DNA methylation markers for diagnosis and prognosis of common cancers

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

[a]Department of Clinical Laboratory Medicine, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China; [b]State Key Laboratory of Oncology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China; [c]Institute for Genomic Medicine, University of California, San Diego, La Jolla, CA 92093; [d]Shanghai Key Laboratory of Ocular Fundus Diseases, Shanghai General Hospital, Shanghai 200080, China; [e]Shanghai Center for Plant Stress Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, Shanghai 210602, China; [f]Guangzhou Youze Biological Pharmaceutical Technology Company Ltd., Guangzhou 510005, China; [g]Department of Pharmacology, University of California, San Diego, La Jolla, CA 92328; and [h]Veterans 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 791 tumor samples and 93 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 prognostication. 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 matched 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 >95% accuracy. This signature also correctly identified 19 of 20 breast cancer metastases and 29 of 30 colorectal cancer metastases to the liver. 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.

Author contributions: X.H., M. Karin, R.-H.X., and K.Z. designed research; X.H., H.L., M. Krawczyk, W. Wei, W. Wang, J.W., K.F., H.Z., S.Y., M.J., D.L., C.C., G.L., W.S., L. Zheng, R.H., Jie Zhu, X.F., E.Z., and C.Z. performed research; J.H., B.A.C., D.D., L. Zhao, and Jian-Kang Zhu analyzed data; and X.H., M. Karin, R.-H.X., and K.Z. wrote the paper.

Reviewers: H.H., Children's Hospital of Philadelphia; and W.Z., Wake Forest Baptist Comprehensive Cancer Center.

The authors declare no conflict of interest.

[1]X.H., H.L., M. Krawczyk, W. Wei, W. Wang, and J.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: haoxkg@fmmu.edu.cn, mkarin@ucsd.edu, xurh@sysucc.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.

Table 1. Confusion table of the TCGA training cohort

Training cohort	Breast cancer	Colon cancer	Liver cancer	Lung cancer	Normal breast	Normal colon	Normal liver	Normal lung	Total
Breast cancer	520				6				
Colon cancer		275				5			
Liver cancer			238				9		
Lung cancer				584				6	
Normal breast			1	1	59	1			
Normal colon						21			
Normal liver							23		
Normal lung								43	
Total	520	275	239	585	65	27	32	49	1,792
Correct	520	275	238	584	59	21	23	43	1,763
False-positive					6	5	9	6	26
False-negative			1						3
Wrong tissue				1		1			3
Correct (%)	100	100	99.6	99.8	90.8	77.8	71.9	87.8	98.4

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

cohorts. We also used methylation patterns to predict survival and analyzed the utility of combining methylation with mutational status in several tumor types. Finally, we correlated specific 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 tumor 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 collected 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 Appendix*, 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 distribution. 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 distinguish 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 (Table 2), and found a slightly decreased but statistically similar correct rate of 97.1%. We also confirmed our results in an independent 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 performed 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 malignancy 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 disequilibrium to study the degree of comethylation among different DNA stands. We used paired-end Illumina sequencing reads to

Table 2. Confusion table of validation cohort 1

Validation cohort 1	Breast cancer	Colon cancer	Liver cancer	Lung cancer	Normal breast	Normal colon	Normal liver	Normal lung	Total
Breast Cancer	268			1	4				
Colon Cancer		129				7			
Liver Cancer			136				4		
Lung Cancer	1			252				5	
Normal Breast	1		1		28				
Normal Colon						11			
Normal Liver			1				14		
Normal Lung				1				20	
Totals	270	129	138	254	32	18	18	25	884
Correct	268	129	136	252	28	11	14	20	858
False-positive					4	7	4	5	20
False-negative	1		1	1					3
Wrong tissue	1		1	1					2
Correct (%)	99.3	100	98.6	99.2	87.5	61.1	77.8	80.0	97.1

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

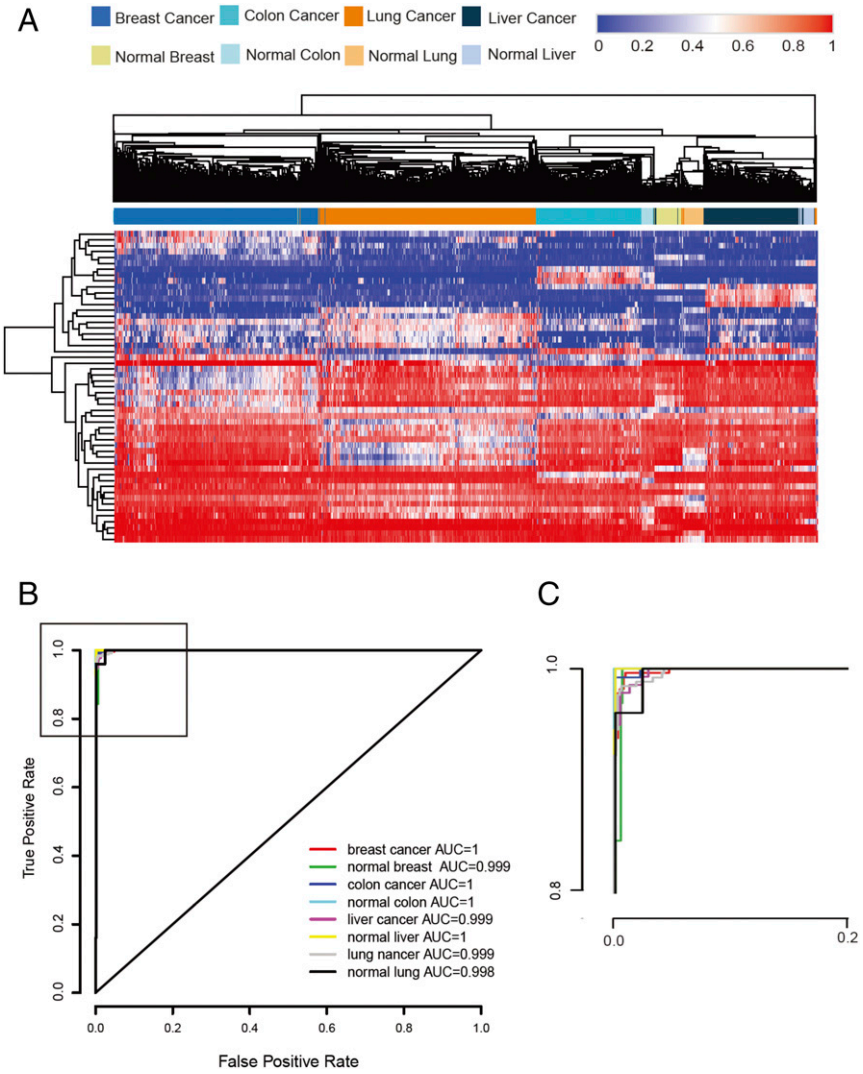

Fig. 1. Methylation signatures can differentiate different cancer types from corresponding normal tissues. (*A*) Unsupervised hierarchical clustering and heat map presentation associated with the methylation profile (according to the color scale shown) in different cancer types. (*B*) ROC curve showing the high sensitivity and specificity in predicting different cancer types. (*C*) Zoom-in view of the block diagram in *B*.

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 ~3,600 MCBs, approximately one-half of which were incomplete/disrupted (*SI Appendix*, Fig. S2) owing to short a span of sequenced reads (~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 β 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 allele-calling 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.

Table 3. Confusion table of validation cohort 2

Validation cohort 2	Breast cancer	Colorectal cancer	Colorectal cancer metastases to liver	Colorectal cancer metastases to lung	Liver cancer	Lung cancer	Normal breast	Normal colon	Normal liver	Normal lung	Total
Breast cancer	65						4			1	
Colorectal cancer	1	154	29	32		1					
Liver cancer		1			44	1			1	3	
Lung cancer						43				1	
Normal breast	7						41				
Normal colon		6		2				161			
Normal liver		2	1		2				72		
Normal lung		1				2				41	
Total	73	164	30	34	46	47	45	161	73	45	718
Correct	65	154	29	32	44	43	41	161	72	41	682
False-positive							4		1	1	6
False-negative	7	9	1	2	2	2					23
Wrong tissue	1	1				2				4	8
Correct (%)	89.4	93.9	96.7	94.1	95.6	91.5	91.1	100	98.6	91.1	95.0

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

Methylation Profiles Predict Prognosis and Survival. 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

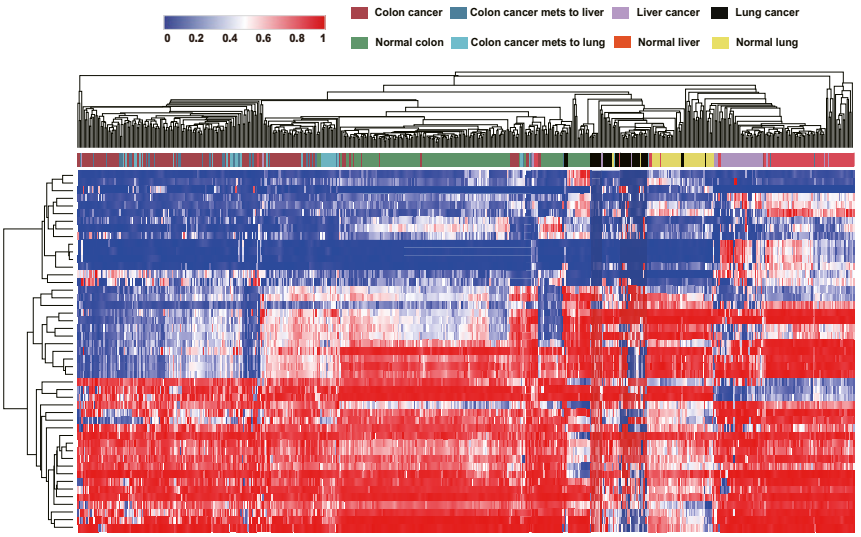

Fig. 2. DNA methylation signatures can identify the cancer of origin in metastasis of colon cancer. Shown are unsupervised hierarchical clustering and heat map associated with the methylation profile of 394 tumor samples and 324 normal samples of primary and metastatic colorectal cancer, liver cancer, and lung cancer in a Chinese cohort with a panel of 46 CpG markers. Each column represents an individual patient, and each row represents an individual CpG marker. The color scale shows relative methylation.

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.

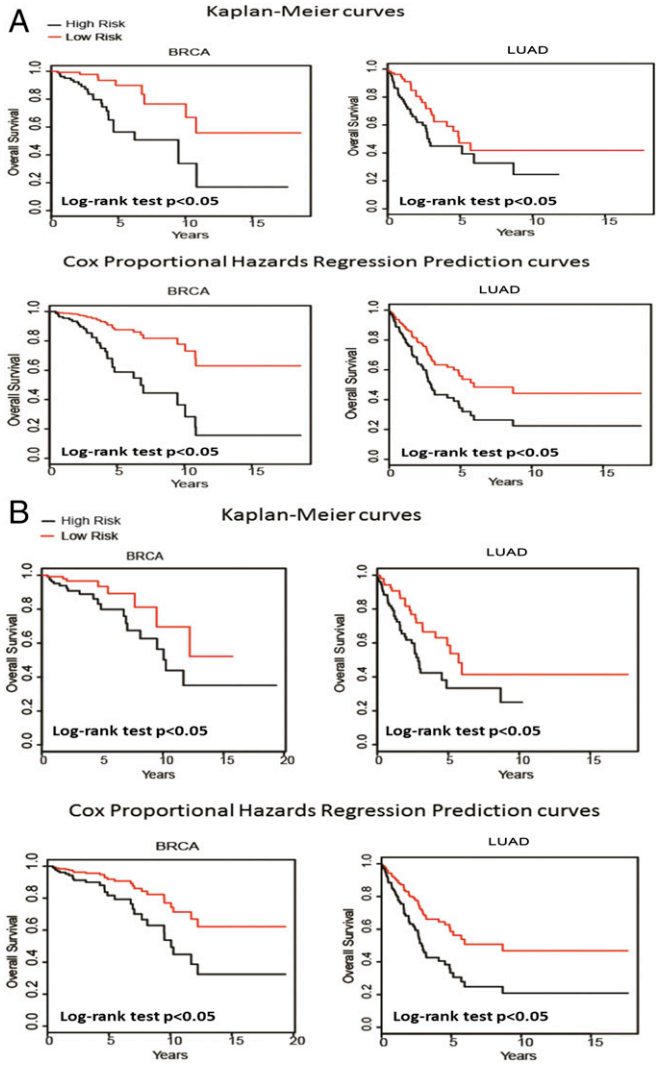

Fig. 3. Methylation markers can predict overall survival of patients in different types of cancers. (*A*) Overall survival curves of BRCA and LUAD patients with a low or high risk of death, according to a combined prognosis score from a lasso analysis. Shown are Kaplan–Meier curves (*Upper*) and Cox proportional hazards regression prediction curves (*Lower*) of overall survival in BRCA (*Left*) and LUAD (*Right*) patients with low or high risk of death. (*B*) Overall survival curves of BRCA and LUAD patients with a low or high risk of death, according to a combined prognosis score from a boosting analysis. Shown are Kaplan–Meier curves (*Upper*) and Cox proportional hazards regression prediction curves (*Lower*) of overall survival of BRCA (*Left*) and LUAD (*Right*) patients with low or high risk of death, according to a combined prognosis score from a boosting analysis.

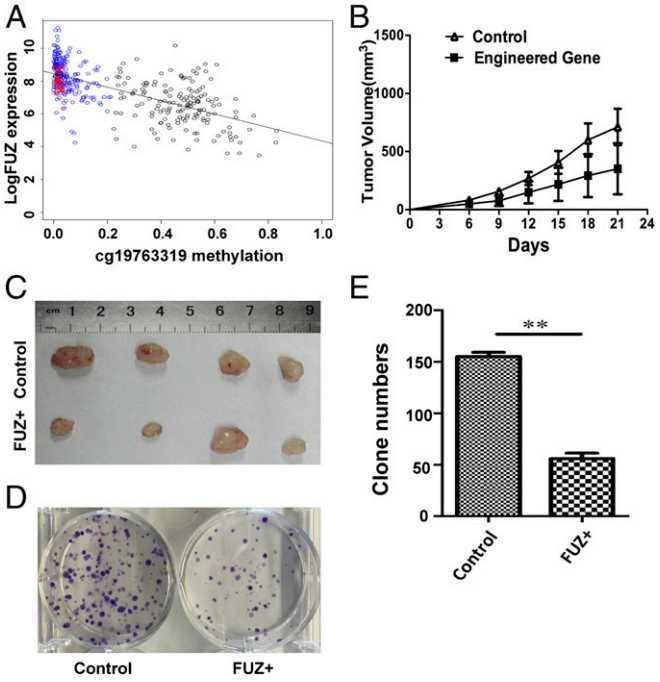

Fig. 4. Linking differentially methylated markers to gene expression in LIHC. (*A*) Relationship between methylation of CpG marker cg19763319 and expression of FUZ in liver cancer. Red dots indicate normal tissue samples; black dots, cancer samples. (*B*) Effect of FUZ expression on growth of liver cancer cell line HEP1. (*C*) Effect of FUZ expression on growth of HEP1 cells in a mouse xenograft model. (*D*) Effect of FUZ expression on colony formation of HEP1 cells. (*E*) Quantified colony formation by FUZ-transduced HEP1 cells compared with control. $^{**}P < 0.001$.

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 datasets 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 Annesser Fund, and the Michael Martin Fund.

1. DeVita VT, et al. (2011) *DeVita, Hellman, and Rosenberg's Cancer: Principles and Practice of Oncology* (Lippincott Williams & Wilkins, Philadelphia), Ed 9.
2. Wang T, et al. (2015) Identification and characterization of essential genes in the human genome. *Science* 350:1096–1101.
3. Han L, et al. (2014) The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun* 5:3963.
4. Akbani R, et al. (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nat Commun* 5:3887.
5. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet* 33:245–254.
6. Vaissière T, Sawan C, Herceg Z (2008) Epigenetic interplay between histone modifications and DNA methylation in gene silencing. *Mutat Res* 659:40–48.
7. Egger G, Liang G, Aparicio A, Jones PA (2004) Epigenetics in human disease and prospects for epigenetic therapy. *Nature* 429:457–463.
8. Herman JG, Baylin SB (2003) Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* 349:2042–2054.
9. Feinberg AP, Tycko B (2004) The history of cancer epigenetics. *Nat Rev Cancer* 4: 143–153.
10. Kandoth C, et al. (2013) Mutational landscape and significance across 12 major cancer types. *Nature* 502:333–339.
11. Paez JG, et al. (2004) EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science* 304:1497–1500.
12. Ogino S, et al.; Alliance for Clinical Trials in Oncology (2013) Predictive and prognostic analysis of PIK3CA Mutation in Stage III Colon Cancer Intergroup Trial. *J Natl Cancer Inst* 105:1789–1798.
13. Koboldt DC, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22:568–576.
14. Dees ND, et al. (2012) MuSiC: Identifying mutational significance in cancer genomes. *Genome Res* 22:1589–1598.
15. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:3.
16. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.
17. Yuan Y, et al. (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 32:644–652.
18. Lehmann-Werman R, et al. (2016) Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci USA* 113:E1826–E1834.