

5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages

Chun-Xiao Song^{1,13,*}, Senlin Yin^{2,3,*}, Li Ma^{4,5}, Amanda Wheeler⁵, Yu Chen², Yan Zhang⁶, Bin Liu^{6,7}, Junjie Xiong⁸, Weihang Zhang⁹, Jiankun Hu⁹, Zongguang Zhou⁹, Biao Dong², Zhiqi Tian¹⁰, Stefanie S Jeffrey⁵, Mei-Sze Chua^{4,5}, Samuel So^{4,5}, Weimin Li¹¹, Yuquan Wei², Jiajie Diao¹⁰, Dan Xie^{2,9,11}, Stephen R Quake^{1,12}

¹Departments of Bioengineering and Applied Physics, Stanford University, Stanford, CA 94305, USA; ²State Key Laboratory of Biotherapy, West China Hospital, Sichuan University and National Collaborative Innovation Center, Chengdu, Sichuan 610041, China; ³Department of Neurosurgery, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China; ⁴Asian Liver Center, Stanford University School of Medicine, Stanford, CA 94305, USA; ⁵Department of Surgery, Stanford University School of Medicine, Stanford, CA 94305, USA; ⁶Department of Thoracic Oncology, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China; ⁷Department of Pulmonary Tumor Ward, Sichuan Cancer Hospital, Chengdu, Sichuan 610041, China; ⁸Department of Pancreatic Surgery, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China; ⁹Department of Gastrointestinal Surgery, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China; ¹⁰Department of Cancer Biology, University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA; ¹¹Center of Precision Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China; ¹²Chan Zuckerberg Biohub, San Francisco, CA 94518, USA

5-Hydroxymethylcytosine (5hmC) is an important mammalian DNA epigenetic modification that has been linked to gene regulation and cancer pathogenesis. Here we explored the diagnostic potential of 5hmC in circulating cell-free DNA (cfDNA) using a sensitive chemical labeling-based low-input shotgun sequencing approach. We sequenced cell-free 5hmC from 49 patients of seven different cancer types and found distinct features that could be used to predict cancer types and stages with high accuracy. Specifically, we discovered that lung cancer leads to a progressive global loss of 5hmC in cfDNA, whereas hepatocellular carcinoma and pancreatic cancer lead to disease-specific changes in the cell-free hydroxymethylome. Our proof-of-principle results suggest that cell-free 5hmC signatures may potentially be used not only to identify cancer types but also to track tumor stage in some cancers.

Keywords: epigenetics; 5-hydroxymethylcytosine; cell-free DNA; cancer diagnostics

Cell Research (2017) 27:1231-1242. doi:10.1038/cr.2017.106; published online 18 August 2017

Introduction

Circulating cell-free DNA (cfDNA) are DNA fragments found in the blood that originate from cell death in different tissues; this phenomenon has formed the basis of noninvasive prenatal diagnostic tests [1], organ trans-

plant rejection diagnostics [2], and cancer detection [3]. Recent work has focused on the identification of 5-methylcytosine (5mC) modifications in cfDNA to characterize a variety of potential health conditions [3-8]. However, there has been no investigation to date of alternative epigenetic DNA modifications in cfDNA, due in part to the challenges of identifying and sequencing alternative modifications in low-input DNA samples.

5-Hydroxymethylcytosine (5hmC) is a recently identified epigenetic mark which impacts a broad range of biological processes ranging from development to pathogenesis [9, 10]. 5hmC is generated from 5mC by the ten-eleven translocation (TET) family dioxygenases [11]. Compared to the repressive effect of 5mC, 5hmC is generally believed to have a permissive effect on gene expression [12-15]. Unlike 5mC which is uniformly dis-

*These two authors contributed equally to this work.

Correspondence: Stephen R Quake^a, Dan Xie^b

^aTel: +1 650 736 7890

E-mail: quake@stanford.edu

^bE-mail: danxie@scu.edu.cn

¹³Current address: Ludwig Institute for Cancer Research and Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7FZ, UK

Received 17 July 2017; revised 23 July 2017; accepted 25 July 2017; published online 18 August 2017

tributed among different tissues in terms of total mass, 5hmC displays a tissue-specific mass distribution [16, 17] and low levels of 5hmC are also frequently observed in many solid tumors compared to corresponding normal tissues [18]. These characteristics suggest that 5hmC may have potential value in cancer diagnostics [10]. However, in contrast to the intensive studies on cell-free 5mC, cell-free 5hmC has remained unexploited, partly due to the low levels of 5hmC in the human genome (10-to 100-fold $< 5mC$) [17] and the lack of a sensitive low-input 5hmC DNA sequencing method that would work with the minute amounts of cfDNA available (typically only a few nanograms per ml of plasma). In this work, we developed a sensitive chemical labeling-based low-input whole-genome 5hmC sequencing method that allows rapid and reliable sequencing of 5hmC in cfDNA, and showed that cell-free 5hmC display distinct features in several types of cancer, which can potentially be used not only to identify cancer types but also to track tumor stage in some cancers.

Results

Development of cell-free 5hmC sequencing

We developed a low-input whole-genome cell-free 5hmC sequencing method based on selective chemical labeling (hMe-Seal) [13]. hMe-Seal is a robust method that uses β -glucosyltransferase (β GT) to selectively label 5hmC with a biotin *via* an azide-modified glucose for pull-down of 5hmC-containing DNA fragments for sequencing [13] (Supplementary information, Figure S1A). Standard hMe-Seal procedure requires micrograms of DNA. In our modified approach, cfDNA is first ligated with sequencing adapters and 5hmC is selectively labeled with a biotin group. After capturing cfDNA containing 5hmC using streptavidin beads, the final library is completed by PCR directly from the beads instead of eluting the captured DNA to minimize sample loss during purification steps (Figure 1A). With this approach we can sequence cell-free 5hmC readily from 1–10 ng of cfDNA. By utilizing a pool of 180 bp amplicons bearing C, 5mC or 5hmC spiked into cfDNA, we demonstrated that only 5hmC-containing DNA can be detected by PCR from the beads after pull-down (Supplementary information, Figure S1B). This result was confirmed in the final sequencing libraries, which showed over 100-fold enrichment in reads mapping to 5hmC spike-in DNA (Figure 1B). Furthermore, our approach performed equally well with cfDNA and bulk genomic DNA (1 μ g whole blood genomic DNA (gDNA)) (Figure 1B). The final cell-free 5hmC libraries are highly complex with a median unique nonduplicate map rate of 0.75 when lightly sequenced

(median 15 million reads, ~ 0.5 fold human genome coverage) (Supplementary information, Figure S1C, S1D and Table S1), and yet technical replicates are highly reproducible (Supplementary information, Figure S1E). We identified 5hmC-enriched regions (hMRs) in the sequence data using a Poisson-based method [19]. hMRs are highly concordant between technical replicates and a pooled sample: over 75% of hMRs in the pooled sample are in common with each of the replicates (Supplementary information, Figure S1F), reaching the ENCODE standard for ChIP-Seq [20]. These results demonstrate that cell-free 5hmC can be readily and reliably profiled by the modified hMe-Seal method.

Genome-wide mapping of 5hmC in cfDNA

We first sequenced cell-free 5hmC from eight healthy individuals (Supplementary information, Tables S1 and S2). We also sequenced 5hmC from whole blood gDNA from two of the individuals as blood is the major contributor to cell-free nucleic acids. Genome-scale profiles showed that the cell-free 5hmC distributions are nearly identical between healthy individuals and are clearly distinguishable from both the whole blood 5hmC distribution and the input cfDNA (Supplementary information, Figure S2A). Previous studies of 5hmC in mouse and human tissues showed that the majority of 5hmC resides in the gene bodies and promoter proximal regions of the genome [12, 14]. Genome-wide analysis of hMRs in our cfDNA data showed that the majority (80%) are intragenic with most enrichment in exons (observed to expected, $o/e = 7.29$), and depletion in intergenic regions ($o/e = 0.46$), consistent with that in whole blood (Supplementary information, Figure S2B and S2C) and in other tissues [12, 14]. The enrichment of 5hmC in gene bodies is known to be correlated with transcriptional activity in tissues such as the brain and liver [12–14]. To determine whether this relationship holds in cfDNA, we performed sequencing of the cell-free RNA from the same individual [21]. By dividing genes into three groups according to their cell-free RNA expression and plotting the average cell-free 5hmC profile along gene bodies (metagene analysis), we discovered an enrichment of 5hmC in and around gene bodies of more highly expressed genes (Figure 1C). These results demonstrate that cell-free 5hmC is derived from various tissue types and contains information from tissues other than the blood.

Since cell-free 5hmC were mostly enriched in the intragenic regions, we next used genic 5hmC fragments per kilobase of gene per million mapped reads (FPKM) to further compare the cell-free hydroxymethylome with the whole blood hydroxymethylome. Indeed, unbiased analysis of genic 5hmC using *t*-distributed stochastic

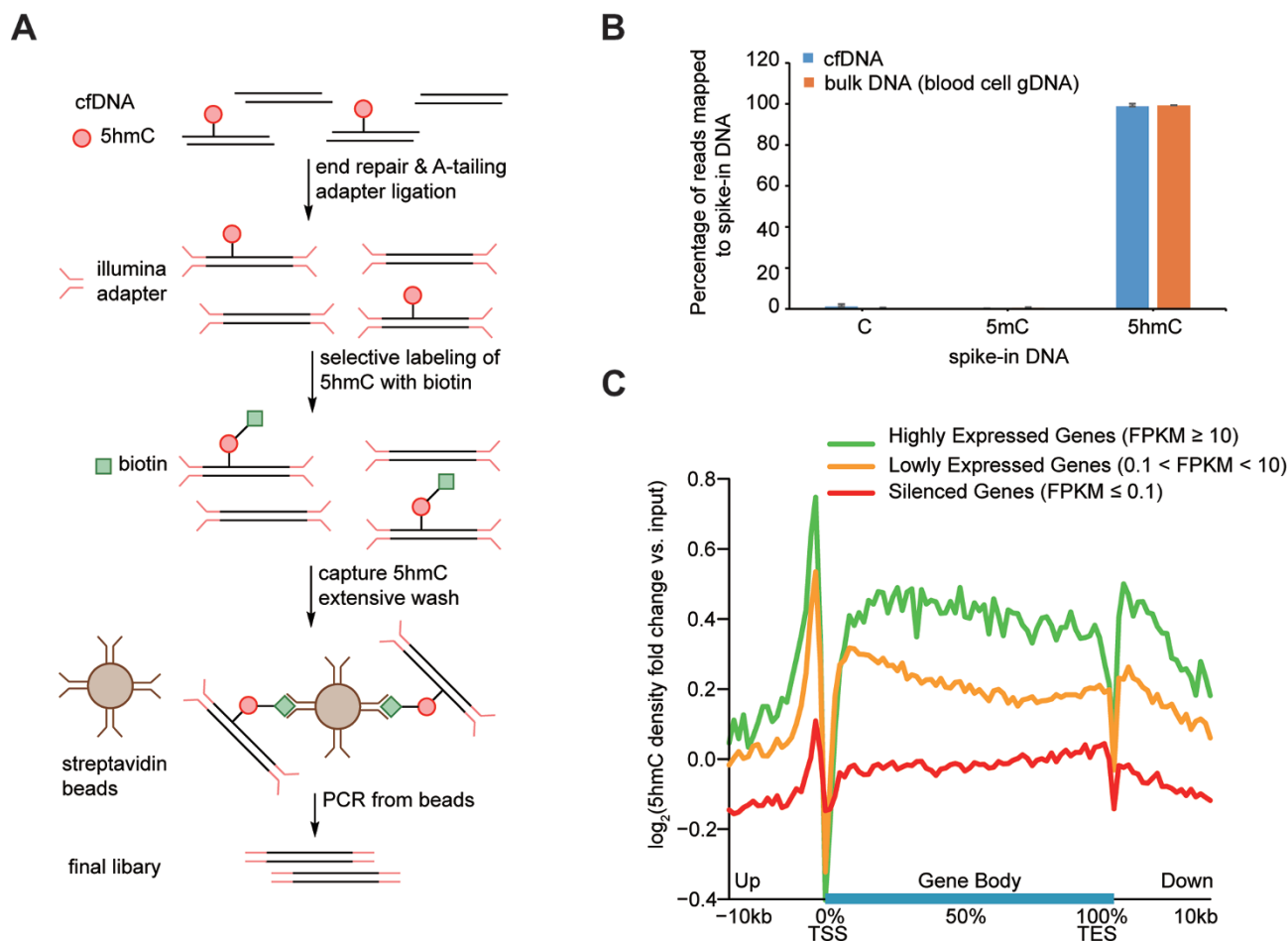


Figure 1 Sequencing of 5hmC in cfDNA. **(A)** General procedure of cell-free 5hmC sequencing. cfDNA is ligated with Illumina adapter and labeled with biotin on 5hmC for pull-down with streptavidin beads. The final library is completed by directly PCR from streptavidin beads. **(B)** Percentage of reads mapped to spike-in DNA in the sequencing libraries. Error bars indicate SD. **(C)** Metagene profiles of \log_2 fold change of cell-free 5hmC to input cfDNA ratio in genes ranked according to their expression in cell-free RNA-Seq.

neighbor embedding (*t*SNE) [22] showed strong separation between the cell-free and whole blood samples (Supplementary information, Figure S2D). We used the limma package [23] to identify 2 082 differentially hydroxymethylated genes between whole blood and cell-free samples (*q*-values (Benjamini and Hochberg adjusted *P*-values) < 0.01 , fold change > 2 , Supplementary information, Figure S3A). Notably, the 735 blood-specific 5hmC enriched genes showed increased expression in whole blood compared to the 1 347 cell-free-specific 5hmC enriched genes [24] (*P*-value $< 2.2 \times 10^{-16}$, Welch *t*-test) (Supplementary information, Figure S3B). In agreement with the differential expression, gene ontology (GO) analysis [25] of blood-specific 5hmC enriched genes mainly identified blood cell-related processes (Supplementary information, Figure S3C), whereas

cell-free-specific 5hmC enriched genes identified much more diverse biological processes (Supplementary information, Figure S3D). Examples of whole blood-specific (FPR1, FPR2) and cell-free-specific (GLP1R) 5hmC enriched genes are shown in Supplementary information, Figure S3E. Together, these results provide further evidence that a variety of tissues contribute 5hmC to cfDNA and that measurement of this is a rough proxy for gene expression.

Stage-dependent loss of 5hmC in lung cancer cfDNA

To explore the diagnostic potential of cell-free 5hmC, we applied our method to sequence cfDNA of a panel of 49 treatment-naïve primary cancer patients, including 15 lung cancer, 10 hepatocellular carcinoma (HCC), 7 pancreatic cancer, 4 glioblastoma (GBM), 5 gastric can-

cer, 4 colorectal cancer, 4 breast cancer patients (Supplementary information, Tables S3-S9). These patients vary from early stage cancer to late stage metastatic cancer. In lung cancer, we observed a progressive global loss of 5hmC enrichment from early stage non-metastatic lung cancer to late stage metastatic lung cancer compared to healthy cfDNA, and it gradually resembled that of the unenriched input cfDNA (Figure 2A). Unbiased gene body analysis using *t*SNE also showed a stage-dependent migration of the lung cancer profile from the healthy profile to one resembling the unenriched input cfDNA

(Supplementary information, Figure S4A). Notably, even the early stage lung cancer samples are highly separated from the healthy samples (Supplementary information, Figure S4A). We further confirmed the global hypohydroxymethylome events using other metrics. First, most differential genes in metastatic lung cancer (q -values $< 1e-7$, 1 159 genes) showed stage-dependent depletion of 5hmC compared to healthy samples (Figure 2B). Second, the metagene profile showed a stage-dependent depletion of gene body 5hmC signal and resemblance of the unenriched input cfDNA (Supplementary information, Figure

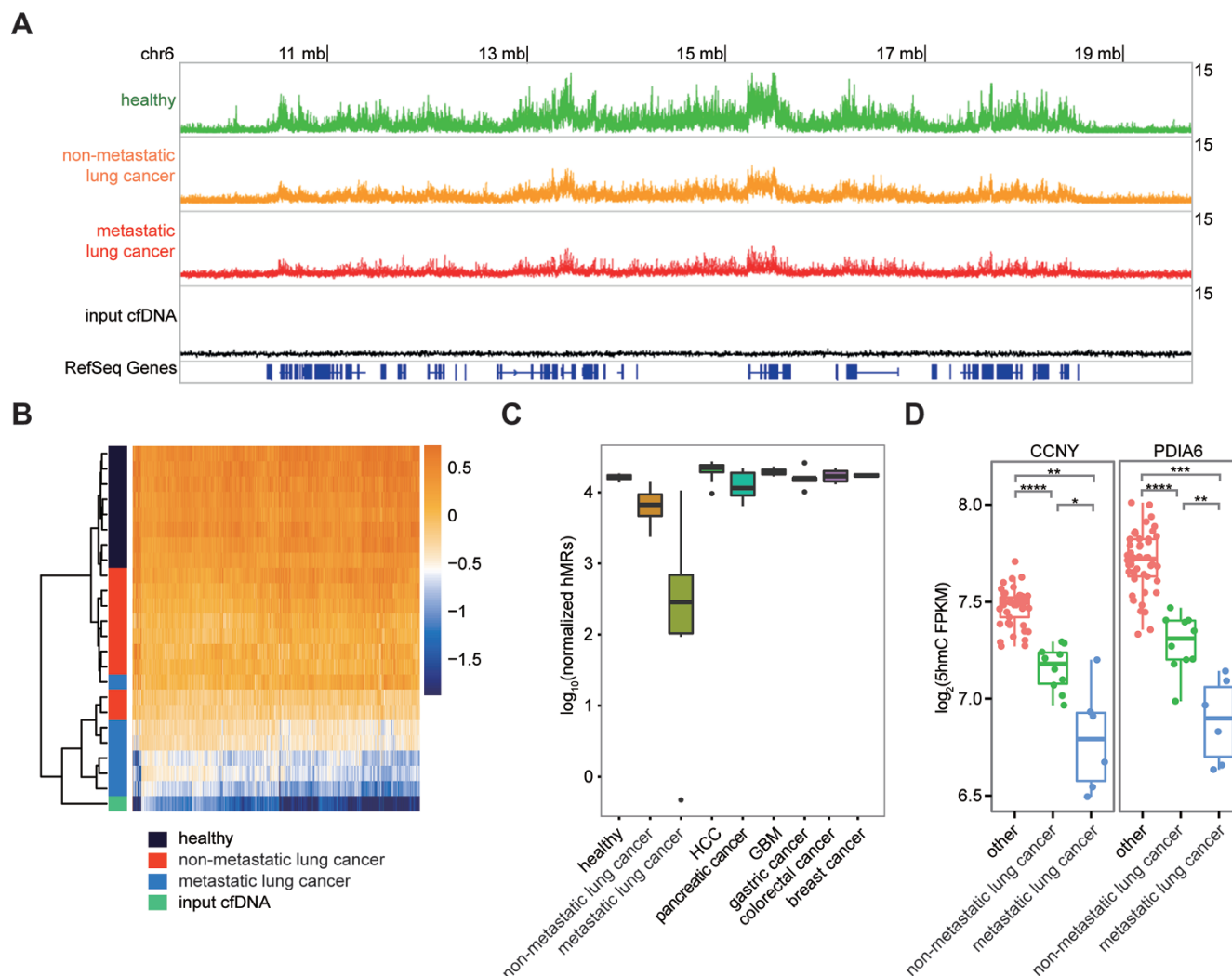


Figure 2 Lung cancer leads to progressive loss of 5hmC enrichment in cfDNA. **(A)** Genome browser view of the cell-free 5hmC distribution in a 10 mb region in chromosome 6. The overlapping tracks of healthy, non-metastatic lung cancer, metastatic lung cancer and input cfDNA samples are shown in line plot. **(B)** Heatmap of 1 159 metastatic lung cancer differential genes in healthy, lung cancer samples and the unenriched input cfDNA. Hierarchical clustering was performed across genes and samples. **(C)** Boxplot of number of hMRs (normalized to 1 million reads) identified in each group. **(D)** Boxplots of CCNY and PDIA6 5hmC FPKM in lung cancer and other cfDNA samples. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 1e-5$, Welch *t*-test.

S4B). Third, there is a dramatic decrease in the number of hMRs identified in lung cancer, especially in metastatic lung cancer compared to healthy and other cancer samples (Figure 2C). These data collectively indicate stage-dependent global loss of 5hmC levels in lung cancer cfDNA.

It should be noted that the global loss of 5hmC enrichment seen in lung cancer cfDNA is not due to the failure of our enrichment method, as the spike-in control in all samples including the lung cancer samples showed high enrichment of 5hmC-containing DNA (Supplementary information, Figure S4C). It is also a phenomenon unique to lung cancer that is not observed in other cancers we tested, evidenced by the number of hMRs (Figure 2C) and the metagene profiles (Supplementary information, Figure S4B). Examples of 5hmC depleted genes in lung cancer are shown in Figure 2D and Supplementary information, Figure S4D. Lung cancer tissue is known to have a low level of 5hmC compared to normal lung tissue [18], and lung has a relatively large contribution to cfDNA [21]. It is plausible that lung cancer, especially metastatic lung cancer, causes large quantities of hypohydroxymethylated gDNA to be released into cfDNA, effectively diluting the cfDNA and leading to the depletion of 5hmC in the cell-free 5hmC landscape. Alternatively or in combination, the cfDNA hypohydroxymethylation could originate from blood gDNA hypohydroxymethylation observed in metastatic lung cancer patients as recently reported [26]. Taken together these results indicate that cell-free 5hmC sequencing may potentially serve as a powerful tool for early lung cancer detection as well as monitoring lung cancer progression and metastasis.

Monitoring treatment and recurrence in HCC

For HCC, we also sequenced cell-free 5hmC from seven patients with hepatitis B (HBV) infection, since most HCC cases are secondary to viral hepatitis infections (Supplementary information, Table S4). Unbiased gene level analysis by *t*SNE revealed that HCC patients in general can be separated from the HBV-infected patients and healthy individuals based on the cell-free 5hmC patterns (Figure 3A). HCC-specific differential genes (q -values < 0.001 , fold change > 1.41 , 1 006 genes) could separate HCC from healthy and most of the HBV samples (Figure 3B). Both HCC specifically enriched and depleted genes can be identified compared to other cfDNA samples (Figure 3B), and the enriched genes (379 genes) showed increased expression in liver tissue compared to the depleted genes (637 genes) [24] (P -values $< 2.2 \times 10^{-16}$, Welch *t*-test) (Supplementary information, Figure S5A), consistent with the permissive effect of 5hmC on gene expression. An example of HCC-specific 5hmC en-

riched genes is AHSR, a secreted protein highly expressed in the liver [24] (Figure 3C; Supplementary information, Figure S5B and S5C), and an example of HCC-specific 5hmC depleted genes is TET2, one of the enzymes that generate 5hmC and a tumor suppressor down-regulated in HCC [27] (Figure 3D; Supplementary information, Figure S5D). Together, these results point to a model where virus infection and HCC development lead to a gradual damage of liver tissue and increased presentation of liver DNA in the blood.

To further explore the potential of cell-free 5hmC for monitoring treatment and disease progression, we followed four of the HCC patients who underwent surgical resection, out of which three of them had recurrent disease (Supplementary information, Table S4). Analysis of serial plasma samples from these patients (pre-operation/pre-op; post-operation/post-op; and recurrence) with *t*SNE revealed that post-op samples clustered with healthy samples, whereas the recurrence samples clustered with HCC (Figure 3E). This pattern was also reflected by changes in the 5hmC FPKM of AHSR and TET2 (Figure 3C and 3D). As an example of using cell-free 5hmC for tracking HCC treatment and progression, we employed linear discriminant analysis (LDA) to define a linear combination of the HCC-specific differential genes (Figure 3D) into a single value (the HCC score) that best separated the pre-op HCC samples from the healthy and HBV samples. We then calculated the HCC score for the post-op and recurrence HCC samples, and showed that the HCC score can accurately track the treatment and recurrence states (Supplementary information, Figure S5E). Together, these results indicate that cell-free 5hmC sequencing presents an opportunity to detect HCC, as well as monitor treatment outcome and disease recurrence.

Pancreatic cancer impacts the cell-free 5hmC

We also found pancreatic cancer produced drastic changes in its cell-free hydroxymethylome, even in some early stage pancreatic cancer patients (Supplementary information, Table S5). Like HCC, pancreatic cancer lead to both up-regulated and down-regulated 5hmC genes compared to healthy individuals (q -value < 0.01 , fold change > 2 , 713 genes) (Supplementary information, Figure S6A). Examples of pancreatic cancer-specific 5hmC-enriched and depleted genes compared to other cfDNA samples are shown in Supplementary information, Figure S6B-S6E. Our results suggest that cell-free 5hmC sequencing can be potentially valuable for early detection of pancreatic cancer.

Copy number variation estimation

Copy number variation (CNV) can be detected from

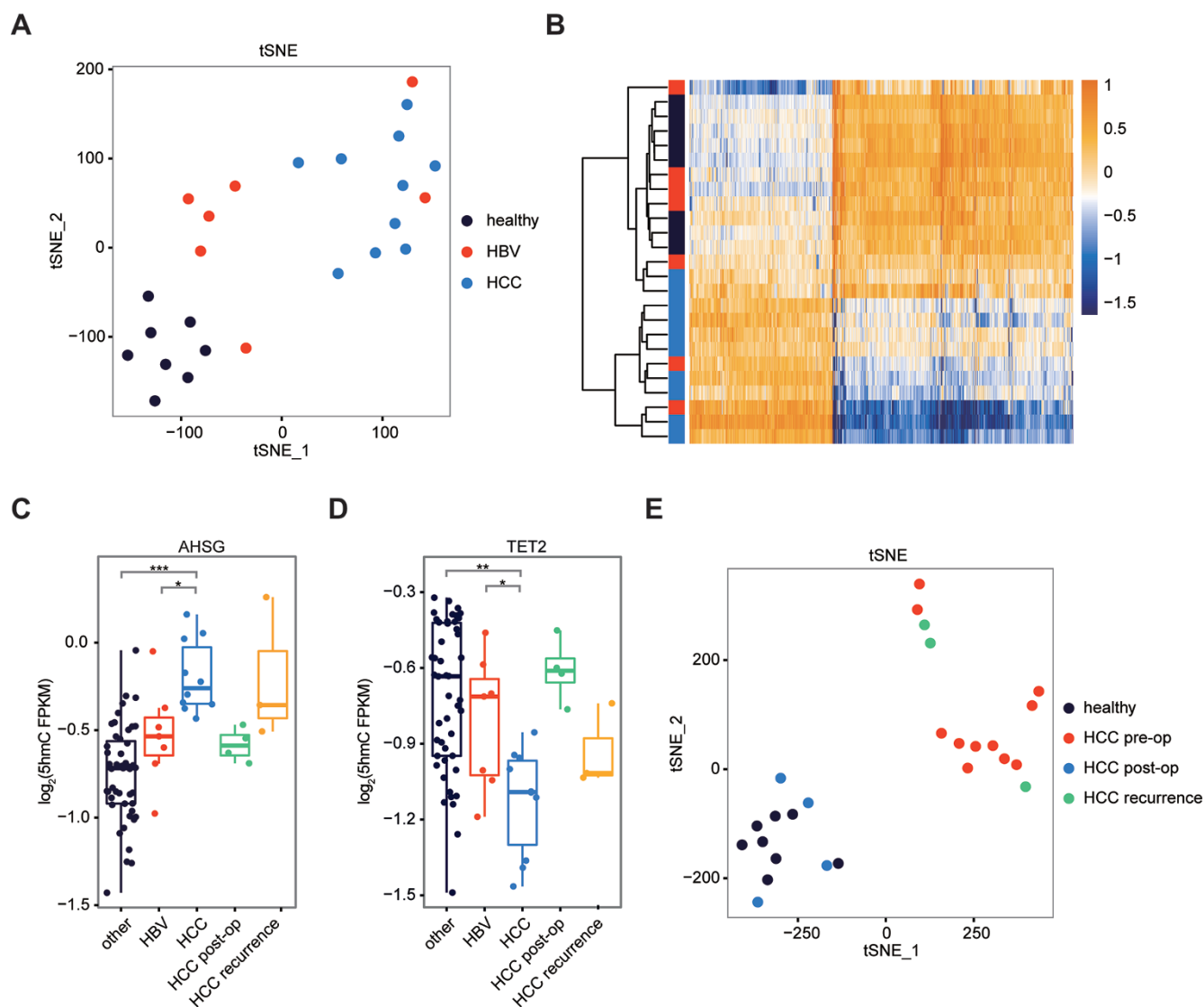


Figure 3 Cell-free 5hmC for monitoring HCC progression and treatment. **(A)** tSNE plot of 5hmC FPKM from healthy, HBV and HCC samples. **(B)** Heatmap of 1 006 HCC differential genes in healthy, HBV and HCC samples. Hierarchical clustering was performed across genes and samples. **(C, D)** Boxplots of AHSG **(C)** and TET2 **(D)** 5hmC FPKM in HBV, HCC (pre-op), HCC post-op, HCC recurrence and other cfDNA samples. * $P < 0.05$, ** $P < 1e-4$, *** $P < 1e-5$, Welch t -test. **(E)** tSNE plot of 5hmC FPKM from healthy, HCC pre-op, HCC post-op and HCC recurrence samples.

cfDNA sequencing, mostly in advanced cancer patients, which provides a way to assess the tumor burden in the cfDNA [3]. To assess the tumor burden in our samples and to explore the relation between CNV contained from unenriched input cfDNA sequencing and the 5hmC enrichment sequencing, we also sequenced the input cfDNA in 47 samples (Supplementary information, Table S10). We analyzed the CNV from these input cfDNA sequencing with 1 mb bin [28], and as expected we could detect large-scale CNV from about 20% of the cancer samples, mostly in late stage cancer samples (Supple-

mentary information, Figure S7A). We then analyzed the CNV from the corresponding 5hmC enrichment sequencing and interestingly, we found matched CNV patterns in several cases (Supplementary information, Figure S7A). For example we could detect chromosome wise CNV in lung293 and lung417, two metastatic lung cancer samples, from input cfDNA sequencing (Supplementary information, Figure S7B and S7C). These samples displayed large-scale cell-free 5hmC changes and correspondingly, the CNV patterns detected from the 5hmC enrichment sequencing mimic the CNV patterns detected

from input cfDNA sequencing (Supplementary information, Figure S7D and S7E). This result supports the notion that 5hmC enriched cfDNA contains significant portion of tumor-derived cfDNA and therefore represents 5hmC patterns in tumor cells. It also shows that 5hmC sequencing and CNV analysis could complement each other in circulating tumor DNA analysis.

Cancer type and stage prediction

Although there has been great interest in using cfDNA as a “liquid biopsy” for cancer detection, it has been challenging to identify the origin of tumor cfDNA and hence the location of the tumor. We discovered from *t*SNE analysis of all seven cancer types that lung cancer, HCC and pancreatic cancer showed distinct signatures

and could be readily separated from each other and healthy samples (Figure 4A). The other four types of cancer displayed relatively minor changes compared to the healthy samples. Using other features such as the promoter region (5 kb upstream of the transcription start site) showed similar patterns (Supplementary information, Figure S8A). We note that no particular cancer type we tested resembled the whole blood profile (Supplementary information, Figure S8B), suggesting that the blood cell contamination is not a significant source of variation. All patients in our panel fall in the same age range as the healthy individuals (Supplementary information, Figure S8C and Tables S2-S9), therefore age is unlikely to be a confounding factor. We also did not observe any batch effect (Supplementary information, Figure S8D).

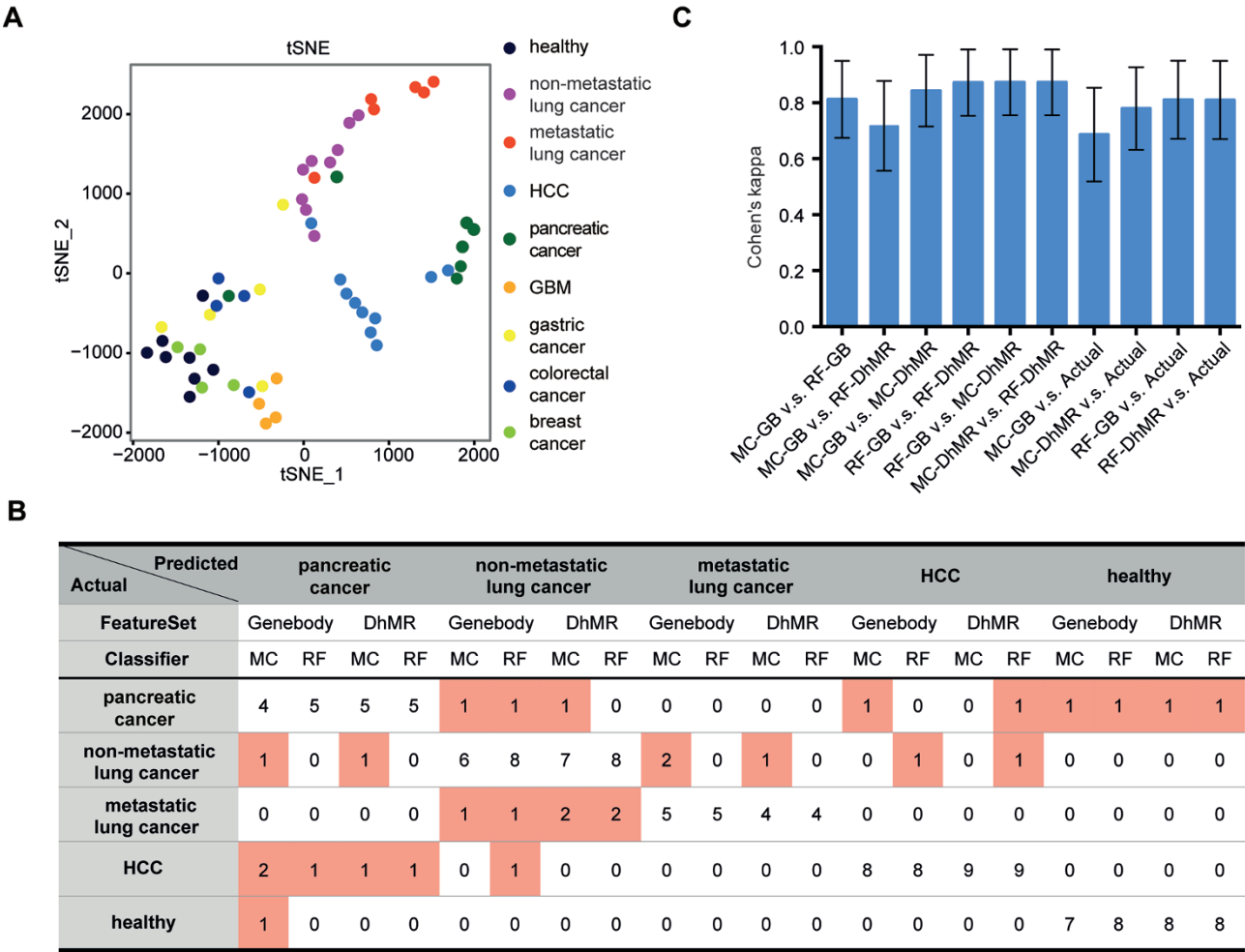


Figure 4 Cancer type and stage prediction with cell-free 5hmC. **(A)** *t*SNE plot of 5hmC FPKM in cfDNA from healthy and various cancer samples. **(B)** The actual and predicted classification by leave-one-out cross-validation using Mclust (MC) and Random Forest (RF) algorithm, based on two feature sets (gene body and DhMR). **(C)** The Cohen's kappa coefficient for measuring inter-classifier agreement (GB for gene body). The error bar indicates 95% confidence interval of the Cohen's kappa estimate.

To further demonstrate the potential of cfDNA 5hmC as a biomarker to predict cancer types, we employed two widely used machine learning methods, the Gaussian mixture model [29] and Random Forest [30]. We focused on the prediction of HCC, pancreatic cancer, non-metastatic and metastatic lung cancer. On the basis of three rules (see Materials and Methods section), we identified genes (Supplementary information, Table S11) whose average gene body 5hmC levels could either distinguish cancer groups from healthy groups or between cancer groups. In addition to using gene body data, the 5hmC on non-coding regions could also potentially serve as a biomarker in predicting cancer types [9]. We therefore designed another set of features by investigating each of the 2 kb windows of the entire genome and identified differential hMRs (DhMRs) for each cancer type (see Materials and Methods section; Supplementary information, Table S12). We trained the two machine learning algorithms using either differential 5hmC genes or DhMRs as features and evaluated the leave-one-out (LOO) cross-validation prediction accuracy. The Gaussian mixture model-based predictor (Mclust) had overall successful prediction rates of 75% and 82.5%, when using gene body and DhMRs as features, respectively (Figure 4B; Supplementary information, Figure S9A and S9B). Mclust-based dimensional reduction showed clear boundaries between the groups (Supplementary information, Figure S9C). When only the type of the cancer is considered, Mclust predictors had higher success rate of 82.5% and 90% when using these two feature sets. The Random Forest predictor achieved LOO cross-validation prediction accuracy of 85%, when using either gene body or DhMRs as features (Figure 4B). When only cancer type is considered, Random Forest predictor achieved 87.5% and 90% prediction accuracy, with gene body and DhMRs as features, respectively. Distinct 5hmC profiles in different cancer types of several DhMRs with high variable importance to random forest prediction model could be observed (Supplementary information, Figures S9D, S9E and S10). Finally, we used Cohen's kappa to evaluate the concordance rate between different prediction models [31]. All combinations showed high agreement (Cohen's kappa ~ 0.8) in inter-classifier comparison and when comparing with the actual classification (Figure 4C). These results support the prospects of using cell-free 5hmC for cancer diagnostics and staging.

Discussion

Recent studies have reported that 5hmC is an important component of the mammalian genome [9, 32]. In this study, we reported an improved hMe-Seal [13] approach

to sequence the low levels of 5hmC in cfDNA, which offers several notable advantages. First, unlike traditional bisulfite sequencing used for cell-free 5mC sequencing, our method does not further degrade the highly fragmented cfDNA. Second, compared to whole genome approaches including mutational sequencing and bisulfite sequencing, the enrichment for 5hmC not only enables cost-effective sequencing (10–20 million reads, ~ 0.5 -fold human genome coverage), but more importantly allows the low-frequency tissue contribution of 5hmC in cfDNA to be amplified from the dominant blood cell contribution in cfDNA.

We sequenced cell-free 5hmC from a panel of seven cancer types and focused our analysis on lung cancer, HCC and pancreatic cancer, the three cancers which displayed the most dramatic impact on the cell-free hydroxymethylome, even in the early stages. Lung and liver are reported to have relatively large contribution to cfDNA [5, 21], and pancreatic cancer is known to invade progressively to the lymph nodes and liver during early stages without remarkable symptoms, which may explain their large impact on the cell-free hydroxymethylome. In lung cancer, we observed a characteristic stage-dependent global loss of cell-free 5hmC enrichment, whereas in HCC and pancreatic cancer, we identified significant finer scale changes of cell-free 5hmC (i.e., gene body and DhMR). In HCC, we also conducted an exploratory study of longitudinal samples whose results suggest that cell-free 5hmC may be used to monitor treatment and recurrence. Further studies will help elucidate how each cancer causes specific changes in the cell-free hydroxymethylome. Importantly, these three types of cancer displayed distinct patterns in their cell-free hydroxymethylome and we could employ machine learning algorithms trained with cell-free 5hmC features to predict the three cancer types with high accuracy.

In summary, we report the first proof-of-principle global analysis of hydroxymethylome in cfDNA. Large-scale clinical trials are required to fully validate the usefulness and understand potential limitations of this approach. Cell-free 5hmC contributes a new dimension of information to liquid biopsy-based diagnosis and prognosis; and we anticipate it may become a valuable tool for cancer diagnostics, as well as potentially for other disease areas, including but not limited to neurodegenerative diseases, cardiovascular diseases and diabetes. We envisage this strategy could be readily combined with other genetic and epigenetic-based cfDNA approaches (e.g., CNV analysis as we demonstrated) for increased diagnostic power. Our method represents the first enrichment-based genome-wide approach applied to cfDNA. The general framework of this method can be readily

adopted to sequence other modifications in cell-free nucleic acids by applying the appropriate labeling chemistry to the modified bases. This would allow a comprehensive and global overview of genetic and epigenetic changes of various disease states, and further increase the power of personalized diagnostics.

Materials and Methods

Study design

The overall goal of this study was to explore the diagnostic potential of 5hmC cfDNA for cancer detection. The objective of the first portion of the study was to determine whether 5hmC can be sequenced from cfDNA using an enrichment-based method. The objective of the second portion of the study was to determine whether cell-free 5hmC contains information that can be used for cancer diagnostics. Samples for healthy subjects were obtained from Stanford blood center. HCC and breast cancer patients were recruited in a Stanford University Institutional Review Board-approved protocol. Lung cancer, pancreatic cancer, GBM, gastric cancer and colorectal cancer patients were recruited in a West China Hospital Institutional Review Board-approved protocol. All recruited subjects gave informed consent. No statistical methods were used to pre-determine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. No samples were excluded from the analysis.

Clinical sample collection and processing

Blood was collected into EDTA-coated Vacutainers. Plasma was collected from the blood samples after centrifugation at $1\,600\times g$ for 10 min at $4\text{ }^{\circ}\text{C}$ and $16\,000\times g$ at 10 min at $4\text{ }^{\circ}\text{C}$. cfDNA was extracted using the Circulating Nucleic Acid Kit (Qiagen). Whole blood genomic DNA was extracted using the DNA Mini Kit (Qiagen) and fragmented using dsDNA Fragmentase (NEB) into average 300 bp. DNA was quantified by Qubit Fluorometer (Life Technologies). Cell-free RNA was extracted using the Plasma/Serum Circulating and Exosomal RNA Purification Kit (Norgen). The extracted cell-free RNA was further digested using Baseline-ZERO DNases (Epicentre) and depleted using Ribo-Zero rRNA Removal Kit (Epicentre) according to a protocol from Clontech.

Spike-in amplicon preparation

To generate the spiked-in control, lambda DNA was PCR amplified by Taq DNA Polymerase (NEB) and purified by AMPure XP beads (Beckman Coulter) in nonoverlapping ~ 180 bp amplicons, with a cocktail of dATP/dGTP/dTTP and one of the following: dCTP, dmCTP or 10% dhmCTP (Zymo)/90% dCTP. Primers sequences are as follows: dCTP FW-5'-CGTTTCCGTTCTTCTTC-GTC-3', RV-5'-TACTCGCACCGAAAATGTCA-3'; dmCTP FW-5'-GTGGCGGGTTATGATGAACT-3', RV-5'-CATAAAATGC-GGGGATTAC-3'; 10% dhmCTP/90% dCTP FW-5'-TGAAAAC-GAAAGGGGATACG-3', RV-5'-GTCCAGCTGGGAGTCGA-TAC-3'.

5hmC library construction, labeling, capture and high-throughput sequencing

cfDNA (1–10 ng) or fragmented whole blood genomic DNA (1 μg) spiked with amplicons (0.01 pg of each amplicon per 10

ng DNA) was end repaired, 3'-adenylated and ligated to DNA Barcodes (Bioo Scientific) using KAPA Hyper Prep Kit (Kapa Biosystems) according to the manufacturer's instructions. Ligated DNA was incubated in a 25 μl solution containing 50 mM HEPES buffer (pH 8), 25 mM MgCl_2 , 60 μM UDP-6- N_3 -Glc (Active Motif) and 12.5 U βGT (Thermo) for 2 h at $37\text{ }^{\circ}\text{C}$. After that, 2.5 μl DBCO-PEG4-biotin (Click Chemistry Tools, 20 mM stock in DMSO) was directly added to the reaction mixture and incubated for 2 h at $37\text{ }^{\circ}\text{C}$. Next, 10 μg sheared salmon sperm DNA (Life Technologies) was added into the reaction mixture and the DNA was purified by Micro Bio-Spin 30 Column (Bio-Rad). The purified DNA was incubated with 0.5 μl M270 Streptavidin beads (Life Technologies) pre-blocked with salmon sperm DNA in buffer 1 (5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl and 0.2% Tween 20) for 30 min. The beads were subsequently undergone three 5-min washes each with buffer 1, buffer 2 (buffer 1 without NaCl), buffer 3 (buffer 1 with pH 9) and buffer 4 (buffer 3 without NaCl). All binding and washing were done at room temperature with gentle rotation. Beads were then resuspended in water and amplified with 14 (cfDNA) or 9 (whole blood genomic DNA) cycles of PCR amplification using Phusion DNA polymerase (NEB). The PCR products were purified using AMPure XP beads. Separate input libraries were made by direct PCR from ligated DNA without labeling and capture. For technical replicates, cfDNA from the same subject was divided into two technical replicates. Pair-end 75 bp sequencing was performed on the NextSeq instrument.

Data processing and gene body analysis

FASTQ sequences were aligned to UCSC/hg19 with Bowtie2 v2.2.5 [33] and further filtered with samtools-0.1.19 [34] (parameters used: samtools view -f 2 -F 1548 -q 30 and samtools rmdup) to retain unique non-duplicate matches to the genome. Pair-end reads were extended and converted into bedgraph format normalized to the total number of aligned reads using bedtools [35], and then converted to bigwig format using bedGraphToBigWig from the UCSC Genome Browser for visualization in Integrated Genomics Viewer [36, 37]. FASTQ sequences were also aligned to the three spike-in control sequences to evaluate the pull-down efficiency. The spike-in control is only used as a validation of successful pull-down in each sample. hMRs were identified with MACS [19] using unenriched input DNA as background and default setting (P -value cutoff $1e-5$). Genomic annotations of hMRs were performed by determining the percentage of hMRs overlapping each genomic regions ≥ 1 bp. Metagenome profile was generated using ngs.plot [38]. 5hmC FPKM were calculated using the fragment counts in each RefSeq gene body obtained by bedtools. For differential analyses, genes shorter than 1 kb or mapped to chromosome X and Y were excluded. Differential genic 5hmC analysis was performed using the limma package in R [23]. GO analyses were performed using DAVID Bioinformatics Resources 6.7 with GOTERM_BP_FAT [25, 39]. Tissue-specific gene expression was obtained from BioGPS [24, 40, 41]. For $t\text{SNE}$ plot, the Pearson correlation of gene body 5hmC FPKM was used as the distance matrix to $t\text{SNE}$. MA-plot, hierarchical clustering, $t\text{SNE}$, LDA and heatmaps were done in R.

Cell-free RNA library construction and high-throughput sequencing

Cell-free RNA library was prepared using ScriptSeq v2 RNA-

Seq Library Preparation Kit (Epicentre) following the FFPE RNA protocol with 19 cycles of PCR amplification. The PCR products were then purified using AMPure XP beads. Pair-end 75 bp sequencing was performed on the NextSeq instrument. RNA-seq reads were first trimmed using Trimmomatic-0.33 [42] and then aligned using tophat-2.0.14 [43]. RPKM expression values were extracted using cufflinks-2.2.1 [44] using RefSeq gene models.

CNV estimation

The hg19 human genome was split into 1 mb bin and bin counts were generated using bedtools intersect. Mappability score of each bin was then assessed by average mappability using mappability track of hg19 from UCSC (kmer = 75). Bins with mappability score under 0.8 were eliminated from further analysis. GC content percentages in each bin were calculated using getGC.hg19 from R package PopSV (1.0.0) and GC bias was corrected by fitting a LOESS model (correct.GC from R package PopSV 1.0.0). The corrected bin counts were then scaled by mean bin count of each sample and centered at 2. For estimation of CNV, we cutoff the corrected bin counts higher than 5 to minimize impact of extreme values. Moving averages with window size of 20 mb were then calculated within each chromosome, as the final estimation of CNV.

Cancer type and stage prediction

Lung cancer, pancreatic cancer, HCC and healthy samples ($n = 40$) were included in the following analyses and LOO cross-validation was performed. With each iteration of LOO, one sample was left out first, and the remaining 39 samples were used for feature selection and as a training data set. The left out sample was then used to test the prediction accuracy of the machine learning model. Two types of feature selection were performed as independent analysis. In the “gene body” approach, cancer type-specific marker genes were selected by performing a Student’s t -test between (1) one cancer group and healthy group, (2) one cancer group and other cancer samples, (3) two different cancer groups. Benjamini and Hochberg correction was then performed for the raw P -value and the genes were then sorted by q -value. The top five genes with smallest q -value from each of these comparisons were selected as feature set to train the classifier. The second approach to finding features (“DhMR”) attempted to achieve higher resolution by first breaking the reference genome (hg19) into 2kb windows *in silico* and then calculating 5hmC FPKM value for each of the window. Blacklisted genomic regions that tend to show artifact signal according to ENCODE were filtered before down-stream analysis [45]. For cancer type-specific DhMRs, Student’s t -test and Benjamini and Hochberg correction of P -values were performed for comparison pairs same as previously performed for identifying cancer-specific genes. The top five DhMRs with smallest q -value from each comparison were chosen for each cancer type. Random forest and Gaussian model-based Mclust classifier were performed on the data set using previously described features (gene bodies and DhMRs). Classifiers were trained on lung cancer, pancreatic cancer, HCC and healthy samples. The same random seed (seed = 5) was used in every random forest analysis for consistency. The top 15 features shared by at least 30 LOO iterations with the highest mean decrease Gini with the highest variable importance were plotted. Gaussian mixture model analysis was performed using Mclust R package [29]. For Mclust model-based classifier training,

a Bayesian information criterion (BIC) plot was performed for visualization of the classification efficacy of different multivariate mixture models. By default, the EEI model (diagonal, equal volume and shape) or VII model (spherical, unequal volume) with EDDA model-type (single component for each class with the same covariance structure among classes) were chosen for Mclust classification. Cohen’s kappa was then calculated for assessment of the interclassifier concordance.

Statistical analysis

We used unpaired two-tailed t -tests (Welch t -test) for normally distributed data in which two comparison groups were involved. In the case of multiple comparisons, Benjamini and Hochberg correction was then performed for the raw P -value to obtain the q -value. Random forest and Gaussian model-based Mclust were used as machine classifiers. Cohen’s kappa was used for evaluating the predictive value of cell-free DNA 5hmC sequencing and interclassifier concordance. t SNE was used for dimension reduction and visualization. Statistical analyses were performed in R 3.3.2.

All sequencing data were deposited in the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) under accession number GSE81314.

Acknowledgments

We would like to acknowledge N Neff and G Mantalas for high-throughput sequencing; L Penland and J Beausang for sample collection; other members of the Quake and Xie Labs for discussions and support; R Altman and W Zhou for critical reading of the manuscript. This work was supported by the National Natural Science Foundation of China (31571327 and 91631111 to DX), National Institutes of Health (U01 CA154209 to SRQ) and Department of Defense (W81XWH110287 to SRQ).

Author Contributions

C-XS, JD, DX and SRQ conceived the study and designed the experiments. C-XS performed the experiments with the help from LM, YC and BD. C-XS analyzed data with help from SY, ZT and DX. LM, AW, YZ, BL, JX, WZ, JH, ZZ, SSJ, M-SC, SS, WL and YW recruited patients, collected blood and organized clinical information. C-XS and SRQ wrote the manuscript with input and comments from SY, DX and M-SC.

Competing Financial Interests

Stanford has filed patent applications based on this work. C-XS and SRQ are founders of Bluestar Genomics.

References

- 1 Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* 2008; **105**:16266-16271.
- 2 Snyder TM, Khush KK, Valentine HA, Quake SR. Universal noninvasive detection of solid organ transplant rejection. *Proc Natl Acad Sci USA* 2011; **108**:6229-6234.
- 3 Wan JC, Massie C, Garcia-Corbacho J, *et al.* Liquid biopsies come of age: towards implementation of circulating tumour

- DNA. *Nat Rev Cancer* 2017; **17**:223-238.
- 4 Chan KC, Jiang P, Chan CW, *et al.* Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci USA* 2013; **110**:18761-18768.
- 5 Sun K, Jiang P, Chan KC, *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci USA* 2015; **112**:E5503-E5512.
- 6 Wen L, Li J, Guo H, *et al.* Genome-scale detection of hypermethylated CpG islands in circulating cell-free DNA of hepatocellular carcinoma patients. *Cell Res* 2015; **25**:1250-1264.
- 7 Lehmann-Werman R, Neiman D, Zemmour H, *et al.* Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci USA* 2016; **113**:E1826-E1834.
- 8 Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet* 2017; **49**:635-642.
- 9 Shen L, Song CX, He C, Zhang Y. Mechanism and function of oxidative reversal of DNA and RNA methylation. *Annu Rev Biochem* 2014; **83**:585-614.
- 10 Vasanthakumar A, Godley LA. 5-hydroxymethylcytosine in cancer: significance in diagnosis and therapy. *Cancer Genet* 2015; **208**:167-177.
- 11 Tahiliani M, Koh KP, Shen Y, *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 2009; **324**:930-935.
- 12 Mellen M, Ayata P, Dewell S, Kriaucionis S, Heintz N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* 2012; **151**:1417-1430.
- 13 Song CX, Szulwach KE, Fu Y, *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* 2011; **29**:68-72.
- 14 Thomson JP, Lempiainen H, Hackett JA, *et al.* Non-genotoxic carcinogen exposure induces defined changes in the 5-hydroxymethylome. *Genome Biol* 2012; **13**:R93.
- 15 Feng J, Shao N, Szulwach KE, *et al.* Role of Tet1 and 5-hydroxymethylcytosine in cocaine action. *Nat Neurosci* 2015; **18**:536-544.
- 16 Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 2009; **324**:929-930.
- 17 Globisch D, Munzel M, Muller M, *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS ONE* 2010; **5**:e15367.
- 18 Jin S-G, Jiang Y, Qiu R, *et al.* 5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations. *Cancer Res* 2011; **71**:7360-7365.
- 19 Zhang Y, Liu T, Meyer CA, *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; **9**:R137.
- 20 Landt SG, Marinov GK, Kundaje A, *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012; **22**:1813-1831.
- 21 Koh W, Pan W, Gawad C, *et al.* Noninvasive *in vivo* monitoring of tissue-specific global gene expression in humans. *Proc Natl Acad Sci USA* 2014; **111**:7361-7366.
- 22 van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**:2579-2605.
- 23 Ritchie ME, Phipson B, Wu D, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43**:e47.
- 24 Su AI, Wiltshire T, Batalov S, *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 2004; **101**:6062-6067.
- 25 Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**:44-57.
- 26 Chowdhury B, Cho IH, Hahn N, Irudayaraj J. Quantification of 5-methylcytosine, 5-hydroxymethylcytosine and 5-carboxylcytosine from the blood of cancer patients by an enzyme-based immunoassay. *Anal Chim Acta* 2014; **852**:212-217.
- 27 Sajadian SO, Ehnert S, Vakilian H, *et al.* Induction of active demethylation and 5hmC formation by 5-azacytidine is TET2 dependent and suggests new treatment strategies against hepatocellular carcinoma. *Clin Epigenetics* 2015; **7**:98.
- 28 Xu H, Zhu X, Xu Z, *et al.* Non-invasive analysis of genomic copy number variation in patients with hepatocellular carcinoma by next generation DNA Sequencing. *J Cancer* 2015; **6**:247-253.
- 29 Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002; **97**:611-631.
- 30 Liaw A, Wiener M. Classification and regression by random-Forest. *R News* 2002; **2**:18-22.
- 31 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**:37-46.
- 32 Huang Y, Rao A. Connections between TET proteins and aberrant DNA modification in cancer. *Trends Genet* 2014; **30**:464-474.
- 33 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**:357-359.
- 34 Li H, Handsaker B, Wysoker A, *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; **25**:2078-2079.
- 35 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**:841-842.
- 36 Robinson JT, Thorvaldsdottir H, Winckler W, *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011; **29**:24-26.
- 37 Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013; **14**:178-192.
- 38 Shen L, Shao N, Liu X, Nestler E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* 2014; **15**:284.
- 39 Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**:1-13.
- 40 Wu C, Orozco C, Boyer J, *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* 2009; **10**:R130.
- 41 Wu C, Macleod I, Su AI. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res* 2013;

- 41:D561-D565.
- 42 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**:2114-2120.
- 43 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013; **14**:R36.
- 44 Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 2013; **31**:46-53.
- 45 Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**:57-74.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)



This work is licensed under a Creative Commons Attribution 4.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2017