

RESEARCH PAPER



EGLN2 DNA methylation and expression interact with HIF1A to affect survival of early-stage NSCLC

Ruyang Zhang^{a,b,c*}, Linjing Lai^{a*}, Jieyu He^{a*}, Chao Chen^a, Dongfang You^a, Weiwei Duan^a, Xuesi Dong^{a,d}, Ying Zhu^a, Lijuan Lin^a, Sipeng Shen^{b,c}, Yichen Guo^{b,e}, Li Su^{b,c}, Andrea Shafer^f, Sebastian Moran^{b,g}, Thomas Fleischer^h, Maria Moksnes Bjaanæs^h, Anna Karlssonⁱ, Maria Planckⁱ, Johan Staaf^j, Åslaug Helland^{b,hj}, Manel Esteller^g, Yongyue Wei^{a,b,c}, Feng Chen^{a,c,k}, and David C. Christiani^{b,c,cf}

^aDepartment of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China; ^bDepartment of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA; ^cChina International Cooperation Center for Environment and Human Health, Nanjing Medical University, Nanjing, Jiangsu, China; ^dDepartment of Epidemiology and Biostatistics, School of Public Health, Southeast University, Nanjing, Jiangsu, China; ^eDepartment of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; ^fPulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA; ^gBellvitge Biomedical Research Institute and University of Barcelona and Institutio Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia, Spain; ^hDepartment of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; ⁱDivision of Oncology and Pathology, Department of Clinical Sciences Lund and CREATE Health Strategic Center for Translational Cancer Research, Lund University, Lund, Skåne, Sweden; ^jInstitute of Clinical Medicine, University of Oslo, Oslo, Norway; ^kJiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Cancer Center, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, Jiangsu, China

ABSTRACT

Hypoxia occurs frequently in human cancers and promotes stabilization and activation of hypoxia inducible factor (HIF). HIF-1 α is specific for the hypoxia response, and its degradation mediated by three enzymes *EGLN1*, *EGLN2* and *EGLN3*. Although *EGLNs* expression has been found to be related to prognosis of many cancers, few studies examined DNA methylation in *EGLNs* and its relationship to prognosis of early-stage non-small cell lung cancer (NSCLC). We analyzed *EGLNs* DNA methylation data from tumor tissue samples of 1,230 early-stage NSCLC patients, as well as gene expression data from The Cancer Genome Atlas. The sliding windows sequential forward feature selection method and weighted random forest were used to screen out the candidate CpG probes in lung adenocarcinomas (LUAD) and lung squamous cell carcinomas patients, respectively, in both discovery and validation phases. Then Cox regression was performed to evaluate the association between DNA methylation and overall survival. Among the 34 CpG probes in *EGLNs*, DNA methylation at cg25923056_{*EGLN2*} was identified to be significantly associated with LUAD survival ($HR = 1.02$, 95% CI: 1.01–1.03, $P = 9.90 \times 10^{-5}$), and correlated with *EGLN2* expression ($r = -0.36$, $P = 1.52 \times 10^{-11}$). Meanwhile, *EGLN2* expression was negatively correlated with *HIF1A* expression in tumor tissues ($r = -0.30$, $P = 4.78 \times 10^{-8}$) and significantly ($P = 0.037$) interacted with *HIF1A* expression on overall survival. Therefore, DNA methylation of *EGLN2*-*HIF1A* is a potential marker for LUAD prognosis and these genes are potential treatment targets for further development of HIF-1 α inhibitors in lung cancer therapy.

ARTICLE HISTORY

Received 26 October 2018
Revised 10 January 2019
Accepted 17 January 2019







KEYWORDS

EGLN2; *HIF1A*; DNA methylation; lung cancer; prognosis

Introduction


Lung cancer is a leading cause of cancer death worldwide [1] and non-small cell lung cancer (NSCLC) accounts for about 85% [2]. Five-year survival in populations with lung cancer varies from 4–17% depending on stage and regional differences [3]. Diagnosed at early stage (TNM stage I, II), when

curative surgical resection is possible, provides a good opportunity for improving survival [4]. However, even for early-stage patients with similar clinical characteristics, significant heterogeneity has been observed, which indicated that there are molecular mechanisms not well understood yet [5]. Molecular characterization such as DNA

CONTACT Feng Chen  fengchen@njmu.edu.cn  Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, SPH Building Room 412, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China; David C. Christiani  dchris@hsph.harvard.edu  Department of Environmental Health, Harvard T.H. Chan School of Public Health, Building I Room 1401, 665 Huntington Avenue, Boston, MA 02115, USA; Yongyue Wei  ywei@njmu.edu.cn  China International Cooperation Center for Environment and Human Health, Nanjing Medical University, SPH Building Room 418, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China

*These authors are contributed equally to this work.

[†]Senior author.

 Supplemental data for this article can be accessed on [publisher's website](#).

methylation is increasingly used to predict tumor prognosis and offers great potential for improving understanding of lung cancer.

DNA methylation, an inheritable reversible epigenetic modification, affects the spatial conformation of DNA, regulates gene expression and interacts with various positive and negative feedback mechanisms [6,7]. Thus, aberrant DNA methylation CpG probes have been considered potential cancer biomarkers and therapeutic targets not only in NSCLC [8,9], but also in other cancers [10,11].

Due to rapid cancer cell division and aberrant angiogenesis, hypoxia occurs frequently in human cancers [12]. Hypoxia in solid tumor tissues promotes stabilization and activation of hypoxia inducible factor (HIF), which is essential for adapting the cell's oxygen homeostasis to hypoxia, physiologically as well as pathologically [13,14]. HIF-1 α is specific for the hypoxia response. In normoxia, HIF-1 α is rapidly degraded, and its low levels do not allow heterodimer formation and transcriptional activation. The hydroxylation of two proline residues (Pro-402 and Pro-564) of HIF1 α by three distinct enzymes allows the specific recognition and ubiquitination of HIF1 α by the tumor suppressor pVHL (von-Hippel-Lindau protein), leading to the proteasomal degradation [15,16]. When hypoxia occurs, this degradation is suppressed and HIF-1 α is stabilized rapidly. These three enzymes are encoded by Egl-9 family hypoxia inducible factor 1 (*EGLN1*, also called *PHD2*: hydroxylase domain-containing proteins 2), *EGLN2* (*PHD1*) and *EGLN3* (*PHD3*) and all of them hydroxylate HIF- α , thus play a vital role in many pathophysiological processes including tumor promotion. *EGLNs* expression has been found to be related to prognosis in many cancers, such as colorectal cancer [17], pancreatic cancer [18] and breast cancer [19]. Due to different cancer type and mechanism, in some studies, the *EGLN/HIF* axis appears to drive tumorigenesis [17,18], but in the others it could play a positive role in tumor suppression [19,20]. All of these studies proved that *EGLNs* are important for many tumor processes. Several studies have reported that *EGLN3* (*PHD3*) hypermethylation might reduce DNA expression in colorectal cancer [21], invasive breast carcinomas [22] and a diverse set of malignant cells [23]. However, few studies

have examined the role of DNA methylation in *EGLNs* and its relationship to prognosis of NSCLC.

Therefore, using multi-center cohorts with DNA methylation as well as gene expression data, we performed a comprehensive analysis of DNA methylation in *EGLN* gene family and *EGLN-HIF1A* interaction on survival of early-stage NSCLC, aiming to find epigenetic biomarkers for potential therapy targets. The two-stage designed study composes a discovery set combining four independent Caucasian cohorts from Harvard, Spain, Norway and Sweden, as well as an independent validation set from The Cancer Genome Atlas (TCGA).

Results

Demographics and clinical characteristics of the study populations are described in Supplementary Table S1. The 34 CpG probes located in *EGLN* gene family members were included in the following analysis (Supplementary Table S2).

Analysis workflow was given in Figure 1. Among lung adenocarcinomas (LUAD) patients, the sliding windows sequential forward feature selection (SWSFS) algorithm identified top 8 and 10 CpG probes in the discovery phase and the validation phase, respectively (Figure 2(a-b)). Two probes, cg25923056 and cg08080060, were simultaneously ranked in the top list of both two phases (Figure 2(c-d)). Meanwhile, cg07040244 and cg08078058 were also identified in lung squamous cell carcinomas (LUSC) patients (Supplementary Figure S1(a-d)). These four CpG probes were further evaluated by Cox regression. After correction for multiple testing, only the probe cg25923056_{*EGLN2*} was significantly associated with survival among LUAD patients in both two phases ($HR = 1.02$, 95% CI: 1.01–1.04, $P = 1.27 \times 10^{-3}$ in the discovery phase; $HR = 1.03$, 95% CI: 1.00–1.05, $P = 0.026$ in the validation phase) and showed stronger association in combined set ($HR = 1.02$, 95% CI: 1.01–1.03, $P = 9.90 \times 10^{-5}$). Therefore, the following analyses were performed in LUAD patients only.

To better illustrate the effect of DNA methylation on overall survival, patients were categorized into two methylation level groups (low and high) based on the median value. Kaplan-Meier survival curves for patients in combined set with

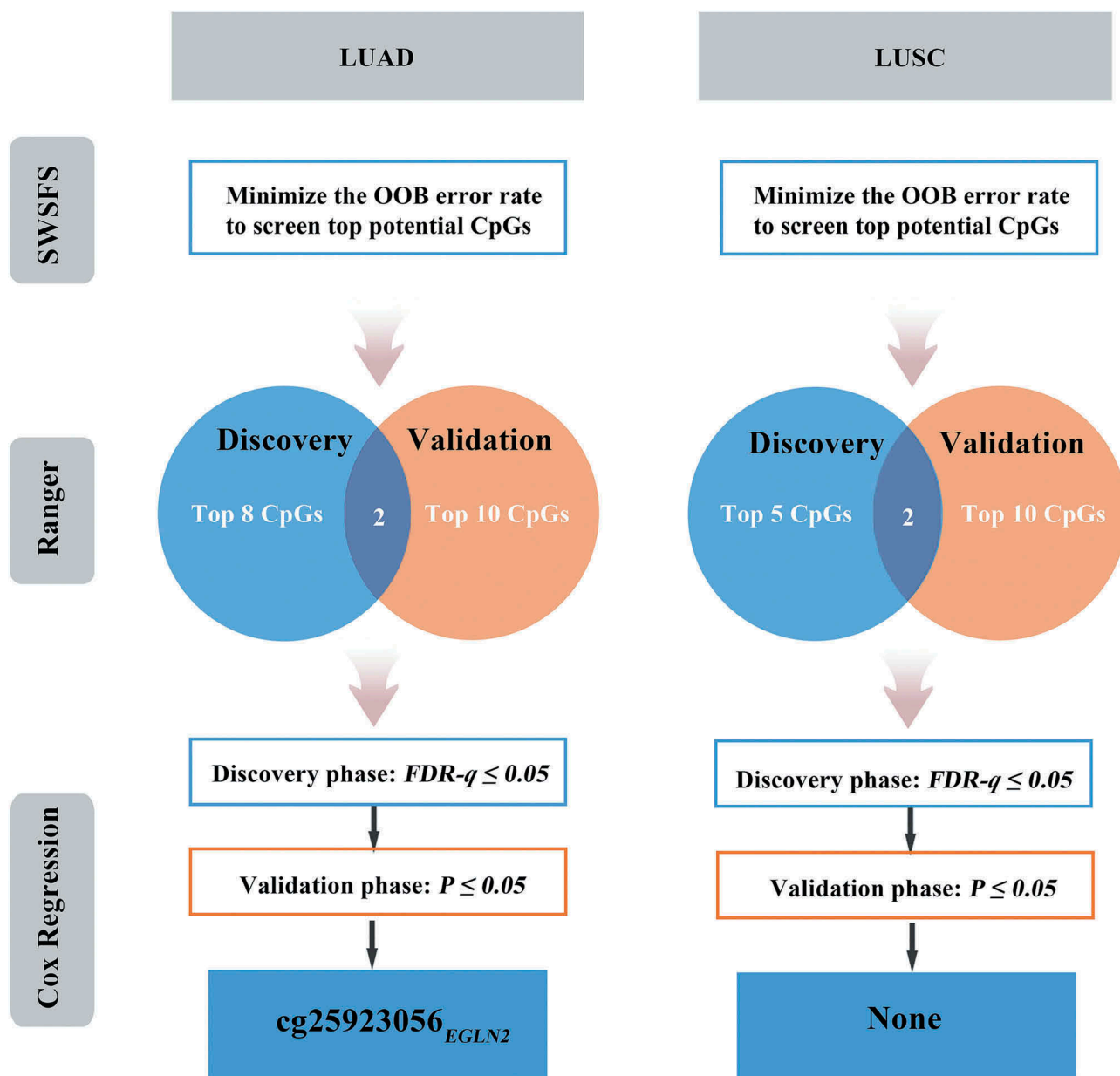


Figure 1. Analysis work flow. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) patients from Harvard, Spain, Norway, and Sweden cohorts were used in the discovery phase for screening. Data from The Cancer Genome Atlas (TCGA) were used for validation. The sliding windows sequential forward feature selection method (SWSFS) was used to identify the top important CpG probes by minimizing the 'out of bag (OOB)' error rate. Ranger is a weighted version of random forest. CpG probes ranked by variable importance score (VIS) in the tops in both discovery and validation phases were selected for further evaluation using Cox regression model. False discovery rate ($FDR \leq 0.05$) in the discovery phase and $P \leq 0.05$ in the validation phase were considered statistically significant.

high- and low-methylation of $cg25923056_{EGLN2}$ were shown in Figure 3 ($HR = 1.71$, 95% CI: 1.33–2.20, $P = 2.46 \times 10^{-5}$).

To consider both linear and non-linear effects of variables and handle complex interactions among them efficiently, RPART, a tree-based method, offers an attractive alternative to Cox models. Among

LUAD patients, $cg25923056_{EGLN2}$ and covariates were used to build a survival classification tree in the combined dataset (Figure 4(a)). Four clusters were identified with significantly different survival curves (Figure 4(b)) and mortality (Figure 4(c)). The probe $cg25923056_{EGLN2}$ was identified as the second most important predictors associated with LUAD

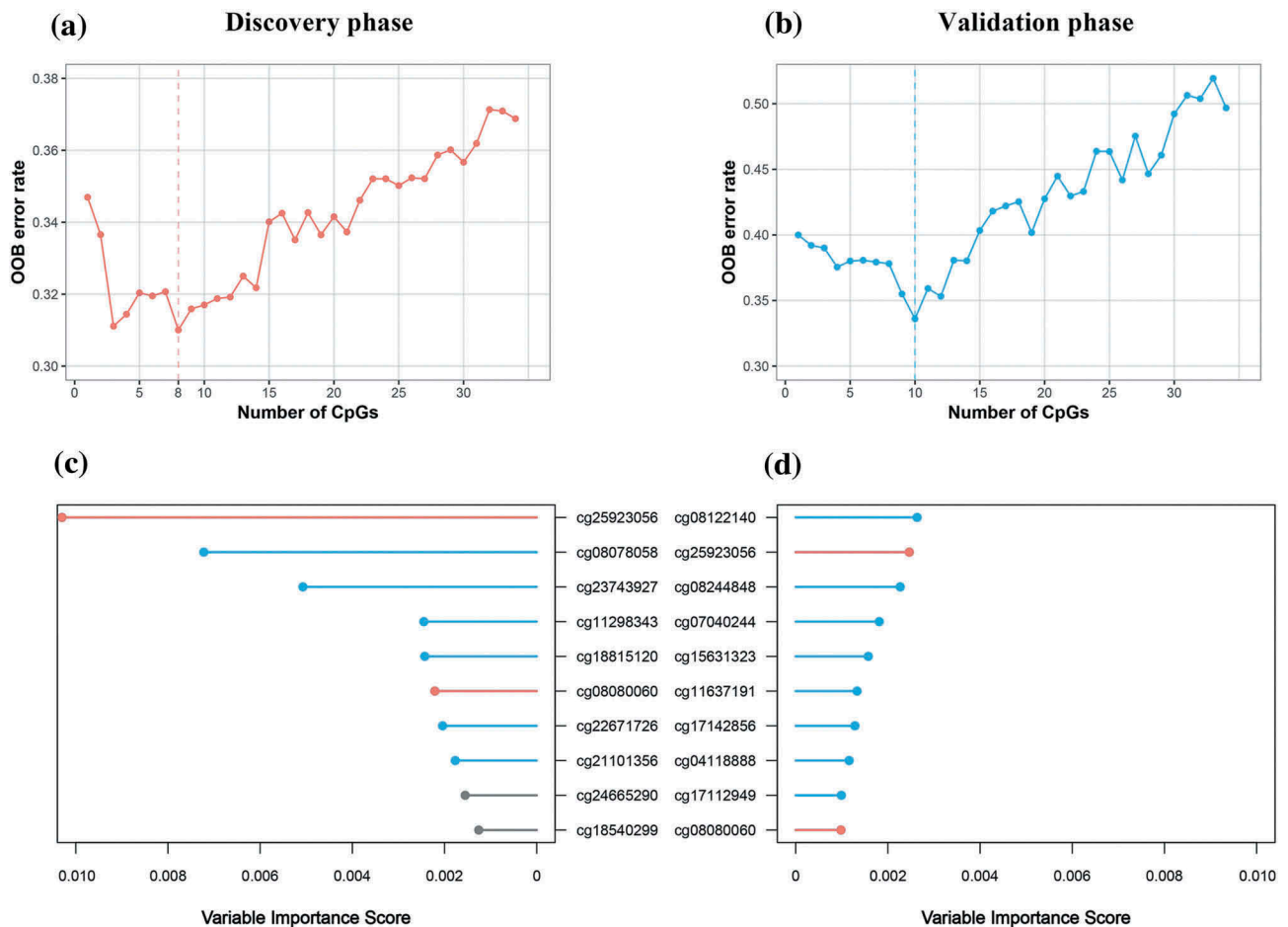


Figure 2. Ranger provides variable importance score (VIS) for each CpG probe for lung adenocarcinoma (LUAD) patients only in the discovery phase and the validation phase. ‘Out of bag (OOB)’ error rate of top CpG probes in the model, when probes were included one by one based on their VIS ranks in the discovery (a) and the validation phase (b). CpG probes (red lollipop) that were both in the top 8 in the discovery (c) and in top 10 in the validation phase (d) were carried forward for further evaluation using Cox regression model.

survival, followed by stage. The model showed that high methylation (≥ 0.48) at cg25923056_{EGLN2} in stage I patients was associated with worse survival compared with the low methylation group.

Interestingly, cg25923056_{EGLN2} was significantly hypermethylated in tumor tissues versus adjacent normal tissues (fold change (FC) = 1.23, $P = 7.05 \times 10^{-5}$) (Figure 5(a)). Meanwhile, *EGLN2* was significantly down-regulated in tumor tissues versus adjacent normal tissues (FC = 0.48, $P = 5.10 \times 10^{-14}$) (Figure 5(b)). Not surprisingly, DNA methylation of cg25923056_{EGLN2} was negatively associated with *EGLN2* gene expression ($r = -0.36$, $P = 1.52 \times 10^{-11}$) (Figure 5(c)). Further, *ENGL2* expression level was negatively correlated with its downstream gene *HIF1A* expression in tumor tissues ($r = -0.30$, $P = 4.78 \times 10^{-8}$) (Figure 5(e)) which was also differentially expressed

in tumor tissues (FC = 2.45, $P = 9.85 \times 10^{-11}$) (Figure 5(d)). Moreover, none of DNA methylation of the 9 CpG probes located in *HIF1A* was associated with prognosis of either LUAD or LUSC patients (Supplementary Table S3). By dichotomizing patients per median *HIF1A* expression, *HIF1A* overexpressed patients had a worse prognosis than the group with low expression ($HR = 2.09$, $P = 5.02 \times 10^{-3}$) (Supplementary Figure S2). *EGLN2* expression was not independently associated with patient overall survival ($P = 0.527$). However, *EGLN2* had a significant interaction effect with *HIF1A* expression on patient outcome ($P = 0.037$). As presented in Figure 5(f), with the decreased expression of *EGLN2*, there was an elevated effect size of *HIF1A* expression on LUAD survival. On the other hand, patients with overexpressed *EGLN2* didn’t retain statistical

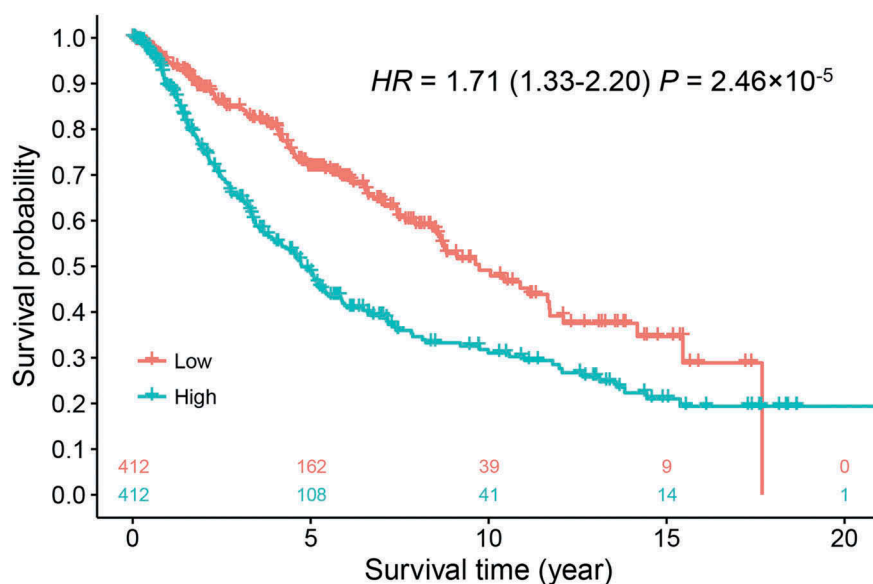


Figure 3. Kaplan-Meier survival curves of different DNA methylation level at cg25923056. Patients were categorized into low- and high-methylation groups using median value of cg25923056. P value was calculated using Cox regression model, and HR indicates hazard ratio.

significance between *HIF1A* expression and overall survival.

We additionally assessed the effect of cg25923056_{EGLN2} on LUAD survival in subgroup patients with different demographic and clinical variables. Almost all these associations remained significant, except some subgroups with small sample size (Supplementary Figure S3).

Discussion

Several epigenetic studies of lung cancer prognosis have identified potential biomarkers relevant to the etiology of NSCLC [8,9]. To the best of our knowledge, this is the first multi-center integrating five independent cohorts and large-scale integration analysis of DNA methylation alterations and expression at the *EGLN* gene family in early-stage NSCLC, as well as association analysis with *HIF1A* expression. Weighted random forest (Ranger) was used to screen DNA methylation CpG probes as well as a survival classification tree to improve statistical power and reveal potential interactions. We identified one probe cg25923056_{EGLN2}, located at the 1st exon region of *EGLN2*, as a biomarker for the prognosis of early-stage LUAD. Nevertheless, no promising individual CpG probe was identified for LUSC, which may be due to underlying epigenetic heterogeneity between LUAD and LUSC [24,25] or to the low power

resulting from the small sample size of LUSC. Previous studies have found that DNA methylation of cg25923056_{EGLN2} was associated with the variant of rs7937 in chronic obstructive pulmonary disease (COPD) patients [26]. Moreover, this association was also reported in a study on meQTLs in blood across the human life course [27]. However, our findings extended the function of cg25923056_{EGLN2} in LUAD patients. High DNA methylation of cg25923056_{EGLN2} is associated with poor LUAD prognosis. The association of promoter DNA methylation with transcriptional silencing is well recognized. Moreover, DNA methylation of the transcription start site (TSS), in the region of the first exon, is much more tightly correlated with transcriptional silencing [28]. Our study consistently found that hypermethylation at cg25923056_{EGLN2} down-regulated the corresponding gene *EGLN2* expression in tumor tissues.

EGLN2 encodes the oxygen-sensing enzyme prolyl hydroxylase 1 (PHD1) responsible for mediating the HIF-1 α degradation and related to tumor progression. *EGLN2* is implicated as a tumor suppressor, since its overexpression could inhibit the tumor growth in colon cancer cell [20]. Additionally, in pancreatic adenocarcinomas, absence of *EGLN2* expression was significantly associated with perineural invasion [29]. In lung carcinoma cells, overexpression of *EGLN2* induces

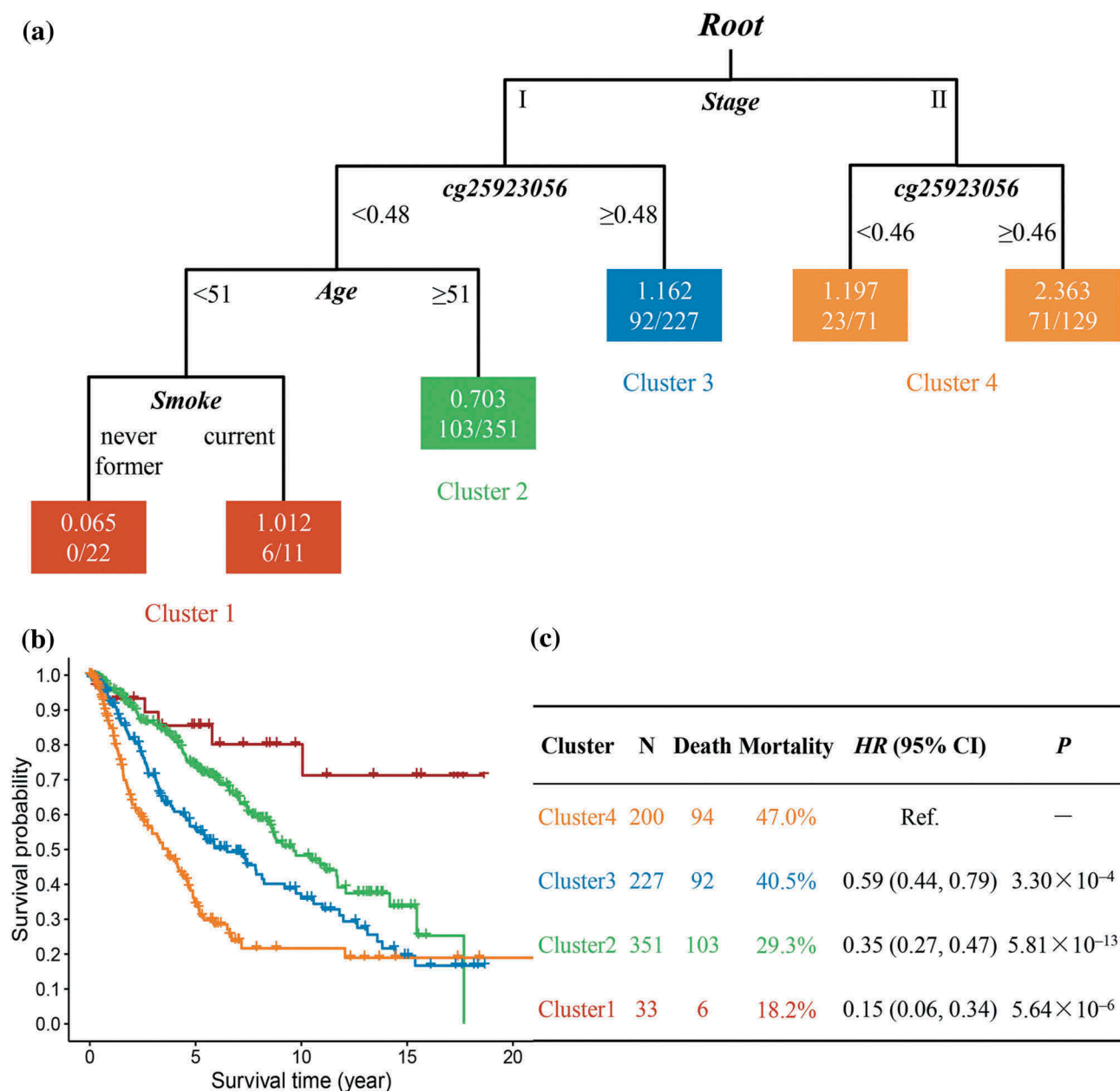


Figure 4. Survival classification tree for lung adenocarcinomas (LUAD). Survival classification tree was built with *cg25923056* as well as covariates in the combined data (a), which identified four clusters with significantly different survival curves (b). Cox regression model was used to compare the outcomes among clusters (cluster 4 as reference) and represented by hazard ratio (HR), 95% confidence interval (95%CI), and the *P* value (c).

cell cycle arrest and suppresses proliferation [30]. Although *EGLN2* has been involved in many cancers, the mechanisms involved are not fully understood.

HIF-1 α is an important regulator in tumor angiogenesis and distant metastases, and plays a pivotal role in the cellular response to tumor hypoxia which represents a major obstacle to the success of radiotherapy and chemotherapy [31]. *HIF1A* is overexpressed in many human cancers and has been associated

consistently with a poor prognosis, including colorectal, oropharyngeal cancers [32–34]. Here, we provide further evidence that this association appeared to be generalized to NSCLC patient. Inhibition of HIF-1 α activity has already become an effective anti-tumor therapy for various tumors and research effort to develop therapeutic drugs have been ongoing for many years but still requires more selectivity and effectiveness [35]. Moreover, DNA methylation of *HIF1A* were not significantly associated with the

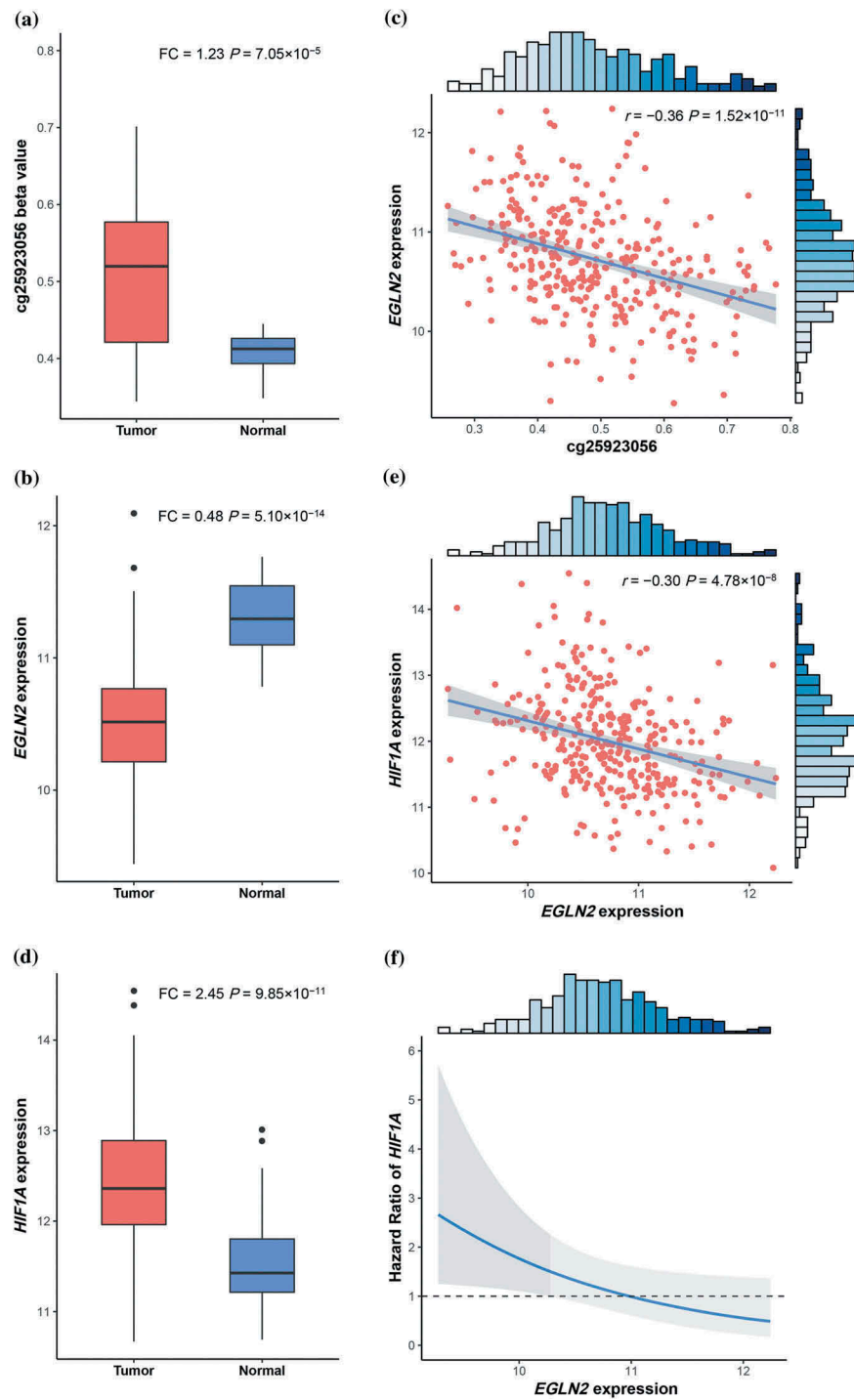


Figure 5. DNA methylation and gene expression analysis for *EGLN2* and *HIF1A*. (a) *cg25923056* methylation differential analysis between tumor and adjacent normal tissues. (b) *EGLN2* expression differential analysis between tumor and adjacent normal tissues. (c) Association between *cg25923056* methylation and *EGLN2* expression. (d) *HIF1A* expression differential analysis between tumor and adjacent normal tissues. (e) Association between *EGLN2* expression and *HIF1A* expression. (f) Hazard ratio of *HIF1A* expression estimated based on different level of *EGLN2* expression. The shallow area represents 95% confidence interval and dark grey area means significant. Gene expression was log2 transformed before analysis. For differential analysis, FC indicates fold change, and P value was calculated using paired student's t test. For correlation analysis, correlation coefficients (r) and hypothesis tests are based on Pearson correlation tests. The histograms on X-axis and Y-axis represent their distributions.

prognosis of lung cancer patients, which suggested that the effect of *HIF1A* expression on overall survival might be modified by other pathways.

Meanwhile, our results indicated that there might be a pathway that possibly accounts for the mechanism of *EGLN2* involved in LUAD: hypermethylation at *cg25923056_{EGLN2}* could suppress *EGLN2* expression, further lead to high *HIF1A* expression and result in a poor prognosis (Figure 6). Our findings are consistent with previous functional studies of *EGLN2* and *HIF1A*. We found that the HR of *HIF1A* expression did not retain statistical significance in patients with overexpressed *EGLN2*, which might result from high expression of *EGLN2* patients with low expression of *HIF1A* and a relatively good prognosis. Moreover, experiments both in vivo and in vitro have confirmed that overexpression in *EGLN2* can inhibit the stabilization of HIF-1 α after hypoxia and inhibit tumor growth [20]. Thus, our study provides evidence for potential development of HIF-1 α inhibitors in LUAD therapy by decreasing DNA methylation of *cg25923056_{EGLN2}*. However, the causation across this path cannot be concluded, which need further exclusive study (e.g. Mendelian randomization analysis) to confirm.

We acknowledge some limitations of our study. First, the censored rate of TCGA cohort is relatively high, which may result in loss of statistical power. However, the association between *cg25923056_{EGLN2}* and survival remained significant in TCGA, indicating that our results are conservative and robust. And early-stage NSCLC patients could be followed longer to obtain more precise estimates in future. In addition, the association between DNA methylation and the corresponding gene expression lacks biological evidence. DNA methylation is believed to play a crucial role in regulating gene expression [36] and may also influence disease development via gene function [37], cell differentiation, and reprogramming [38]. However, further functional experiments are needed to confirm these associations. Finally, our study was performed mainly in Caucasian populations (89.19%). The findings of this study should be interpreted with caution among other populations.

In conclusion, our study identified that the *EGLN2-HIF1A* axis interacts in affecting the prognosis of LUAD. These results elucidate some of the

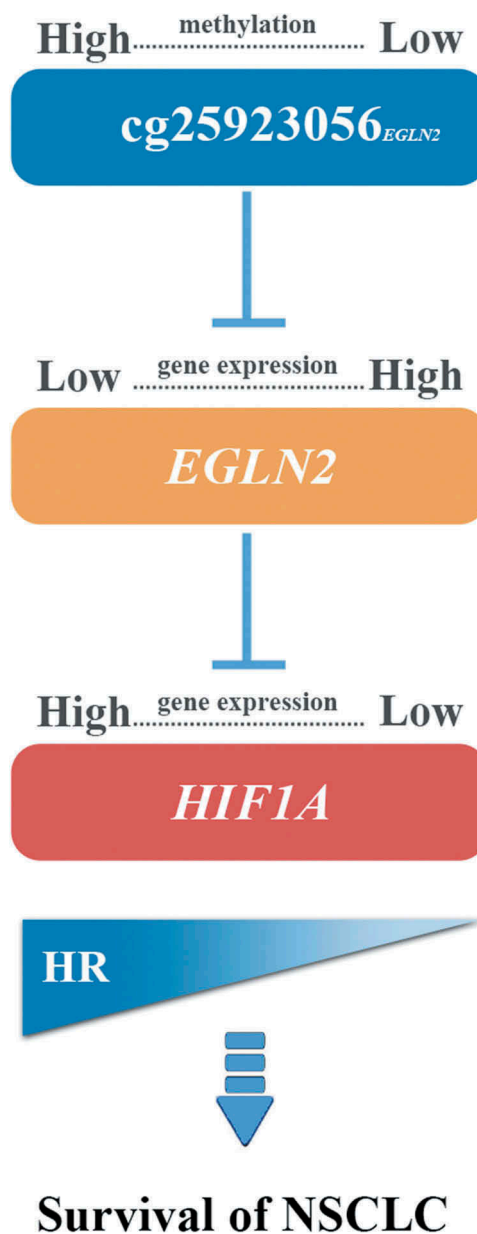


Figure 6. Diagram for DNA methylation-*EGLN2*-*HIF1A*-survival pathway for LUAD patients.

molecular mechanisms underlying LUAD and provide potential reversible therapeutic targets for HIF-1 α inhibitors.

Patients and methods

Study population

Harvard. The Harvard Lung Cancer Study cohort was described previously [39]. All cases were recruited at Massachusetts General Hospital (MGH) since 1992 and were newly diagnosed, histologically confirmed primary NSCLC. Snap-frozen tumor samples were

collected from NSCLC patients during curative surgery with complete resection. There were 151 early-stage (TNM stage I, II) cases selected for this study which had complete survival information. Tumor DNA was extracted from 5- μ m-thick histopathologic sections. Each specimen was evaluated by a MGH pathologist for amount (tumor cellularity > 70%) and quality of tumor cells and histologically classified using WHO criteria.

Spain. Study population was reported previously [40]. In brief, tumors were collected by surgical resection from patients who provided consent and under approval by the institutional review boards. Tumor DNA was extracted from fresh-frozen tumor specimens (10 μ m-thick, tumor cellularity >50%) which were collected by surgical resection. The median clinical follow-up was 7.2 years. The study was approved by the Bellvitge Biomedical Research Institute institutional review board. All patients provided written informed consent.

Norway. As described previously [41], participants were 16 LUAD patients with operable lung cancer tumors who were seen at Oslo University Hospital-Riks hospitalet, Norway, from 2006 to 2011. None of the enrolled patients had received chemotherapy or radiotherapy prior to surgery. Tumor tissues obtained during surgery were snap frozen in liquid nitrogen and stored at -80°C until DNA isolation. Only early-stage (stage I, II) patients were selected for the current study.

Sweden. We collected tumor tissue specimens from 103 early-stage lung cancer patients who underwent an operation at the Skane University Hospital, Lund, Sweden [42]. The study was approved by the Regional Ethical Review Board in Lund, Sweden (Registration no. 2004/762 and 2008/702). All patients provided written informed consent.

TCGA. We used The Cancer Genome Atlas (TCGA) resources for validation, including 332 early-stage lung adenocarcinomas (LUAD) and 285 early-stage lung squamous cell carcinomas (LUSC) which had survival information and common covariates. Level-1 HumanMethylation450 DNA methylation data (image data) of each patient were downloaded on 1 October 2015.

In the TCGA cohort, 328 lung adenocarcinoma (LUAD) patients had complete mRNA sequencing data. TCGA mRNA sequencing data processing and quality control was done by the TCGA workgroup. Raw counts were normalized using RNA Sequencing by Expectation Maximization (RSEM). Level-3 (gene level) gene quantification data were downloaded from TCGA data portal and were further checked for quality. Besides, we extracted 29 early-stage LUAD patients from the TCGA cohort with both tumor and adjacent normal tissues DNA methylation data and 57 early-stage LUAD patients with both tumor and adjacent normal tissues gene expression data for differential methylation and differential expression analysis, respectively. Expression of *EGLN2* genes was extracted and log₂-transformed before analysis.

Quality control procedures

DNA methylation was profiled using Infinium HumanMethylation450 BeadChips (Illumina Inc., San Diego, CA, USA) for all patients. All centers followed the same quality control (QC) procedures before association studies. Raw image data were transformed into beta values to perform background subtraction and control normalization. Unqualified probes were excluded if they met either one of the following criteria: (i) failed detection $P > 0.05$ over 5% of patients; (ii) coefficient of variance (CV) < 5%; (iii) methylated or unmethylated in all samples; (iv) common single nucleotide polymorphisms (SNP) located in the probe sequence or 10-bp flanking regions; (v) cross-reactive probes or cross-hybridizing probes; (vi) or did not pass quality control in all centers. Samples with >5% undetectable probes were excluded. Methylation signals were further processed for quantile normalization, design bias correction for type I and II probes, and batch effects adjustment. Details of QC processes are described in Supplementary Figure S4.

Statistical analysis

Continuous variables were summarized as mean \pm standard deviation (SD), and categorized variables were described by frequency (n) and proportion (%). We used paired Student's t-test to compare the differential expression values and DNA methylation

beta values between tumor and adjacent normal tissues. We used Pearson correlation (r) to explore relationships between DNA methylation and gene expression. False-discovery-rate (FDR) correction q -value was used to adjust for multiple comparisons. Statistical analyses were performed using R version 3.4.4 (The R Foundation of Statistical Computing).

Among LUAD and LUSC patients, we employed Ranger, a weighted version of random forest, in the discovery and the validation set, to evaluate the importance of each individual DNA methylation CpG probe with R package *ranger*. A weight of 100% was given to each covariate to ensure a 100% chance to be selected into each tree. Variable importance score (VIS) for the 34 CpG probe in *EGLNs* was estimated and ranked in a descending order. The sliding windows sequential forward feature selection method (SWSFS) was used to identify the top important CpG probes [43]. The SWSFS method includes the CpG probes one by one to the random forest (RF) model by the order of VIS. Then, we plotted the ‘out of bagging (OOB)’ error, which measured the performance of each model consisting of a specific number of CpG probes. The top potential CpG probes were screened out for further analysis when the RF model having the lowest error rate. CpG probes that were in tops in both discovery and validation set were identified as candidates.

Then, these candidate CpG probes were further evaluated with a two-stage design, as well as a series of stratified analyses. In the discovery phase, we applied a Cox proportional hazards model adjusted for age, gender, smoking status, clinical stage and study center to test the association between a DNA methylation CpG probe and overall survival in LUAD and LUSC patients, respectively. The hazard ratio (HR) per 1% methylation increment and 95% confidence interval (CI) were estimated for each probe. Probes with $FDR-q \leq 0.05$ were further replicated in TCGA. Robustly significant probes were finally retained if they met the all following criteria: (i) $P \leq 0.05$ in the validation phase; (ii) consistent effect direction in both discovery and validation phases.

In addition, survival tree construction was done using the recursive partitioning and regression tree (RPART) [44], which extends the classification and regression trees (CART), to identify clusters with heterogeneous survival outcome with R package

rpart. Kaplan-Meier method was used to illustrate the survival curves of different clusters.

Acknowledgments

The authors thank TCGA for contributing clinical, DNA methylation, and RNA sequencing data, as well as all study subjects who participated in the five study cohorts.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by National Key Research and Development Program of China [2016YFE0204900]; National Natural Science Foundation of China [81530088 and 81473070]; National Institutes of Health [CA209414, CA092824, and ES000002]; Natural Science Foundation of the Jiangsu Higher Education Institutions of China [18KJB310011 and 14KJA310002]; China Postdoctoral Science Foundation [2018M633767] and the Priority Academic Program Development of Jiangsu Higher Education Institutions. Y. W. and R.Z. were partially supported by the Outstanding Young Teachers Training Program of Nanjing Medical University.

Declarations

Ethics approval and consent to participate

The Harvard study protocol was approved by the Institutional Review Boards at Harvard School of Public Health and MGH. The Spain study was approved by the Bellvitge Biomedical Research Institute Institutional Review Board. The Norway project was approved by Oslo University Institutional Review Board and Regional Ethics Committee (S-05307). The Sweden study was approved by the Regional Ethical Review Board in Lund, Sweden (registration no. 2004/762 and 2008/702). All patients provided written informed consent.

Consent for publication

All participants or their surrogate care providers gave written informed consent. All authors have reviewed the manuscript and consented for publication.

Availability of data and materials

TCGA: <https://tcga-data.nci.nih.gov>; now hosted at GDC: <https://portal.gdc.cancer.gov>.

Authors' contributions

R.Z., L.L., J.H., Y.W., F.C. and D.C.C. contributed to the study design. R.Z., S.M., T.F., M.M.B., A.K., M.P., J.S., A. H., M.E., A.S. and L.S. contributed to data and sample collection. R.Z., L.L., J.H. performed statistical analysis, interpretation and drafted the manuscript. C.C., D.Y., W. D., X.D., S.S. and Y.G. revised the manuscript. All authors contributed to critical revision of the final manuscript and approved the final version of the manuscript. Financial support and study supervision: F.C. and D.C.C.

ORCID

Sipeng Shen  <http://orcid.org/0000-0003-0436-4736>
 Sebastian Moran  <http://orcid.org/0000-0003-4192-8983>
 Åslaug Helland  <http://orcid.org/0000-0002-5520-0275>
 David C. Christiani  <http://orcid.org/0000-0002-0301-0242>

References

- [1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin.* 2018;68:7–30.
- [2] Chen Z, Fillmore CM, Hammerman PS, et al. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer.* 2014;14:535–546.
- [3] Hirsch FR, Scagliotti GV, Mulshine JL, et al. Lung cancer: current therapies and new targeted treatments. *Lancet.* 2016;389:299.
- [4] Goldstraw P, Chansky K, Crowley J, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (Eighth) edition of the TNM classification for lung cancer. *J Thorac Oncol.* 2016;11:39–51.
- [5] Tang S, Pan Y, Wang Y, et al. Genome-wide association study of survival in early-stage non-small cell lung cancer. *Ann Surg Oncol.* 2015;22:630–635.
- [6] Egger G, Liang G, Aparicio A, et al. Epigenetics in human disease and prospects for epigenetic therapy. *Nature.* 2004;429:457–463.
- [7] Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer.* 2004;4:143–153.
- [8] Shen S, Zhang R, Guo Y, et al. A multi-omic study reveals BTG2 as a reliable prognostic marker for early-stage non-small cell lung cancer. *Mol Oncol.* 2018;12:913–924.
- [9] Wei Y, Liang J, Zhang R, et al. Epigenetic modifications in KDM lysine demethylases associate with survival of early-stage NSCLC. *Clin Epigenetics.* 2018;10:41.
- [10] Györfy B, Bottai G, Fleischer T, et al. Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes. *Int J Cancer.* 2015;138:87–97.
- [11] Shen S, Wang G, Shi Q, et al. Seven-CpG-based prognostic signature coupled with gene expression predicts survival of oral squamous cell carcinoma. *Clin Epigenetics.* 2017;9:88.
- [12] Vaupel P, Mayer A. Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metast Rev.* 2007;26:225–239.
- [13] Keith B, Simon MC. Hypoxia-inducible factors, stem cells, and cancer. *J Cell Mol Med.* 2009;13:4319–4328.
- [14] Semenza GL. HIF-1 mediates metabolic responses to intratumoral hypoxia and oncogenic mutations. *J Clin Invest.* 2013;123:3664–3671.
- [15] Epstein AC, Gleadle JM, McNeill LA, et al. *C. elegans* EGL-9 and mammalian homologs define a family of dioxygenases that regulate HIF by prolyl hydroxylation. *Cell.* 2001;107:43–54.
- [16] Maxwell PH, Wiesener MS, Chang GW, et al. The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. *Nature.* 1999;399:271–275.
- [17] Xie G, Zheng L, Ou J, et al. Low expression of prolyl hydroxylase 2 is associated with tumor grade and poor prognosis in patients with colorectal cancer. *Exp Biol Med.* 2012;237:860–866.
- [18] Couvelard A, Deschamps L, Rebours V, et al. Overexpression of the oxygen sensors PHD-1, PHD-2, PHD-3, and FIH Is associated with tumor aggressiveness in pancreatic endocrine tumors. *Clin Cancer Res.* 2008;14:6634–6639.
- [19] Peurala E, Koivunen P, Bloigu R, et al. Expressions of individual PHDs associate with good prognostic factors and increased proliferation in breast cancer patients. *Breast Cancer Res Treat.* 2012;133:179–188.
- [20] Erez N, Milyavsky M, Eilam R, et al. Expression of prolyl-hydroxylase-1 (PHD1/EGLN2) Suppresses hypoxia inducible factor-1 α activation and inhibits tumor growth. *Cancer Res.* 2003;63:8777.
- [21] Rawluszko AA, Bujnicka KE, Horbacka K, et al. Expression and DNA methylation levels of prolyl hydroxylases PHD1, PHD2, PHD3 and asparaginyl hydroxylase FIH in colorectal cancer. *BMC Cancer.* 2013;13:526.
- [22] Huang KT, Mikeska T, Dobrovic A, et al. DNA methylation analysis of the HIF-1 α prolyl hydroxylase domain genes PHD1, PHD2, PHD3 and the factor inhibiting HIF gene FIH in invasive breast carcinomas. *Histopathology.* 2010;57:451–460.
- [23] Place TL, Fitzgerald MP, Venkataraman S, et al. Aberrant promoter CpG methylation is a mechanism for impaired PHD3 expression in a diverse set of malignant cells. *PLoS One.* 2011;6:e14617.
- [24] Devarakonda S, Morgensztern D, Govindan R. Genomic alterations in lung adenocarcinoma. *Lancet Oncol.* 2015;16:e342–e351.

- [25] Network CGA. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;517:576.
- [26] Nedeljkovic I, Lahousse L, Carnero ME, et al. COPD GWAS variant at 19q13.2 in relation with DNA methylation and gene expression. *Hum Mol Genet*. 2018;27:396–405.
- [27] Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol*. 2016;17:61.
- [28] Brenet F, Moh M, Funk P, et al. DNA methylation of the first exon is tightly linked to transcriptional silencing. *PLoS One*. 2012;6:e14524.
- [29] Gossage L, Zaitoun A, Fareed KR, et al. Expression of key hypoxia sensing prolyl-hydroxylases PHD1, -2 and -3 in pancreaticobiliary cancer. *Histopathology*. 2010;56:908–920.
- [30] Xiao X, Haibo X, Fangbao D, et al. Over-expression of prolyl hydroxylase-1 blocks NF- κ B-mediated cyclin D1 expression and proliferation in lung carcinoma cells. *Cancer Genet*. 2014;207:188–194.
- [31] Triner D, Shah YM. Hypoxia-inducible factors: a central link between inflammation and cancer. *J Clin Invest*. 2016;126:3689–3698.
- [32] Zhong H, De Marzo AM, Laughner E, et al. Overexpression of hypoxia-inducible factor 1 α in common human cancers and their metastases. *Cancer Res*. 1999;59:5830–5835.
- [33] Baba Y, Nosho K, Shima K, et al. HIF1A overexpression is associated with poor prognosis in a cohort of 731 colorectal cancers. *Am J Pathol*. 2010;176:2292–2301.
- [34] Aebersold DM, Burri P, Beer KT, et al. Expression of hypoxia-inducible factor-1 α : a novel predictive and prognostic parameter in the radiotherapy of oropharyngeal cancer. *Cancer Res*. 2001;61:2911–2916.
- [35] Semenza GL. Targeting HIF-1 for cancer therapy. *Nat Rev Cancer*. 2003;3:721–732.
- [36] Bird A. Perceptions of epigenetics. *Nature*. 2007;447:396.
- [37] Schübeler D. Function and information content of DNA methylation. *Nature*. 2015;517:321.
- [38] Khavari DA, Sen GL, Rinn JL. DNA methylation and epigenetic control of cellular differentiation. *Cell Cycle*. 2010;9:3880–3883.
- [39] Asomaning K, Miller DP, Liu G, et al. Second hand smoke, age of exposure and lung cancer risk. *Lung Cancer*. 2008;61:13–20.
- [40] Sandoval J, Mendez-Gonzalez J, Nadal E, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J Clin Oncol*. 2013;31:4140.
- [41] Bjaanaes MM, Fleischer T, Halvorsen AR, et al. Genome-wide DNA methylation analyses in lung adenocarcinomas: association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol Oncol*. 2016;10:330–343.
- [42] Karlsson A, Jonsson M, Lauss M, et al. Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin Cancer Res*. 2014;20:6127–6140.
- [43] Jiang R, Tang W, Wu X, et al. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*. 2009;10:1–12.
- [44] Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the rpart routine Technical Report 61. Rochester Mayo Foundation. 1997.