DNA methylation markers for diagnosis and prognosis of common cancers

Xiaoke Hao,a,1,2 Huiyan Luo,b,c,d, Michal Krawczyk,a,1 Wei Wei,b,c,1 Wenyu Wang,c,d,1 Juan Wang,a,1 Ken Flagg,e, Jiayi Hou,a Heng Zhang,a Shaohua Yi,a Maryam Jafari,a Danni Lin,f Christopher Chung,g Bennett A. Caughey,h Gen Li,j Debjan Dhar,k William Shi,k Lianghong Zheng,k Rui Hou,k Jie Zhu,k Liang Zhao,k Xin Fu,a Edward Zhang,k Charlotte Zhang,k Jian-Kang Zhu,k Michael Karin,g,h 1, Rui-Hua Xub,h and Kang Zhangc,h,2

*Department of Clinical Laboratory Medicine, Xijing Hospital, Fourth Military Medical University, Xi’an 710032, China; †State Key Laboratory of Oncology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China; ‡Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92037; 
§Shanghai Key Laboratory of Ocular Fundus Diseases, Shanghai General Hospital, Shanghai 200080, China; ‖Shanghai Center for Plant Stress Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai 210602, China; ¼Guangzhou Youze Biological Pharmaceutical Technology Company Ltd., Guangzhou 510005, China; ‡Department of Pharmacology, University of California, San Diego, La Jolla, CA 92328; and ½Veterans Administration Healthcare System, San Diego, CA 92093

Contributed by Michael Karin, May 24, 2017 (sent for review March 3, 2017; reviewed by Hakon Hakonarson and Wei Zhang)

The ability to identify a specific cancer using minimally invasive biopsy holds great promise for improving the diagnosis, treatment selection, and prediction of prognosis in cancer. Using whole-genome methylation data from The Cancer Genome Atlas (TCGA) and machine learning methods, we evaluated the utility of DNA methylation for differentiating tumor tissue and normal tissue for four common cancers (breast, colon, liver, and lung). We identified cancer markers in a training cohort of 1,619 tumor samples and 173 matched adjacent normal tissue samples. We replicated our findings in a separate TCGA cohort of 591 tumor samples and 91 matched adjacent normal tissue samples, as well as an independent Chinese cohort of 394 tumor samples and 324 matched adjacent normal tissue samples. The DNA methylation analysis could predict cancer versus normal tissue with more than 95% accuracy in these three cohorts, demonstrating accuracy comparable to typical diagnostic methods. This analysis also correctly identified 29 of 30 colorectal cancer metastases to the liver and 32 of 34 colorectal cancer metastases to the lung. We also found that methylation patterns can predict prognosis and survival. We correlated differential methylation of CpG sites predictive of cancer with expression of associated genes known to be important in cancer biology, showing decreased expression with increased methylation, as expected. We verified gene expression profiles in a mouse model of hepatocellular carcinoma. Taken together, these findings demonstrate the utility of methylation biomarkers for the molecular characterization of cancer, with implications for diagnosis and prognosis.

DNA methylation | cancer diagnosis | cancer prognosis | gene expression | survival analysis

Accurate diagnosis of cancer based on histological subtype, as well as other markers identified via histology and immunohistochemistry, is crucial for choosing the proper treatment and for predicting survival (1). For some primary tumors, complex anatomy may prevent accurate identification of the tissue of origin or tumor type. Tissue must be obtained from these tumors either from surgical resection or from a tissue biopsy. Diagnosis in these cases may be limited by the patient’s tolerance of surgery or by inaccessibility of the tumor, preventing acquisition of a tissue sample of adequate size and quality that preserves tissue architecture. Even when high-quality biopsy specimens are obtained, diagnostic uncertainty may persist, hindering treatment decisions and prognosis. Thus, there is a need for strategies to improve diagnostic certainty. Molecular characterization is increasingly used to predict tumor prognosis and response to therapy and offers great potential for improving understanding of an individual patient’s tumor (2–4). Importantly, these methods may have specific utility in scenarios of limited tissue availability or quality.

Methylation of CpG sites is an epigenetic regulator of gene expression that usually results in gene silencing (5, 6). Extensive perturbations of DNA methylation have been noted in cancer, causing changes in gene regulation that promote oncogenesis (7–9). Understanding both epigenetic changes and somatic DNA mutations show promise for improving the characterization of malignancy to predict treatment response and prognosis (3, 10–12). Some changes in methylation are reproducibly found in nearly all cases of a specific type of cancer. In contrast, somatic mutations are often neither specific nor sensitive for a particular type of cancer. Even within commonly mutated genes, individual mutations may be found across tens or hundreds of kilobases, limiting the utility of targeted sequencing of molecular markers (10, 13, 14).

Consequently, to explore the utility of DNA methylation analysis for cancer diagnosis, we analyzed whole-genome methylation profiles of tumors and normal tissue from patients with four of the most common cancers to identify potential cancer-specific DNA methylation markers. We then verified these methylation markers in two other independent patient

Significance

The ability to identify a specific cancer using minimally invasive biopsy holds great promise for improving diagnosis and prognosis. We evaluated the utility of DNA methylation profiles for differentiating tumors and normal tissues for four common cancers (lung, breast, colon, and liver) and found that they could differentiate cancerous tissue from normal tissue with more than 95% accuracy. This signature also correctly identified 29 of 30 breast cancer metastases and 32 of 34 colorectal cancer metastases to the lung. We report that methylation patterns can predict the prognosis and survival, with good correlation between differential methylation of CpG sites and expression of cancer-associated genes. Their findings demonstrate the utility of methylation biomarkers for the molecular characterization, diagnosis, and prognosis of cancer.


Reviewers: H.H., Children’s Hospital of Philadelphia; and W.Z., Wake Forest Baptist Comprehensive Cancer Center. The authors declare no conflict of interest.

1X.H., H.L., M. Krawczyk, W. Wei, W. Wang, and J.W. contributed equally to this work. 2To whom correspondence may be addressed. Email: kaoe@fmmu.edu.cn, mkarin@ucsd.edu, xuhb@tsguc.org.cn, or kang.zhang@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1703577114/-/DCSupplemental.
cohorts. We also used methylation patterns to predict survival and analyzed the utility of combining methylation with muta-
tional status in several tumor types. Finally, we correlated spe-
cific methylation patterns with gene expression in genes known
to be important in cancer biology.

Results
Characteristics of Patients and Tissues. Clinical characteristics and
molecular profiling, including methylation data for a training
cohort of 1,619 tumor samples and 173 matched adjacent normal
tissue samples, as well as a validation cohort of 791 tumor and
93 matched normal samples, were obtained from The Cancer
Genome Atlas (TCGA). A separate validation cohort of 394 tu-
mor samples and 324 matched normal samples was obtained
from Chinese patients with cancer treated at the Sun Yat-sen
University Cancer Center, West China Hospital, and Xijing
Hospital. Matched adjacent normal tissue samples were col-
lected simultaneously with tumor tissue from the same patient
and were verified by histology to have no evidence of cancer.
Clinical characteristics of all patients are summarized in SI Ap-
pendix, Tables S1–S3.

Methylation Profiling Identifies Cancer-Specific Methylation Signatures.
To identify a cancer type-specific signature, we randomly split
the full TCGA dataset into training and test cohorts with a
2:1 ratio in each of the eight types of sample groups. We first
performed the prescreening procedure to remove excessive
noise on the training data using the moderated $t$ statistic (15).
For multinomial classification, we used lasso (least absolute
shrinkage and selection operator) under a multinomial distri-
bution. A multiclass prediction system (16) was constructed to
predict the group membership of samples using a panel of
markers. Hierarchal clustering of these samples according to
differential methylation of CpG sites in this fashion could dis-
tinguish the cancer tissue of origin, as well as differentiate
cancer tissue from normal tissue in our TCGA training cohort
(Table 1). The overall correct diagnosis rate was 98.4%. We
then applied these markers to a TCGA validation cohort (Ta-
ble 2), and found a slightly decreased but statistically similar
correct rate of 97.1%. We also confirmed our results in an in-
dependent cohort of Chinese cancer patients (Table 3), which
also showed a decreased but similar correct rate of 95.0%. Of
note, the methylation analysis of the Chinese cohort was per-
formed using an alternative bisulfite sequencing technique in a
different ethnic and geographic background than the TCGA
cohorts. Overall, these results demonstrate the robust nature of
these methylation patterns in identifying the presence of ma-
lignancy as well as its site of origin (Fig. 1 and SI Appendix,
Table S4 and Fig. S1).

Methylation Block Structure for Improved Allele Calling Accuracy.
We used the well-established concept of genetic linkage disequi-
librium to study the degree of comethylation among different
DNA stands. We used paired-end Illumina sequencing reads to

Table 1. Confusion table of the TCGA training cohort

<table>
<thead>
<tr>
<th>Training cohort</th>
<th>Breast cancer</th>
<th>Colon cancer</th>
<th>Liver cancer</th>
<th>Lung cancer</th>
<th>Normal breast</th>
<th>Normal colon</th>
<th>Normal liver</th>
<th>Normal lung</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>520</td>
<td>275</td>
<td>238</td>
<td>584</td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>1,792</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>59</td>
<td>1</td>
<td>21</td>
<td>23</td>
<td>43</td>
</tr>
<tr>
<td>Liver cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal breast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal colon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal liver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal lung</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>520</td>
<td>275</td>
<td>238</td>
<td>584</td>
<td>65</td>
<td>27</td>
<td>32</td>
<td>49</td>
<td>1,763</td>
</tr>
<tr>
<td><strong>Correct</strong></td>
<td>520</td>
<td>275</td>
<td>238</td>
<td>584</td>
<td>59</td>
<td>21</td>
<td>23</td>
<td>43</td>
<td>1,763</td>
</tr>
<tr>
<td><strong>False-positive</strong></td>
<td>6</td>
<td>5</td>
<td>9</td>
<td>6</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False-negative</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wrong tissue</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correct (%)</strong></td>
<td>100</td>
<td>100</td>
<td>99.6</td>
<td>99.8</td>
<td>90.8</td>
<td>77.8</td>
<td>71.9</td>
<td>87.8</td>
<td>98.4</td>
</tr>
</tbody>
</table>

Orange indicates cancer sample, purple indicates normal sample, and gray indicates correctly diagnosed sample number of each training cohort.

Table 2. Confusion table of validation cohort 1

<table>
<thead>
<tr>
<th>Validation cohort 1</th>
<th>Breast cancer</th>
<th>Colon cancer</th>
<th>Liver cancer</th>
<th>Lung cancer</th>
<th>Normal breast</th>
<th>Normal colon</th>
<th>Normal liver</th>
<th>Normal lung</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>268</td>
<td>129</td>
<td>136</td>
<td>252</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>884</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>28</td>
<td>11</td>
<td>14</td>
<td>20</td>
<td>858</td>
</tr>
<tr>
<td>Liver Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Breast</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Colon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Liver</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal Lung</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td>270</td>
<td>129</td>
<td>136</td>
<td>252</td>
<td>32</td>
<td>18</td>
<td>18</td>
<td>25</td>
<td>884</td>
</tr>
<tr>
<td><strong>Correct</strong></td>
<td>268</td>
<td>129</td>
<td>136</td>
<td>252</td>
<td>28</td>
<td>11</td>
<td>14</td>
<td>20</td>
<td>858</td>
</tr>
<tr>
<td><strong>False-positive</strong></td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False-negative</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wrong tissue</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correct (%)</strong></td>
<td>99.3</td>
<td>100</td>
<td>98.6</td>
<td>99.2</td>
<td>87.5</td>
<td>61.1</td>
<td>77.8</td>
<td>80.0</td>
<td>97.1</td>
</tr>
</tbody>
</table>

Orange indicates cancer sample, purple indicates normal sample, and gray indicates correctly diagnosed sample number of each validation cohort.
identify each individual methylation block (mBlock). We applied a Pearson correlation method to quantify the comethylation of mBlock. We compiled all common mBlocks of a region by calculating different mBlock fractions (Methods). We then partitioned the genome into blocks of tightly comethylated CpG sites that we termed methylation-correlated blocks (MCBs), using an $R^2$ cutoff of 0.5. We surveyed MCBs in cancer and normal tissues and found that MCBs were highly consistent among different cancer and normal tissues. Overall, we found $\sim$3,600 MCBs, approximately one-half of which were incomplete/disrupted (SI Appendix, Fig. S2) owing to short a span of sequenced reads ($\sim$100 base pairs).

We next determined methylation values within MCBs. SI Appendix, Fig. S3 shows an example of MCBs found on chromosome 1 in both normal tissues (breast, colon, liver, and lung) and corresponding tumor tissues: breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), and lung adenocarcinoma (LUAD). We found similar $\beta$ values across multiple CpG sites within a MCB, and thus calculated a compound methylation value for one entire MCB. We used them instead of single CpG sites in downstream bioinformatics pipelines, which significantly enhanced the allellexcalling accuracy.

**Methylation Profiles Can Identify Cancer Metastases to Liver.** Because identifying the tissue of origin is crucial in selecting the optimum treatment strategy for patients presenting with metastases, we investigated the utility of DNA methylation analysis for diagnosis of cancer metastases to liver and lung in our Chinese cohort. In addition to the aforementioned primary tumors, we analyzed 30 colorectal cancer metastases to liver and 34 colorectal cancer metastases to lung. We found that unsupervised hierarchical clustering could differentiate these metastases from colon cancer or normal tissue (Fig. 2). The methylation signature could correctly diagnose 29 of 30 colorectal cancer metastases to liver and 32 of 34 colorectal cancer metastases to lung (Table 3); one of the three misdiagnoses were identified as normal liver and two of the three misdiagnoses were identified as normal colorectal tissue, suggesting that the error was due to tissue contamination. These findings support the potential for using the DNA methylation signature to improve the diagnosis of metastatic disease in addition to primary cancers.

![Fig. 1. Methylation signatures can differentiate different cancer types from corresponding normal tissues. (A) Unsupervised hierarchical clustering and heat map presentation associated with the methylation profile (according to the color scale shown) in different cancer types. (B) ROC curve showing the high sensitivity and specificity in predicting different cancer types. (C) Zoom-in view of the block diagram in B.](image-url)
We next assessed the prognostic utility of a methylation signature for each type of cancer. Clinical and demographic characteristics, including age, sex, race, and American Joint Committee on Cancer stage, were included in the analysis as well, because the prognostic power can be greatly improved by combining this information with informative molecular data (17). For each cancer category, we used two different statistical learning algorithms, lasso and boosting, to reduce the dimensionality of markers and construct a predictive model. We evaluated the prognostic utility on TCGA training and validation cohorts at a 2:1 ratio. Our method performed well in differentiating low-risk and high-risk groups in Kaplan–Meier analyses and in associated log-rank tests with significant P values, demonstrating significant prognostic utility of the methylation signatures in BRCA and LUAD (Fig. 3 and SI Appendix, Tables S5 and S6).

A Cancer Methylation Profile Correlated with Its Gene Expression Pattern and Function. Given that DNA methylation is an essential epigenetic regulator of gene expression, we sought to investigate how differential methylation of sites in genes in cancer versus normal tissue correlated with gene expression. Specifically, we were interested in those methylation sites that predicted the presence of malignancy in our aforementioned signatures. As described in Methods, we used both methylation and RNA sequencing data to select top CpG markers in LIHC for which methylation was significantly correlated with gene expression. As expected, we typically observed an inverse correlation between promoter methylation and gene expression and identified several genes known to be important in carcinogenesis, as well as genes with relatively unknown functional relevance in LIHC. Among the genes hypermethylated with decreased expression, we selected one gene for LIHC [fuzzy planar cell polarity protein (Fuz); Fig. 4]. Overexpression of FUZ suppressed LIHC cell line growth (Fig. 4).

We further attempted to validate a list of top genes whose methylation patterns were closely correlated with gene expression in a mouse model of LIHC. We found a good correlation between the gene expression profiles in human and mouse LIHC (SI Appendix, Fig. S4 and Table S7). These results support a
functional role of these methylation markers in promoting carcinogenesis and provide biological validation for their use in methylation studies to characterize cancers.

**Discussion**

The present study demonstrates the potential for using methylation signatures to identify cancer tissue of origin and predict prognosis. Although we focused on four common cancers here, we expect that DNA methylation analysis can be readily expanded to aid diagnosis of a much larger number of cancers. Our results may be particularly helpful for identifying cancers in cases with an inadequate tissue yield or quality for histological diagnosis, which requires preservation of the tissue architecture. In contrast, DNA methylation analysis requires only a small amount of tissue to obtain adequate DNA, thus potentially allowing the use of lower-quality biopsies. These studies also may have significant utility in assigning diagnoses from analysis of metastatic lesions, especially when the tumor is of an unknown primary cancer type.

Through sequencing of bisulfate-converted DNA (bis-DNA), we identified many previously unknown CpG markers differentially methylated in cancer tissues versus normal tissues. Lehmann-Werman et al. (18) described multiple adjacent CpG sites that share the same tissue-specific methylation pattern. We further explored this concept of the mBlock and found that many nearby methylation markers are highly correlated. This information allowed us to identify additional markers and improve the accuracy of sequencing for determining significant methylation differences. This method has substantial potential for improving the accuracy and utility of DNA methylation analysis for the four study cancer types and other cancers, as well as for expanding the number of diagnostic markers available for interrogation. However, the length of an MCB, which is related to how long a DNA methyl-transferase binds to and exerts its enzymatic effect on modifying adjacent and surrounding CpG sites on a DNA strand, is not clear, because its underlying biochemical basis is not fully defined.

DNA methylation analysis has the potential to improve outcomes, given that accurate diagnosis is often crucial to treatment selection. Our application of methylation signatures to prognosis revealed subsets of patients with positive and negative prognoses. This finding raises the possibility that methylation may help identify relatively indolent or aggressive tumors and may aid decision making regarding the selection of more aggressive or less aggressive treatment and monitoring. Further studies are warranted to fully explore the clinical applications of methylation sequencing to guide personalized care for patients with cancer.
Methods

Training and first validation cohorts were performed on patient data obtained from TCGA. Patient characteristics are summarized in SI Appendix, Tables S1 and S2. Complete clinical, molecular, and histopathological data-sets are available at the TCGA website (https://tcga-data.nci.nih.gov/docs/publications/tcga/). Individual institutions that contributed samples coordinated the consent process and obtained informed written consent from each patient in accordance with their respective institutional review boards. A second independent (Chinese) cohort consisted of patients of the Sun Yat-sen University Cancer Center, the West China Hospital in Chengdu, China, and Xijing Hospital. Those who presented with lung adenocarcinoma, liver hepatocellular carcinoma, breast adenocarcinoma, and colorectal adenocarcinoma, including metastatic disease, were selected and enrolled in this study. Patient characteristics are also summarized in SI Appendix, Tables S1 and S3. This project was approved by the IRB of the Sun Yat-sen University Cancer Center, West China Hospital, and Xijing Hospital. Informed consent was obtained from all patients. Tumor and normal tissues were obtained as clinically indicated for patient care and were retained for this study with patients’ informed consent.

Information on data sources, statistical analyses, probe design, bis-DNA capture, sequencing and data analysis, DNA extraction, cell culture, colony formation assays, and tumor xenografts is provided in SI Appendix.

ACKNOWLEDGMENTS. This study was supported in part by the Carol and Dick Hertzberg Fund, the Richard Annenser Fund, and the Michael Martin Fund.