Cell Research www.nature.com/cr

LETTER TO THE EDITOR

Circulating tumor DNA 5-hydroxymethylcytosine as a novel diagnostic biomarker for esophageal cancer

Cell Research (2018) 28:597–600; https://doi.org/10.1038/s41422-018-0014-x

Dear Editor,

Esophageal cancer is a serious malignancy with high rates of incidence and mortality. It ranked the eighth most common cancer and the sixth leading cause of cancer death worldwide.[1] About 87% of esophageal cancers are esophageal squamous cell carcinomas (ESCCs), with the highest incidence found in South-Eastern and Central Asia.[2] Notably, almost half of the global esophageal cancers occur in China.[3] The 5-year survival rate of esophageal cancer patients is only 15 to 20%, partly due to late clinical presentation and lack of early diagnostic biomarkers.[4] Therefore, exploring a highly sensitive and specific early diagnosis method is in urgent need.

One of the early events that occur during carcinogenesis is epigenetic alterations, including aberrant DNA and histone modifications.[5] Circulating cell-free DNA (cfDNA) in plasma has been shown to reflect the epigenetic features in cancer patients. 5-hydroxymethylcytosine (5hmC), the oxidative product of 5-methylcytosine (5mC) catalyzed by ten-eleven translocation protein family, is not only a relatively stable intermediate of active DNA demethylation, but also regarded as a novel epigenetic hallmark of cancer.[6,7] Two recent studies have revealed that 5hmC patterns in cfDNA provide tumor-associated signatures of several human cancers.[8,9] Therefore, 5hmCs in cfDNA have potential to be promising biomarkers for minimally invasive diagnosis in esophageal cancer.

Here, we utilized our recently established nano-hmC-Seal method to map the 5hmC profiles in cfDNA from a cohort of 150 newly diagnosed esophageal cancer patients and 177 healthy individuals (Fig. 1a; Supplementary information, Figure S1 and Table S1).[9,10] In addition to ESCCs, there were nine adenocarcinoma, three small cell carcinoma, and one neuroendocrine carcinoma samples in our cohort, identified by hematoxylin and eosin staining (Fig. 1b).

We first identified the 5hmC-enriched peaks and found that the peak numbers were more stable in control samples than esophageal samples (Supplementary information, Figure S2A). The heterogeneity of tumor samples may contribute to this result. Then we evaluated the distribution of 5hmC along the gene bodies in tumor and control groups. Compared to controls, tumor groups showed some increased 5hmC levels in gene bodies (Supplementary information, Figure S2B). Then, we identified differentially hydroxymethylated regions (DhMRs) and detected 5hmC-gain regions (9,047) and 5hmC-loss regions (10,460) in tumor groups by comparing tumor groups with control groups. 5hmC-gain regions, but not the 5hmC-loss regions, were particularly enriched in promoter and UTR regions. (Supplementary information, Figure S2C, D). Meanwhile, we found the enrichment of 5hmC-gain regions in short interspersed nuclear element, long tandem repeat, and satellite repeats. (Supplementary information, Figure S2D). All these results suggest that cfDNA 5hmC profiles of healthy individuals and esophageal cancer patients indeed display significant differences. To better understand the correlation between regulatory sequence codes and 5hmC changes, we performed de novo motif analysis in DhMRs. Most of the 5hmC peaks in 5hmC-gain regions were enriched in CEBP motif (CCAAT/enhancer-binding protein epsilon, $p = 1e^{-168}$), which was highly related to transcriptional misregulation in cancer. In contrast, ARNT motif was observed in the 5hmC-loss regions. It was known that the heterodimer composed of ARNT and HIF1A acted as a transcriptional regulator of adaptive response to hypoxia. Thus, the enrichment of this heterodimer may be the result of hypoxia (Supplementary information, Figure S2E).

To further explore the 5hmC signal changes between tumor and non-tumor samples, we then detected the differentially regulated 5hmC genes (i.e., genes with differential 5hmC levels) in esophageal cancer samples with DESeq2 package.[11] The results showed that esophageal cancer could lead to both upregulated and downregulated 5hmC levels in genes compared to control individuals (up 1,344 and down 231). To further validate the classification effects of 5hmC signal for esophageal cancer and control samples, we clustered the genes with differentially regulated 5hmC levels by hierarchical clustering method and the results showed that the majority of cancer samples were distinct from non-cancer samples (Fig. 1c). However, both cancer and non-cancer samples in cluster 2 showed similar patterns. Then we carried out the PCA (principal component analysis) analysis for genes with differentially regulated 5hmC levels and found that esophageal cancer samples showed distinct signatures and could be readily separated from control samples (Fig. 1d). Meanwhile, the four categories obtained by the clustering analyses could also be separated from each other in PCA result (Supplementary information, Figure S3A). In contrast, the samples from different age range were not separated (Supplementary information, Figure S3B). The hierarchical clustering and PCA analysis based on top variance genes showed similar results (Supplementary information, Figure S3C, D).

Next, to explore the function of differentially regulated 5hmC genes in esophageal cancer, we did the functional enrichment analysis for genes with upregulated and downregulated 5hmC levels in esophageal cancer, respectively (Fig. 1e, f). Our analyses showed cancer-related and metastasis-related pathways such as Hippo signaling pathway, platelet homeostasis, PI3k–Akt signaling pathway, and MAPK signaling pathway were enriched. Together, these results indicate that the 5hmC signal within genes display obvious difference between the esophageal cancer and control and these differentially regulated 5hmC genes are enriched in pathways associated with cancer and metastasis. In addition, we further compared our data to

Received: 26 November 2017 Revised: 18 January 2018 Accepted: 30 January 2018
Published online: 21 February 2018

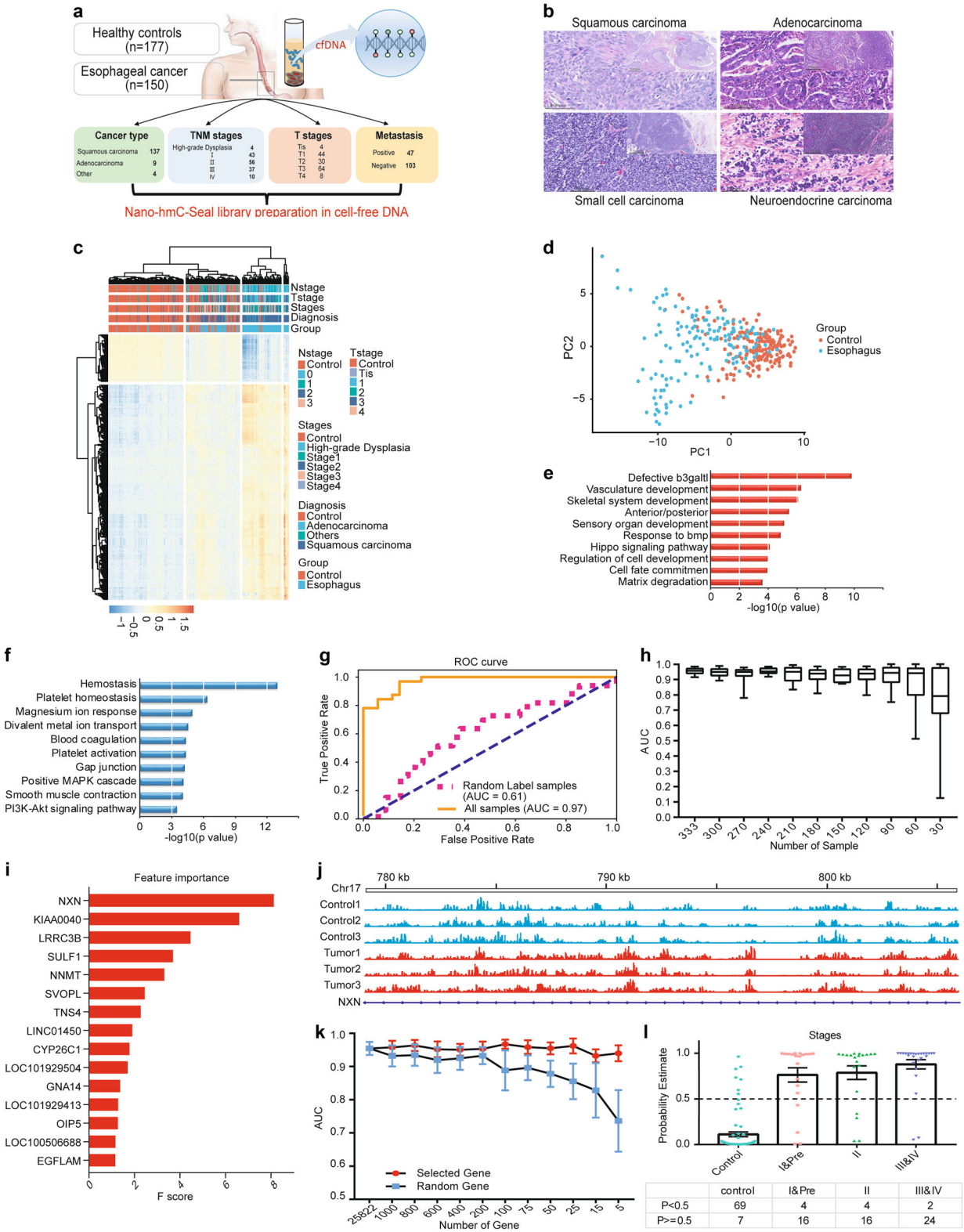

Fig. 1 cfDNA 5-hydroxymethylcytosine as a novel diagnostic biomarker for esophageal cancer. **a** Schematic overview of sample collections from esophageal cancer patients and healthy controls. **b** Different histological types of esophageal cancer identified by hematoxylin and eosin (H&E) staining. **c** Heat map showing clustering of 183 control and 150 esophagus cancer samples based on differential 5hmC ($|\log_2 FC| \geq 0.5$ and $p \leq 1e^{-20}$). Stage 0 indicates precancerous lesion. "Others" in "diagnosis legend" indicates endocrine carcinoma and small cell carcinoma. **d** PCA plot of normalized 5hmC reads from control and esophagus cancer samples. **e**, **f** GO enrichment analysis of genes with significant 5hmC increase (**e**) or decrease (**f**) in esophagus cancer samples. **g** ROC curve indicating the performance and credibility of our model. **h** The classifier performance gradually reduced along with samples number decreasing. **i** Bar plot showing the F score of top 15 important genes based on their contribution (feature importance) in model training. **j** The normalized 5hmC values of NXN (top one important machine learning gene) between control and esophagus cancer samples. **k** Classifier performance is stable upon reducing the number of contributive genes. **l** The predicted cancer probability based on 5hmC classifier from plasma cfDNA shows a significant trend associated with clinical stage

published data set about colorectal and gastric cancers.[9] Most of the genes with differential 5hmC levels were unique to esophageal cancer (81.7%, 1,287 genes; Supplementary information, Figure S4).

Next, we aim to utilize 5hmC characteristics detected in cfDNA for cancer classification. Using XGBoost method,[12] the 5hmC values in gene bodies of 333 samples, including 177 healthy controls plus 6 replicative samples and 150 esophagus cancer patients, were used to construct a classifier model. We separated samples into three groups (199 training samples, 67 validation samples, and 67 testing samples) for model training and evaluation. The prediction performance achieved a sensitivity of 93.75% and specificity of 85.71% (AUC = 0.972) (AUC, area under curve) in independent testing set (dark-orange line, Fig. 1g). In contrast, models trained with random labeled samples showed a poor performance (AUC = 0.61, deep pink line, Fig. 1g), indicating that the model does not overfit. To evaluate the performance, the model training was repeated 100 times and received an average AUC of 0.947 on testing set and 0.956 on validating set.

In order to determine whether the sample size is sufficient for model training, we reduced the sample number to train and test the classifier performance and found that the performance gradually reduced under 240 samples (Fig. 1h). Next, we used random sub-sampling (0.40) to train and valid 100 times in 199 training sets. Top genes were selected according to the contribution of each gene in the model (Fig. 1i). To verify that the selected genes are effective for the diagnosis of cancer, we reduced the number of genes to carry out training and testing (Fig. 1k). Classifier performance is stable with five genes that show highest contribution. As expected, the model's performance decreased dramatically when we use the random genes. In addition, we confirmed that the 5hmC signals at these genes are different (Fig. 1j; Supplementary information, Figure S3E).

We next studied the utility of our model in staging of esophageal cancer (191 samples for the training set and 142 samples for the validation set). Box plots were used to show the predicted probability of esophageal cancer patients and healthy samples (Fig. 1l). With the progression of cancer stage, the probability of being predicted as cancer gradually increased ($F_{(3,138)} = 80.572$, $P < 0.001$, ANOVA). Meanwhile, esophageal cancer with lymph node metastatic carcinoma (LNMC) samples were given higher probability than non-LNMC samples ($F_{(2,139)} = 122.233$, $P < 0.001$, ANOVA) (Supplementary information, Figure S3F). It is consistent with our understanding of cancer process. However, a small number of control samples were predicted to be of high probability of being cancer samples. These samples might be obtained from undiagnosed cancer patients or those at a high risk, and this is subject to our follow-up tracking and research. In addition, to further confirm the effect of age in our model, we picked a subset of 73 cancer and 74 healthy samples with matched age; and ROC curve showed clear separation between cancer and healthy samples (Supplementary information, Figure S5).

In summary, we utilized nano-hmC-Seal to generate the 5hmC profiles in esophageal cancer patients and identified robust esophageal cancer-associated 5hmC signatures in cfDNA. Meanwhile, we discovered 5hmC-based biomarkers in circulating cfDNA of esophageal cancer. Recent studies have reported that 5-hydroxymethylcytosine signatures in circulating cell-free DNA can be used as diagnostic biomarkers for some human cancers, however, the 5hmC patterns in cfDNA of esophageal cancer are unknown. Our study revealed the genome-wide pattern of cancer-associated 5hmC changes in plasma cfDNA in esophageal cancer. The identified 5hmC biomarkers could be used as detection tools for esophageal cancer. Additional patient studies in future will further improve the performance of this approach. The strategy presented here provides a foundation for effective future biopsy-based diagnosis using body fluids.

ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China 2016YFC0900300, NSFC 31670895/U1504831/31741074, Major Science and Technology Project of Henan Province (161100310100), CAS Strategic Priority Research Program XDA16010108/XDB14030300, CAS Hundred Talent Program, Youth Innovation Promotion Association, CAS 2016097. Shanghai Epican Genetech sponsor part of the sequencing cost.

AUTHOR CONTRIBUTIONS

Q.K., Y.-G.Y., and C.H. conceived the idea. X.T., J.Z., X.L., Z.Q., X.Z., Y.X., R.L., X.L. Y.-G.Y., and Q.K. design the experiments, recruited patients, collected blood and organized clinical information. Shanghai Epican Genetech finished all the hmC-Seal experiments with their protocol. C.C., C.G., B.S., L.W., and X.H. perform bioinformatics analysis under the supervision of D.H. B.S., C.C., C.G., J.Z., D.H. and Y.-G.Y. wrote the manuscript with input from all authors.

ADDITIONAL INFORMATION

Supplementary information accompanies this paper at https://doi.org/10.1038/s41422-018-0014-x.

Competing interests: The authors declare that they have no conflict of interest.

Xin Tian[1,2], Baofa Sun[3,4], Chuanyuan Chen[3,4], Chunchun Gao[3,4], Ji Zhang[1,2], Xingyu Lu[5,6], Linchen Wang[3,4], Xiangnan Li[7], Yurong Xing[7], Ruijuan Liu[1,2], Xiao Han[3,4], Zheng Qi[7], Xiaojian Zhang[1,2], Chuan He[8], Dali Han[3,4], Yun-Gui Yang[3,4] and Quancheng Kan[1,2]

[1]Department of Pharmacy, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450052, China; [2]Henan Key Laboratory of Precision Clinical Pharmacy, Zhengzhou University, Zhengzhou, Henan 450052, China; [3]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China; [4]College of Future Technology, Sino-Danish College, University of Chinese Academy of Sciences, Beijing 100049, China; [5]Shanghai Epican Genetech, Co. Ltd., Zhangjiang Hi-Tech Park, Shanghai 201203, China; [6]Shanghai Epican Biotech, Co. Ltd., Qingpu, Shanghai 201799, China; [7]The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052 Henan, China and [8]Department of Chemistry, Department of Biochemistry and Molecular Biology, and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, Chicago, IL 60637, USA

These authors contributed equally: Xin Tian, Baofa Sun, Chuanyuan Chen, Chunchun Gao, Ji Zhang and Xingyu Lu

Correspondence: Dali Han (handl@big.ac.cn) or Yun-Gui Yang (ygyang@big.ac.cn) or Quancheng Kan (kanqc@zzu.edu.cn)

REFERENCES

1. Ferlay, J. et al Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359 (2015).
2. Arnold, M. et al. Predicting the future burden of esophageal cancer by histological subtype: international trends in incidence up to 2030. *Am. J. Gastroenterol.* **112**, 1247–1255 (2017).
3. Zeng, H. et al Esophageal cancer statistics in China, 2011: estimates based on 177 cancer registries. *Thorac. Cancer* **7**, 232 (2016).
4. Napier, K. J., Scheerer, M. & Misra, S. Esophageal cancer: a review of epidemiology, pathogenesis, staging workup and treatment modalities. *World J. Gastrointest. Oncol.* **6**, 112 (2014).
5. Ma, K., Cao, B. & Guo, M. The detective, prognostic, and predictive value of DNA methylation in human esophageal squamous cell carcinoma. *Clin. Epigenetics* **8**, 1–9 (2016).
6. Chen, K. et al Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. *Cell Res.* **26**, 103 (2016).

7. Vasanthakumar, A. & Godley, L. A. 5-hydroxymethylcytosine in cancer: significance in diagnosis and therapy. *Cancer Genet.* **208**, 167–77 (2015).
8. Song, C. et al. 5-hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* **27**, 1231–42 (2017).
9. Li, W., Xu, Z. & Lu, X. 5-hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res.* **27**, 1243–57 (2017).
10. Han, D. L. et al A highly sensitive and robust method for genome-wide 5hmC profiling of rare cell populations. *Mol. Cell.* **63**, 711 (2016).
11. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
12. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In: *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA.* 785–794 (ACM, New York, NY, 2016).