Received: 14 January 2024 | Revised: 26 July 2024 | Accepted: 18 August 2024

DOI: 10.1002/cac2.12606

CANCER COMMUNICATIONS Open Access

LETTER TO THE JOURNAL

# Epigenomic exploration of disease status of *EGFR*-mutated non-small cell lung cancer using plasma cell-free DNA hydroxymethylomes

Non-small cell lung cancer (NSCLC) represents about 85% of histological diagnoses of lung cancer [1]. Epidermal growth factor receptor (*EGFR*) mutations occur in 12.7%-40.3% of NSCLC [2], and 5-hydroxymethylcytosine (5hmC) signatures and pathways can be inhibited by EGFR signaling [3]. The epigenome of plasma cell-free DNA (cfDNA), including 5hmC, has demonstrated promise as a cancer biomarker [4]. Currently, it remains unknown whether cfDNA 5hmC can identify disease status of NSCLC. Here, we performed 5hmC Seal-sequencing of 302 plasma cfDNA samples from 113 patients with metastatic *EGFR*-mutated NSCLC, which included 240 samples reflecting stable disease (SD) and 62 samples reflecting progressive disease (PD) (Figure 1A, Supplementary Table S1). SD and PD were clinically defined by the treating physician (Supplementary Methods).

High quality was ensured, 11 samples as outliers were discarded, and batch effects were removed effectively (Supplementary Figures S1, Supplementary Tables S2-S3). The remaining 291 samples were classified by disease status and various potential confounding factors (Figure 1A, Supplementary Tables S4-S7). The relative frequency of disease status in each group was nearly identical to that of the overall 291 samples (Supplementary Figure S4). cfDNA 5hmC peaks of each sample displayed proper reproducibility (Supplementary Figure S5A). Interestingly, 123 cfDNA 5hmC peaks were located on the *EGFR* gene (Supplementary Figure S5B, Supplementary Table S8). Genomewide cfDNA 5hmC levels were overall similar between PD and SD samples, as well as various potential confounders (Supplementary Figures S5C-E and S6).

A substantial portion of 5hmC peaks displayed high heterogeneity of 5hmC levels among the 291 samples (Supplementary Figure S7A), which were not derived from disease status and potential confounders (Supplementary Figure S7B-E, Supplementary Table S9). With 1,000 bp bins instead of peaks, similar results were observed (Supplementary Figure S8). We found that *EGFR* mutations were associated with 5hmC heterogeneity (Supplementary Figure S9A) and identified 4,743 cfDNA 5hmC peaks (Supplementary Table S10) with 5hmC levels differing among intergroups of *EGFR* mutation subtypes more than that of intragroups ($P < 0.005$) (Supplementary Figure S9B). Interestingly, the 4,743 cfDNA 5hmC peaks were strongly associated with the function of *EGFR* (Supplementary Figure S10A), but not associated with disease status (Supplementary Figure 10B-E). This result was further confirmed by a nearly identical 5hmC level between PD and SD samples (Supplementary Figure 10F), as well as distribution of false discovery rate and *P* values (Figure 1B).

Disease status-associated 5hmC peaks were completely different from potential confounder-associated 5hmC peaks, except for smoking status (Supplementary Figure S11). Consistently, 5hmC levels of SD and PD samples were significantly different on smoking status-associated peaks, but not on sex-, age-, or race-associated peaks (Supplementary Figure S12A). Comparisons between either two of the three smoking statuses or between the two disease statuses shared 123, 282, 106, and 58 differential 5hmC peaks, respectively (Supplementary Figure S12B-C, Supplementary Tables S11-S14). The shared 4 groups of 5hmC peaks showed differences of 5hmC levels between PD and SD samples (Figure 1C), and can classify both disease statuses and smoking statuses (Figure 1D,

**List of Abbreviations:** 5hmC, 5-hydroxymethylcytosine; ANOVA, analysis of variance; AUC, area under the receiver operating characteristic curve; cfDNA, cell free DNA; CV, coefficient of variation; CV, cross validation; DPs, differential peaks; *EGFR*, epidermal growth factor receptor; Ex20 LI, exon 20 loop insertion; *FBXL7*, f-box and leucine rich repeat protein 7; *LEPR*, leptin receptor; lgG3, immunoglobulin G3; lgM, immunoglobulin M; MDS, multidimensional scaling; *NORPEG (RAI14)*, novel retinal pigment epithelial cell protein; NSCLC, non-small cell lung cancer; PACC, P-loop and αC-helix compressing; PCA, principal component analysis; PEER, probabilistic estimation of expression residuals; *RAI14 (NORPEG)*, retinoic acid induced 14; TFs, transcription factors; *THRB*, thyroid hormone receptor beta.


**FIGURE 1** Identification, characterization, biological functionalization, and diagnostic application of disease status-dependent and patients' characteristics-independent cfDNA 5hmC in *EGFR*-mutated non-small cell lung cancer. (A) A flowchart for study design. Multiple samples might be collected from one patient at different time points. At different time points, some patients might have different disease

Supplementary Figure S13D-E). Overall, although 5hmC levels varied based on patients' characteristics, only smoking status affected disease status-associated 5hmC.

The hyper- or hypo-hydroxymethylated 5hmC peaks from PD versus SD samples (Supplementary Tables S15-S16) could not identify subtypes of sex, race, age, smoking status, or *EGFR* mutation (Figure 1E, Supplementary Figure S13A-C). They were correlated only with disease status, but not the potential confounders (Figure 1F, Supplementary Figure S13D-E). Functional enrichment analysis showed that the hyper-5hmC peaks were closely associated with lung development, vital capacity, and smoking (Supplementary Figure S14A-C). Interestingly, the hypo-5hmC peaks were not associated with lung function directly but may affect the disease status through the immune system, such as T cell activation, leukocyte adhesion, and lgM levels (Supplementary Figure S14D-E). Like the hyper-5hmC peaks, the hypo-5hmC peaks were also associated with smoking behaviors and forced expiratory volume (Supplementary Figure S14F).

The lung function- and immune system-associated 5hmC peaks were mainly located on gene bodies, but

statuses (SD or PD). More details for sample information and groups are listed in Supplementary Tables S1 and S3-S7. (B) Scatter plot displaying average 5hmC level in the two groups with different disease status (left), and distribution of *P* value and FDR of Wilcoxon-rank-sum test for comparing 5hmC level of SD versus PD samples on the 4,743 peaks (right). Each dot represents one of the 4,743 peaks. (C) Boxplots display distribution of 5hmC level on the smoking status-associated 5hmC peaks for comparing different disease status. The four boxplots represent the four highlighted numbers by red in Supplementary Figure S12C, from left to right: the 123 overlapped hypo-5hmC peaks from Never vs. Heavy, the 282 overlapped hypo-5hmC peaks from Never vs. Light, the 106 overlapped hyper-5hmC peaks from Light vs. Heavy, and the 58 overlapped hyper-5hmC peaks from Never vs. Light. (D) Heatmap showing hierarchical clustering for the 291 samples based on 5hmC level of the pooled peaks of the 4 sets in panel (C). Each row and column represent one sample and one peak, respectively. The sum of the number of the 4 sets in the figure (C) is 569 (123 + 282 + 106 + 58), but after merging the overlapping peaks, 523 peaks were generated. (E) Hierarchical clustering of the 291 samples using 5hmC levels on the 9,038 hyper-hydroxymethylated (left) and 2,244 hypo-hydroxymethylated (right) cfDNA 5hmC peaks which were generated by comparing SD and PD. Each row represents one 5hmC peak, each column represents one sample, and cfDNA 5hmC level was row scaled. Samples were colored based on different classification criteria which are illustrated in Supplementary Tables S6-S7. (F) Scatter plots illustrate 5hmC levels of the 9,038 hyper- (cyan dots) and 2,244 hypo-hydroxymethylated (pink dots) 5hmC peaks were correlated with disease status, but not potential confounders (sex, age, race, *EGFR* mutation subtypes, and smoking status). Each dot represents one cfDNA 5hmC peak. (G) Genomic distribution of the 9,038 hyper- and 2,244 hypo-5hmC peaks (middle panel), and gene ontology (biological processes) enrichment results for the hyper- (upper panel) and hypo-5hmC (lower panel) peaks which were located at gene body. Lung cancer-associated and immune-associated terms were highlighted by red (upper panel) and purple (lower panel), respectively. (H) Box plots showing distribution of cfDNA 5hmC levels on 8 hyper-5hmC peaks located at gene body of 8 lung cancer associated genes. Each panel represents one hyper-5hmC peak, its associated gene was shown. The *P* value between two groups (SD versus PD) was determined by Wilcoxon rank-sum test. (I) Human lung enhancer enrichment analysis for the hyper- and hypo-5hmC peaks by using LOLA. Based on genomic location, differential 5hmC peaks were classified into 3'-UTR, 5'-UTR, intergenic region, gene body, and promoter. The proportion (upper) and *P* value of Chi-square test (lower) of human lung enhancers overlapping with each class are shown by bar graph. (J) Enrichment of lung function associated transcription factors (TFs) in hyper-5hmC peaks located at gene body and intergenic. Motif of TFs binding (left), name of TFs (middle), and *P* value (right) was shown respectively. (K) Sensitivity and specificity metrics plotted against cutoff values (upper panels), and distribution of outputs of logistic regression (lower panels) were plotted to classify disease status by using hyper-, hypo-, or all the differential 5hmC peaks, respectively. The black dashed line represents score cutoff (output of logistic regression). The three logistic models were optimized based on 5hmC levels of the differential peaks by 10-fold cross-validation (10-fold CV). (L) Hierarchical clustering of the 291 samples using 5hmC levels on the 888 5hmC peaks which were the optimized results of the logistic model by 10-fold CV. Each column represents one 5hmC peak, each row represents one sample, and cfDNA 5hmC level was row scaled. Samples were colored based on different classification criteria which were described in Supplementary Tables S6-S7. (M) Receiver operating characteristic (ROC) curves generated by using logistic model to classify disease status (SD versus PD), age (younger versus older), sex (female versus male), race (white versus black), and smoking status (heavy, light, and never), respectively (left). Sensitivity, specificity, precision, and accuracy of the 7 logistic models were also shown (right). The logistic models were optimized based on the 5hmC level of the 888 peaks by 10-fold CV. AUC for each classification was also shown in parentheses. (N) The 888 differential 5hmC peaks (DPs) are classified into two groups based on their genomic locations: gene body or intergenic; 626 and 262 of the 888 DPs are located at gene body and intergenic, respectively. The 626 gene body DPs are associated with 590 different genes. The 590 genes are further separated into 3 groups based on their number of DPs: 1) 557 genes containing only one 5hmC peak 2) 30 genes each containing two 5hmC peaks, and 3) each of the three genes containing three 5hmC peaks. (O) Sensitivity, specificity, precision, accuracy (left), and FDR (right) of the two logistic models: using all the 888 DPs or the top 63 DPs. The top 63 5hmC peaks were selected from the 888 peaks based on their logistic regression coefficients. Abbreviations: 5hmC, 5-hydroxymethylcytosine; NSCLC, non-small cell lung cancer; *EGFR*, epidermal growth factor receptor; cfDNA, cell-free DNA; SD, stable disease; PD, progressive disease; ANOVA, analysis of variance; DPs, differential peaks; vs., versus; UTR, untranslated region; LOLA, locus overlap analysis; TFs, transcription factors; CV, cross-validation; ROC, receiver operating characteristic; AUC, area under the receiver operating characteristic curve; FDR, false discovery rate.

not intergenic (Figure 1G), such as a hyper-5hmC peak at the intron of thyroid hormone receptor beta (*THRB)* gene (Supplementary Figure S15A) which regulates lung development [5]. Some important lung function-associated genes were hyper-hydroxymethylated (Figure 1H, Supplementary Figure S15B), whereas hypo-5hmC peaks were located on the gene body of immune-associated genes (Supplementary Figure S15C, Supplementary Table S17). Regulatory elements and lung enhancers were enriched in the gene body, promoter, or intergenic regions of the hyper- or hypo-5hmC peaks (Figure 1I, Supplementary Figure S15D). Motifs and binding regions of some lung function-associated transcription factors (TFs) were also enriched in the hyper-5hmC peaks (Figure 1J, Supplementary Figure S15E). Taken together, disease status-dependent and patient characteristics-independent cfDNA 5hmC peaks can be linked to lung development, smoking behavior, and immune response, as well as lung function-associated enhancers and TF-binding sites (Supplementary Figure S16).

We optimized 888 peaks (Supplementary Table S18) from the differential 5hmC peaks to build a logistic regression model with an area under the receiver operating characteristic curve (AUC) of 0.998 using appropriate cutoffs of the output probabilities (Figure 1K, Supplementary Figure S17A-B). Based on the 888 peaks, unsupervised clustering could discriminate PD and SD samples with 100% accuracy, while not being able to discriminate different groups from sex, race, age, smoking status, or *EGFR* mutation subtypes (Figure 1L). The AUC of the model for predicting disease status was much greater than those for classifying age, sex, race, or smoking status (Figure 1M, Supplementary Figure S17C). Our cfDNA 5hmC-based logistic regression model could discriminate disease status accurately, sensitively, and specifically, and was independent of potential confounding factors in NSCLC (Figure 1M, Supplementary Figure S17D-E). The 888 peaks could not distinguish the 10 treatment-naïve samples and the 49 previously treated samples (Supplementary Figure S17F-G).

As expected, most of the 888 peaks were located at gene bodies, and multiple peaks might locate on the same gene (Figure 1N). Interestingly, three genes with three of the 888 peaks (Figure 1N) were strongly associated with lung function and lung cancer [6, 7], and highly expressed in various cancers including lung cancer (Supplementary Figure S18A). More importantly, lung cancer patients with different expression levels of the retinoic acid-induced 14 (*RAI14* or *NORPEG*) gene demonstrated a 16% difference of survival probability (Supplementary Figure S18B). Furthermore, some of the 30 genes with t peaks (Figure 1N) were associated with lung function such as leptin receptor (*LEPR)* and f-box and leucine rich repeat protein 7 (*FBXL7)* [8, 9], survival probability of patients with lung cancer (Supplementary Figure S18C-D), and exhibited high expression level in lung cancer (Supplementary Figure S18E). To determine 5hmC biomarkers for disease status, the top 63 cfDNA 5hmC peaks (Supplementary Table S19) with maximum absolute values of logistic regression coefficients were selected from the 888 peaks. The 63-5hmC peak-based logistic model could also achieve high performance (Figure 1O, Supplementary Figure S19A-B). Some of the 63-peak-associated genes not only played an important role for lung function but also correlated with lung cancer survival probability (Supplementary Figure S19C-E).

Overall, we found that smoking status affected disease status-associated cfDNA 5hmC. We unveiled that lung function and regulatory elements were enriched in disease status-associated 5hmC peaks which could discriminate progressive and stable NSCLC with high sensitivity and specificity. Our results conferred the epigenomic distinguishability of different treatment responses and nominated cfDNA 5hmC profiling as a non-invasive, cost-effective, and universally applicable approach to monitor disease status.

## AUTHOR CONTRIBUTIONS

C.M.B. and C.H. contributed to project conceptualization and study design. J.K. collected samples and prepared libraries for sequencing. Y.P. performed the bioinformatic analysis. J.D.P., E.E.V., M.C.G., K.L., Z.Z, W.Z, and M.C. helped samples collection, and provided project supervision and suggestions. Y.P., C.M.B., J.K., and C.H. wrote the manuscript with input from all authors.

## ACKNOWLEDGMENTS

We have no additional acknowledgements.

## CONFLICT OF INTEREST STATEMENT

C.M.B. reports research funding to the institution from AstraZeneca and BMS; advisory boards and personal consulting payments from Amgen, AstraZeneca, BMS, CVS, Daiichi Sankyo, EMD Serono, Gilead, Guardant, JNJ, Mirati, Novocure, Sanofi, Tempus and Turning Point Therapeutics. M.C.G. reports funding to the institution from Eli Lilly, MSD, Pfizer (MISP); AstraZeneca, MSD International GmbH, BMS, Boehringer Ingelheim Italia S.p.A, Celgene, Eli Lilly, Ignyta, Incyte, MedImmune, Novartis, Pfizer, Roche, Takeda, Tiziana, Foundation Medicine, Glaxo Smith Kline GSK, Spectrum pharmaceuticals. MCG reports advisory boards and personal consulting payments from AstraZeneca, MSD International GmbH, Bayer, BMS, Boehringer Ingelheim Italia S.p.A, Celgene, Eli Lilly, Incyte, Novartis, Pfizer, Roche, Takeda, Seattle Genetics, Mirati, Daiichi Sankyo, Regeneron, Merck, Blueprint, Jansenn, Sanofi, AbbVie, BeiGenius, Oncohost. The remaining authors report no competing interests.

## FUNDING INFORMATION

No funding was received for this project.

## DATA AVAILABILITY STATEMENT

Data have been deposited in the NCBI Gene Expression Omnibus (GEO) and are accessible through GEO series accession number GSE231296. Bioinformatics pipeline for 5hmC Seal-sequencing data analysis and scripts used for plotting figures are available at https://github.com/CTLife/cfDNA-5hmC_LungCancer

## ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the local Institutional Review Board according to the U.S. Common Rule ethical guidelines. All patients were consented to a general thoracic biobanking study under IRB 18-1319, which allowed for utilization of samples collected under IRB 9571.

Yong Peng[1,2,3,4,5]
Jason Karpus[1]
Jyoti D. Patel[6]
Everett E. Vokes[7]
Marina Chiara Garassino[7]
Kirsteen Lugtu[7]
Zhou Zhang[8]
Wei Zhang[8]
Mengjie Chen[4,5]
Chuan He[1,2,3]
Christine M. Bestvina[7]

[1]*Department of Chemistry, The University of Chicago, Chicago, Illinois, USA*
[2]*Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois, USA*
[3]*Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois, USA*
[4]*Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, Illinois, USA*
[5]*Department of Human Genetics, The University of Chicago, Chicago, Illinois, USA*
[6]*Division of Hematology and Oncology, Department of Medicine, Northwestern University, Chicago, Illinois, USA*
[7]*Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Chicago, Illinois, USA*
[8]*Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA*

**Correspondence**
Christine M. Bestvina, Section of Hematology/Oncology, Department of Medicine, The University of Chicago, Chicago, IL, USA.
Email: cbestvina@medicine.bsd.uchicago.edu

Chuan He, Department of Chemistry, The University of Chicago, Chicago, IL, USA.
Email: chuanhe@uchicago.edu

Yong Peng and Jason Karpus contributed equally.

## ORCID

*Yong Peng* https://orcid.org/0000-0003-3405-8836
*Zhou Zhang* https://orcid.org/0000-0002-9639-052X

## REFERENCES

1. Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS. Lung cancer. Lancet. 2021;398(10299):535-54.
2. Zhang YL, Yuan JQ, Wang KF, Fu XH, Han XR, Threapleton D, et al. The prevalence of EGFR mutation in patients with non-small cell lung cancer: a systematic review and meta-analysis. Oncotarget. 2016;7(48):78985-93.
3. Beadell AV, Zhang Z, Capuano AW, Bennett DA, He C, Zhang W, et al. Genome-Wide Mapping Implicates 5-Hydroxymethylcytosines in Diabetes Mellitus and Alzheimer's Disease. J Alzheimers Dis. 2023;93(3):1135-51.
4. Guler GD, Ning Y, Ku CJ, Phillips T, McCarthy E, Ellison CK, et al. Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. Nat Commun. 2020;11(1):5270.
5. Cao S, Feng H, Yi H, Pan M, Lin L, Zhang YS, et al. Single-cell RNA sequencing reveals the developmental program underlying proximal-distal patterning of the human lung at the embryonic stage. Cell Res. 2023:1-13.
6. Yuan C, Hu H, Kuang M, Chen Z, Tao X, Fang S, et al. Super enhancer associated RAI14 is a new potential biomarker in lung adenocarcinoma. Oncotarget. 2017;8(62):105251-61.
7. Jiang Z, Zhao J, Zou H, Cai K. CircRNA PTPRM Promotes Non-Small Cell Lung Cancer Progression by Modulating the miR-139-5p/SETD5 Axis. Technol Cancer Res Treat. 2022;21:15330338221090090.
8. Unsal M, Kara N, Karakus N, Tural S, Elbistan M. Effects of leptin and leptin receptor gene polymorphisms on lung cancer. Tumour Biol. 2014;35(10):10231-6.
9. Zhou J, Lin Y, Kang X, Liu Z, Zou J, Xu F. Hypoxia-mediated promotion of glucose metabolism in non-small cell lung cancer correlates with activation of the EZH2/FBXL7/PFKFB4 axis. Cell Death Dis. 2023;14(5):326.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.