0.76)</td>
<td>0.4870</td>
</tr>
<tr>
<td>BRUNOL5 CpG 3</td>
<td>0.586 (0.41-0.76)</td>
<td>0.3310</td>
</tr>
<tr>
<td>FXYD6 CpG 4</td>
<td>0.667 (0.50-0.84)</td>
<td>0.0490</td>
</tr>
<tr>
<td>MAGEB3 CpG 1</td>
<td>0.669 (0.50-0.84)</td>
<td>0.0740</td>
</tr>
<tr>
<td>TET1 CpG 1</td>
<td>0.689 (0.52-0.86)</td>
<td>0.0350</td>
</tr>
<tr>
<td>TET1 CpG 2</td>
<td>0.588 (0.41-0.77)</td>
<td>0.2550</td>
</tr>
<tr>
<td>TET1 CpG 3</td>
<td>0.612 (0.44-0.79)</td>
<td>0.1720</td>
</tr>
<tr>
<td>Model 9 CpGs</td>
<td><strong>0.914 (0.82-1.00)</strong></td>
<td><strong>0.0050</strong></td>
</tr>
</tbody>
</table>

Sensitivity at 80% Specificity  
90% (70-99)  
82% (65-93)

Sensitivity at 70% Specificity  
90% (70-99)  
88% (72-97)