



## ORIGINAL RESEARCH

# Non-small Cell Lung Cancer Epigenomes Exhibit Altered DNA Methylation in Smokers and Never-smokers



Jennifer A. Karlow<sup>1,2,#,§</sup>, Erica C. Pehrsson<sup>1,2,#</sup>, Xiaoyun Xing<sup>1,2</sup>, Mark Watson<sup>3</sup>,  
Siddhartha Devarakonda<sup>4</sup>, Ramaswamy Govindan<sup>4</sup>, Ting Wang<sup>1,2,5,\*</sup>

<sup>1</sup> Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA

<sup>2</sup> The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA

<sup>3</sup> Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110, USA

<sup>4</sup> Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA

<sup>5</sup> McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA

Received 16 March 2022; revised 11 February 2023; accepted 14 March 2023

Available online 22 September 2023

Handled by Hui Shen

## KEYWORDS

Non-small cell lung cancer;  
Epigenomics;  
M&M;  
methylCRF;  
DNA methylation

**Abstract** Epigenetic alterations are widespread in cancer and can complement genetic alterations to influence cancer progression and treatment outcome. To determine the potential contribution of **DNA methylation** alterations to tumor phenotype in **non-small cell lung cancer** (NSCLC) in both smoker and never-smoker patients, we performed genome-wide profiling of DNA methylation in 17 primary NSCLC tumors and 10 matched normal lung samples using the complementary assays, methylated DNA immunoprecipitation sequencing (MeDIP-seq) and methylation sensitive restriction enzyme sequencing (MRE-seq). We reported recurrent methylation changes in the promoters of several genes, many previously implicated in cancer, including *FAM83A* and *SEPT9* (hypomethylation), as well as *PCDH7*, *NKX2-1*, and *SOX17* (hypermethylation). Although many methylation changes between tumors and their paired normal samples were shared across patients, several were specific to a particular smoking status. For example, never-smokers displayed a greater proportion of hypomethylated differentially methylated regions (hypoDMRs) and a greater number of recurrently hypomethylated promoters, including those of *ASPSCR1*, *TOP2A*, *DPP9*, and *USP39*, all previously linked to cancer. Changes outside of promoters were also widespread and

\* Corresponding author.

E-mail: [twang@genetics.wustl.edu](mailto:twang@genetics.wustl.edu) (Wang T).

# Equal contribution.

§ Current address: Department of Pathology, Dana-Farber Cancer Institute / Harvard Medical School, Boston, MA 02215, USA.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2023.03.006>

1672-0229 © 2023 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

often recurrent, particularly methylation loss over repetitive elements, highly enriched for ERV1 subfamilies. Recurrent hypoDMRs were enriched for several transcription factor binding motifs, often for genes involved in signaling and cell proliferation. For example, 71% of recurrent promoter hypoDMRs contained a motif for NKX2-1. Finally, the majority of DMRs were located within an active chromatin state in tissues profiled using the Roadmap Epigenomics data, suggesting that methylation changes may contribute to altered regulatory programs through the adaptation of cell type-specific expression programs.

## Introduction

Lung cancer is the leading cause of cancer deaths in the USA, with a 5-year relative survival rate of 22.9% [1]. Non-small cell lung cancer (NSCLC) represents 85% of lung cancer cases and is highly heterogeneous, comprising lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and large cell carcinoma. In addition to cytotoxic chemotherapies, currently approved therapeutics for NSCLC include inhibitors of EGFR, ALK, ROS1, VEGF, and other proteins, as well as checkpoint blockade therapies targeting PD-1 and PD-L1 [2]. LUAD and LUSC both exhibit high somatic mutation rates relative to other cancers [3].

Although most NSCLC cases are associated with tobacco smoking [4], the proportion of NSCLC cases attributed to never-smokers is rising and now represents 10%–40% of NSCLC cases worldwide [5]. Lung cancer in never-smokers is linked to genetic susceptibilities and certain environmental exposures [4]. It is most common among women and East Asians, and most frequently of the LUAD subtype [4]. Never-smoker tumors exhibit distinct molecular characteristics, including a higher frequency of *EGFR* and *HER2* mutations and *ALK/RET/ROS* fusions that improve their response to targeted therapy. In contrast, smokers exhibit a higher rate of mutations in *KRAS*, *TP53*, *STK11*, *BRAF*, *JAK2*, *JAK3*, and mismatch repair genes [5,6]. The mutation burden in smokers is  $> 10\times$  higher [5] and is characterized by C  $>$  A nucleotide transversions caused by direct benzo[*a*]pyrene exposure. This signature is clonal, suggesting that mutations occur before transformation, and correlates with pack years smoked [7]. Smokers also exhibit a higher rate of copy number alterations, non-synonymous mutations, and neoepitopes [7,8], which may explain their greater response to immunotherapy.

In the past, cancer analyses have focused primarily on genomic mutations. However, cancer is also characterized by massive epigenetic dysregulation. Epigenetic modifiers are frequently mutated in cancer, suggesting a role in tumorigenesis. Cancer exhibits global DNA hypomethylation coupled with focal promoter hypermethylation [9], which can silence tumor suppressor genes in lung cancer [6]. In addition, epigenetic alterations can lead to genomic instability, the dysregulation of genomic architecture [10], histone modification spreading, and widespread DNA hypomethylation, resulting in aberrantly activated transposable elements (TEs). Cryptic promoters within de-repressed TEs can drive oncogene expression [11–13] and create chimeric transcripts that may produce neoepitopes [14,15]. Additionally, demethylated endogenous retroviruses produce double-stranded RNA and trigger an antiviral immune response [16,17], potentiating treatment with

checkpoint blockade therapy [18]. In sum, epigenetic alterations can complement genomic alterations and have profound effects on cancer progression and treatment outcome.

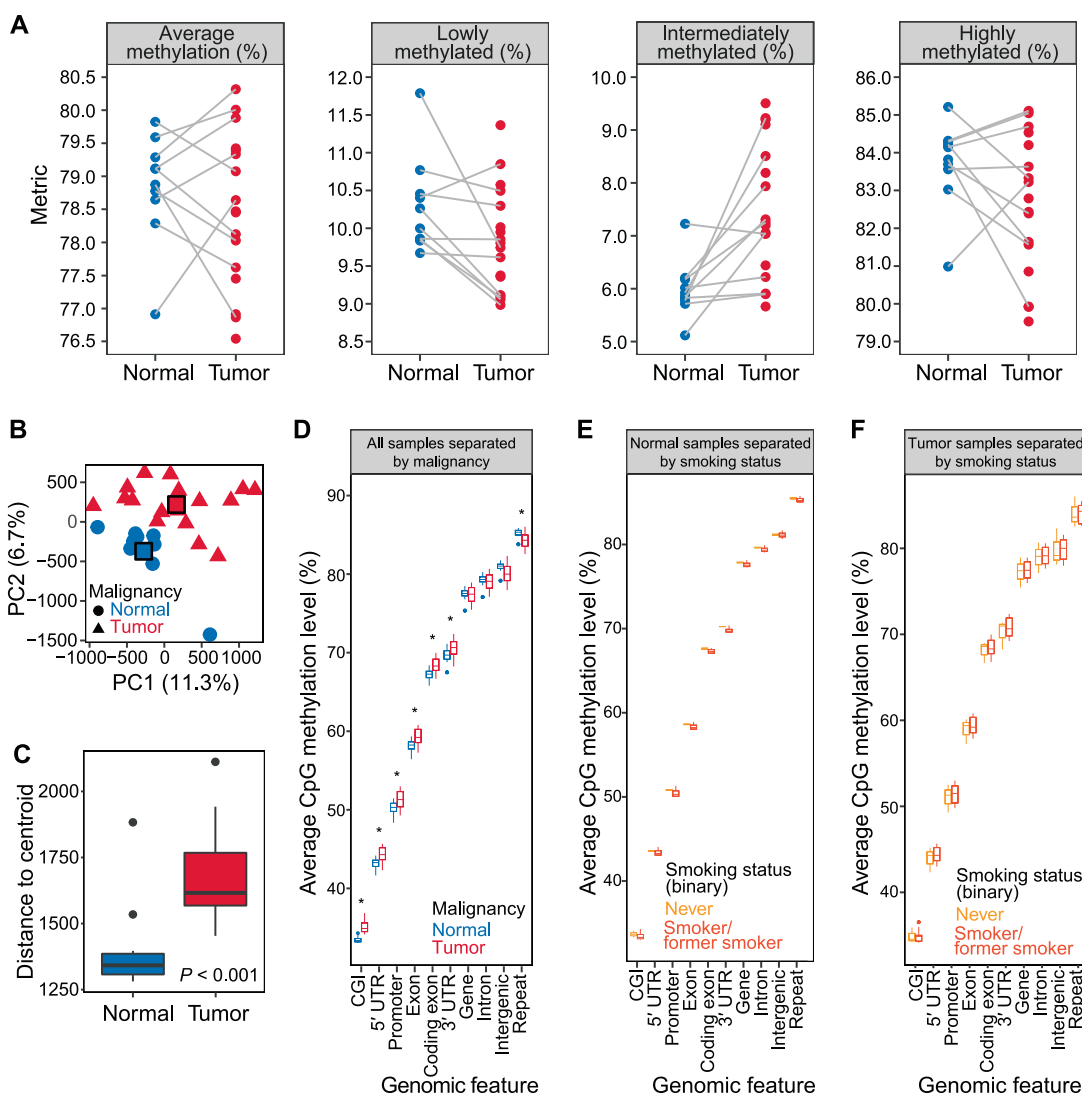
The Cancer Genome Atlas (TCGA) has profiled thousands of NSCLC samples and matched normal lung samples [19,20]. However, DNA methylation data were generated using the Human Methylation 450K array, which covers only a fraction of the CpGs in the genome. To supplement this analysis, we profiled genome-wide DNA methylation in 17 primary LUAD tumors, as well as matched normal lung samples from 10 of the patients (see Materials and methods). Clinicopathologic data are summarized in Table S1. For each sample, we performed methylated DNA immunoprecipitation sequencing (MeDIP-seq) and methylation sensitive restriction enzyme sequencing (MRE-seq), which capture methylated and unmethylated CpGs, respectively (Figure S1). Data generated by these complementary assays were then integrated using methylCRF [21] to estimate the methylation level at over 28 million CpGs genome-wide, resulting in comprehensive, environmentally matched pairs of normal lung and primary tumor methylomes. We also used the M&M algorithm, which integrates MeDIP-seq and MRE-seq from two comparative samples [22], to identify differentially methylated regions (DMRs), providing a complete profile of common DNA methylation alterations across a heterogeneous set of primary NSCLC tumors.

## Results

### Global methylation profiles of primary tumors exhibit systematic deviations from non-malignant lung

We first profiled genome-wide DNA methylation changes between primary NSCLC and patient-matched, histologically non-malignant lung tissues. Although there was substantial variation across samples (Figure S2A), on average, primary tumor samples exhibited a shift in overall methylation density, with fewer highly methylated CpGs and lowly methylated CpGs and a significant increase in intermediately methylated CpGs (Wilcoxon test,  $P = 0.001423$ ; **Figure 1A**, **Figure S2B**; File S1). The genome-wide methylation changes observed in this study are in line with data from other studies (Karlow et al., unpublished data; [23]), and analysis of MRE-seq data alone confirmed that tumors lose methylation over intergenic regions and repeats (Figure S3A–H; File S1), consistent with previous studies [24].

At a finer resolution, local methylation changes separated normal lung samples from tumor samples. Principal component analysis (PCA) on the average methylation level over 1-kb windows separated normal lung samples from primary



**Figure 1 Genome-wide CpG methylation alterations in primary NSCLC**

**A.** Average genome-wide methylation and proportion of CpGs at each methylation level per sample (lowly methylated, < 30% methylation; intermediately methylated, 30%–70% methylation; highly methylated, > 70% methylation). **B.** PCA on normal lung and primary tumor samples using the average methylation level over genome-wide 1-kb windows as features. Group centroids are indicated with squares. Axis titles display the amount of variance explained by each PC. P2385\_N\_UC is the outlier on PC2. **C.** Distance to centroid by sample malignancy (Wilcox test,  $P = 0.0004497$ ). Normal outliers are P4999\_N\_S and P2385\_N\_UC; tumor outlier is P14658\_T\_NS. **D.** Average CpG methylation level across genomic features in all samples, colored by sample malignancy and ordered by median CpG methylation level. \*,  $P < 0.05$  (Wilcox test). **E.** Average CpG methylation level across genomic features in normal samples, separated by smoking status. **F.** Average CpG methylation level across genomic features in tumor samples, separated by smoking status. NSCLC, non-small cell lung cancer; PCA, principal component analysis; PC, principal component; CGI, CpG island; UTR, untranslated region.

NSCLC samples along principal component 1 (PC1) and PC2 (Figure 1B). The degree of correlation between tumors and paired normals was not different when separating patients based on smoking status (Wilcox test,  $P = 1$ ; Figure S4A–C) or tumor stage (early vs. late; Wilcox test,  $P = 1$ ; Figure S4D–G), and was not correlated with tumor purity (Pearson correlation,  $P = 0.6079$ ; Figure S4H–J). The two patients with adenosquamous carcinomas demonstrated a greater divergence from their paired normal samples than

patients with adenocarcinomas, although this also did not meet statistical significance (Wilcox test,  $P = 0.09524$ ; Figure S4K–M). Interestingly, interpatient normal-to-primary tumor distances were comparable to inpatient distances (Wilcox test,  $P = 0.06136$ ; Figure S4N–Q), suggesting that tumors undergo a great deal of epigenetic repatterning relative to non-malignant tissues.

Non-malignant lung tissue was equally homogenous among smokers and never-smokers (Wilcox test,  $P = 0.10052$ ,

distance to centroid; Figure S5A–C), and was significantly more homogenous than primary NSCLC (Wilcoxon test,  $P < 0.001$ , distance to centroid; Figure 1C), suggesting that epigenetic changes during tumor progression can take a number of paths. Primary tumors were also equally homogenous among smokers and never-smokers (Wilcoxon test,  $P = 0.8363$ , distance to centroid; Figure S5D–F), across tumor stages (early vs. late; Wilcoxon test,  $P = 0.8981$ , distance to centroid; Figure S5G–I), and across subtypes (adenocarcinoma vs. adenocarcinoma; Wilcoxon test,  $P = 0.7676$ , distance to centroid; Figure S5J–L). No correlation was found between the degree of variation of tumor samples and their tumor purity (Pearson correlation,  $P = 0.7664$ ; Figure S5M–O).

Finally, we looked at methylation alterations in primary NSCLC by genomic location (Figure 1D). Compared to normal lung, primary NSCLC gained CpG methylation over promoters, exons [both untranslated regions (UTRs) and coding exons], and CpG islands (CGIs), but lost methylation over repeats (Wilcoxon test,  $P < 0.05$ ), in agreement with previous studies [24]. Interestingly, the two normal lung samples from never-smokers tended to exhibit higher CpG methylation across each profiled genomic feature compared to normal lung samples from smokers, although the small sample sizes prohibited the significance (Figure 1E). On the other hand, tumors separated by smoking status did not show any differences in methylation across genomic features (Figure 1F).

#### DMRs are largely patient-specific with several recurrent examples pertaining to a particular smoking status

Next, we identified DMRs across the genome between patient-matched normal lung and primary NSCLC, as well as between normal lung samples and between normal lung and tumor samples from different patients. A  $Q$  value threshold of 0.001 was selected (Figure S6A–G; File S1).

The number of DMRs per patient was highly variable, ranging from 239 to 16,676 (Figure 2A), independent of tumor purity (Figure S7A). However, all but one patient-matched comparison exhibited more DMRs than that between normal lung samples (59–1708 DMRs in normal vs. normal comparisons) (Figure S7B). The tumor in question (Patient\_4999) was a bronchioloalveolar carcinoma, a minimally invasive subtype of NSCLC that may have accumulated fewer methylation changes due to a slower rate of cell division [25]. Based on genome-wide methylation profiles, normal lung samples were more homogenous than primary NSCLC samples (Figure S7C–F; File S1). The proportion of DMRs that were hypomethylated in normal lung compared to tumor also varied by patient, from 2% to 73% (Figure 2A), and did not correlate with the total number of DMRs (Pearson correlation,  $P > 0.5$ ).

To determine whether smoking created DNA methylation field effects prior to malignant transformation, we identified DMRs between normal lung samples based on smoking category ( $n = 1214$  unique DMRs). Normal samples of the same smoking status did not have fewer DMRs, as would be expected if there were characteristic changes in CpG methylation due to smoking (Wilcoxon test,  $P > 0.05$ ) (Figure S8). Furthermore, DMRs which were found between never-smoker and smoker normal samples, but not between the two never-

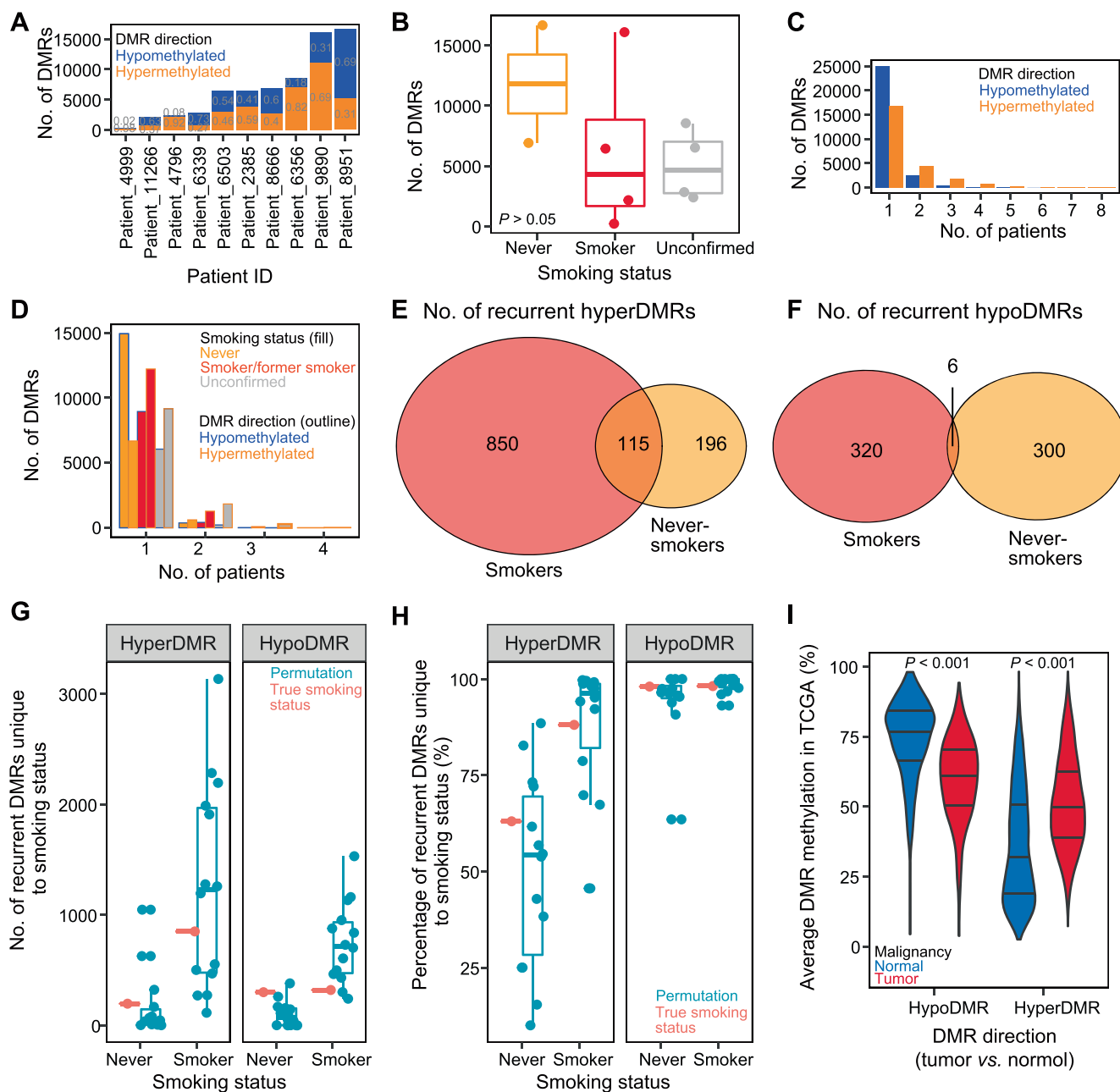
smoker normal samples ( $n = 785$ ), were not significantly enriched in any Gene Ontology (GO) biological processes. However, there were several examples of methylation loss in only smoker samples, including DMRs in the promoters of *PM20D1* (chr1:205818500–205819000) and *ARRB2* (chr17:4612500–4613000). *ARRB2* variants have been previously linked to smoking status [26], and the depletion of the gene promoted lung cancer growth in a mouse model [27]. Additionally, there were DMRs overlapping with the gene bodies of *METAP1D*, *RGPD8*, and *PKP3*, which have been shown to be up-regulated in LUAD and promote cancer growth [28], and whose methylation status has been linked to *in utero* nicotine exposure [29]. The number of DMRs identified in non-malignant lung tissue specifically between smokers and never-smokers was not significantly different from the number of DMRs identified among the entire cohort, suggesting that unlike somatic mutations, smoking history may not contribute to a systematic difference in methylation status.

Although underpowered to reach statistical significance (Wilcoxon test,  $P = 0.2667$ ), there were fewer DMRs between paired normal and tumor samples from smokers ( $n = 4$  patients) than those from never-smokers ( $n = 2$  patients), with the exception of Patient\_9890 (Figure 2B). Although never-smoker patients tended to have higher numbers of both hypermethylated differentially methylated regions (hyperDMRs) and hypomethylated differentially methylated regions (hypoDMRs), they tended to have a higher percentage of hypoDMRs relative to smokers (Figure S9A–C; Wilcoxon test,  $P > 0.05$ ). The number of DMRs as well as the percentage of hypoDMRs did not correlate with tumor stage (Figure S9D and E), subtype (Figure S9F and G), or total methylation change (Figure S9H and I).

The majority of patient-matched DMRs were exclusive to a single patient (Figure 2C), with 11% (3158/28,204) of hypoDMRs and 31% (7457/24,242) of hyperDMRs shared by multiple individuals. However, hypoDMRs were shared by up to six patients, and hyperDMRs were shared by up to eight. When considering patients according to smoking status, 8% (612/7289) of hyperDMRs and 2% (378/15,305) of hypoDMRs were shared by both never-smokers, while 10% (1380/13,593) of hyperDMRs and 5% (432/9377) of hypoDMRs were shared by at least two smokers (Figure 2D). Of the DMRs identified in more than one patient (recurrent), 850 hyperDMRs and 320 hypoDMRs, respectively, were shared by at least two smokers and absent in never-smokers, 196 hyperDMRs and 300 hypoDMRs were shared by both never-smokers and absent in smokers, and 115 hyperDMRs and 6 hypoDMRs were recurrent in both smokers and never-smokers (Figure 2E and F). Interestingly, the number of DMRs recurrent in never-smokers but not recurrent in smokers was generally greater than when randomly subsetting patients, suggesting that these recurrent epigenetic changes may contribute to never-smoker cancer progression (Figure 2G and H).

Finally, we confirmed the DMRs observed in this study using a much larger cohort from TCGA. Although only 14% of patient-matched hypoDMRs and 52% of hyperDMRs overlapped with a CpG from the 450K array ( $n = 3815$  and 12,668, respectively), the average methylation level of the DMRs in TCGA LUAD and matched normal lung samples recapitulated the DMR directions in our study (Figure 2I, Figure S10A). Furthermore, hyperDMRs identified in smoker patients from our study (at all or recurrently) exhibited





**Figure 2 DMRs between normal lung and primary NSCLC**

**A.** Number of DMRs between paired normal and tumor samples in each patient. The proportions of hypoDMRs and hyperDMRs are also shown. **B.** Number of patient-matched DMRs in never-smoker, smoker, and unconfirmed patients. Wilcox test comparing never-smoker patients ( $n = 2$ ) to smoker patients ( $n = 4$ ) is not significant ( $P = 0.2667$ ). **C.** Number of patient-matched DMRs shared by different numbers of patients by DMR direction. **D.** Number of patient-matched DMRs shared by different numbers of patients by both DMR direction (outline color) and smoking status (fill color). **E.** Number of hyperDMRs recurrent in smokers, never-smokers, or both. **F.** Number of hypoDMRs recurrent in smokers, never-smokers, or both. **G.** and **H.** Number (**G**) and percentage (**H**) of recurrent hyperDMRs (left) or hypoDMRs (right) unique to the specified smoking status. Red indicates values using true patient smoking statuses; blue indicates values using permutations of patients (smoker,  $n = 4$ ; never-smoker,  $n = 2$ ). **I.** Average methylation level of all hypoDMRs or hyperDMRs shared by at least two patients across all TCGA LUAD samples, split by sample type. Violin plot lines indicate quartiles.  $P < 0.001$  (Wilcox test) between LUAD normal lung and primary tumor for both hypoDMRs and hyperDMRs. DMR, differentially methylated region; hyperDMR, hypermethylated differentially methylated region; hypoDMR, hypomethylated differentially methylated region; TCGA, The Cancer Genome Atlas; LUAD, lung adenocarcinoma.

significantly higher methylation levels in TCGA patients with confirmed smoking history compared to never-smokers/patients without data (Figure S10B and C). HypoDMRs found in smoker patients from our study exhibited lower methylation levels in TCGA LUAD samples with confirmed smoking history (Figure S10D) but failed to exhibit a significant difference compared to TCGA LUAD never-smokers/patients without data when restricted to recurrent smoker hypoDMRs (Figure S10E). Differences in average methylation between TCGA LUAD tumor samples from patients with confirmed smoking history and never-smokers/patients without data were not observed over hyperDMRs or hypoDMRs identified in never-smokers from our study (Figure S10F–I), with the exception of significantly lower methylation in TCGA patients with confirmed smoking status over hypoDMRs identified in never-smokers from our study (Figure S10G).

### Recurrent methylation changes are observed at the genome-wide level, including within the promoters of several genes previously implicated in lung cancer

The locations of DMRs between patient-matched primary and non-tumor samples recapitulated recognized patterns. In general, hyperDMRs in tumors (analyzing 500-bp bins containing CpGs) were enriched in promoters, exons, CGIs, and partially methylated domains (PMDs), and were depleted over intergenic regions and repeats, whereas hypoDMRs were enriched to a lesser extent in promoters, exons, CGIs, and PMDs (Figure 3A, Figure S11A). Stratifying DMRs based on identification in smokers, never-smokers, or both revealed similar trends. Enrichment was slightly higher over promoters, exons, and CGIs for hyperDMRs identified in smokers than in never-smokers, whereas hypoDMRs identified in never-smokers demonstrated a slightly higher enrichment over those regions (Figure S11B and C). Enrichment over genic features was stronger for DMRs shared between patients than for patient-exclusive DMRs, where the pattern was again more pronounced for hyperDMRs (Figure S11D).

The enrichment of DMRs over genic features also varied by transcript type. Figure S11E displays DMR enrichment over protein-coding, antisense, long intergenic non-coding RNA (lincRNA), and processed transcripts, which together comprise 72% of all GENCODE-annotated transcripts. HyperDMRs were more strongly enriched over the promoters and exons of protein-coding and antisense transcripts, with a 5' vs. 3' UTR bias evident only in protein-coding transcripts. In contrast, hypoDMRs were most strongly enriched over lincRNA, although still to a lesser degree than hyperDMRs. Stratifying DMRs based on identification in smokers, never-smokers, or both revealed similar enrichment trends (Figure S11F). Finally, hyperDMRs were more strongly enriched over protein-coding transcripts with a CGI in the promoter, particularly over the promoter (Figure S11G), where smoker hyperDMRs showed a slightly higher enrichment in CGI-associated promoter transcripts, and never-smoker hyperDMRs showed a slightly higher enrichment in non-CGI-associated transcripts (Figure S11H).

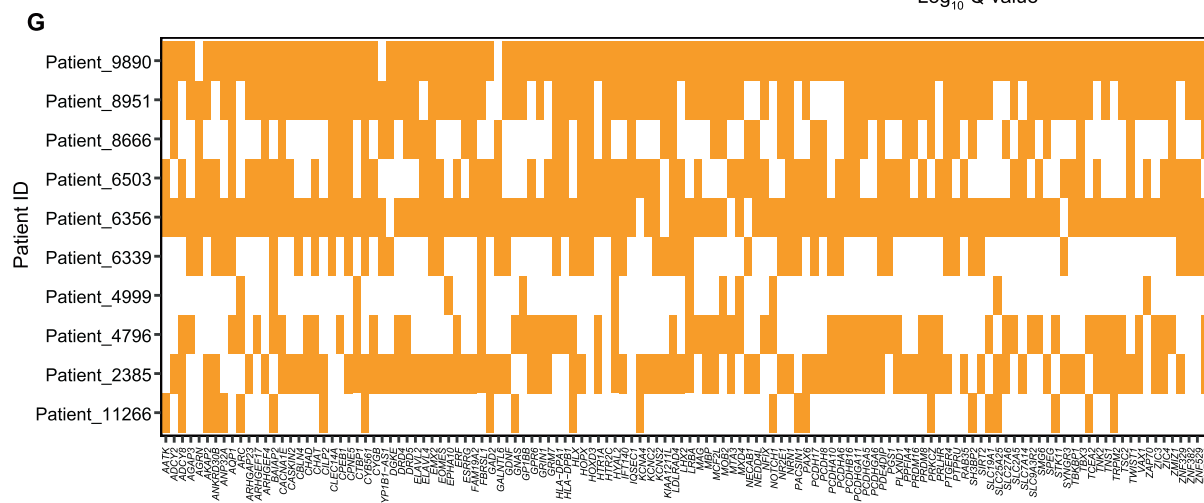
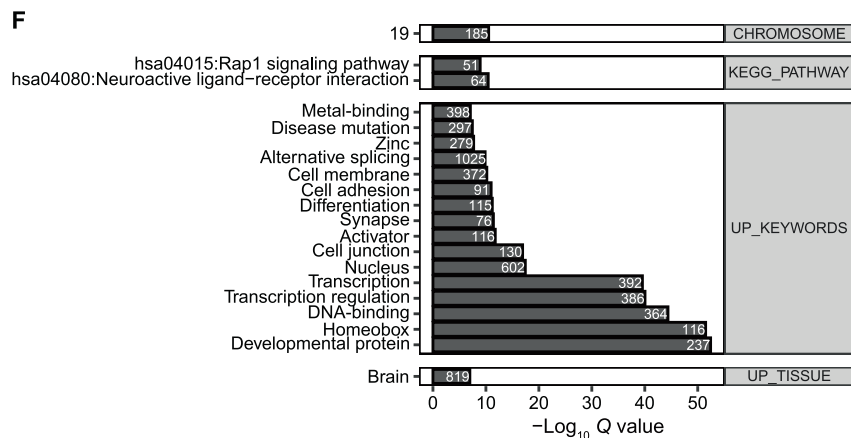
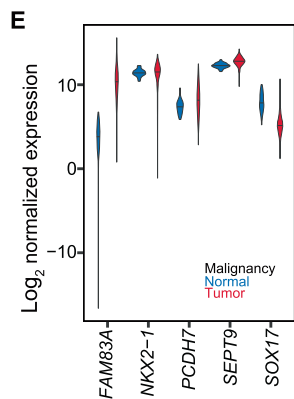
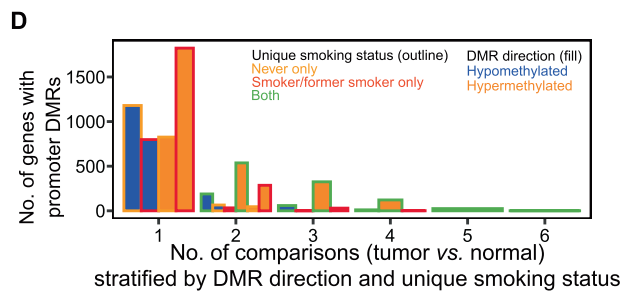
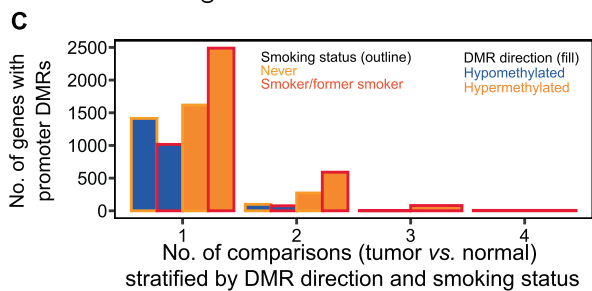
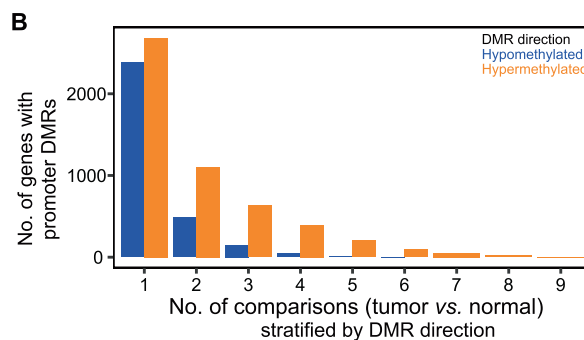
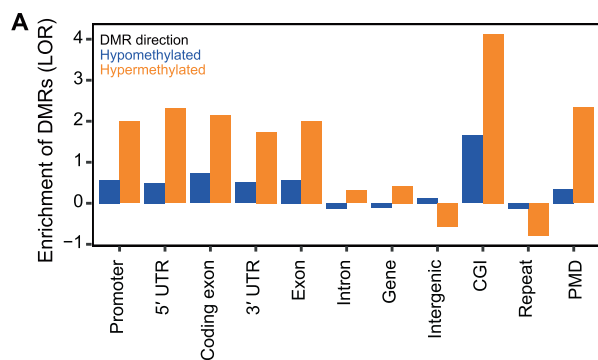
Both recurrent never-smoker-specific hyperDMRs ( $n = 2$  patients) and smoker-specific hyperDMRs ( $n \geq 3$  patients) were enriched in functions related to transcription and embryogenesis (Figure S12A and B), but smoker-specific

hyperDMRs were also enriched in terms related to cell signaling, epithelial tissue, and the respiratory system. Several genes, including *NOTCH1*, *STK11*, *AKT1*, *CDKN2B*, *DLX5*, *EYAI*, *GATA3*, and *SKI*, were represented by many of the terms. In contrast, recurrent never-smoker-specific hypoDMRs and smoker-specific hypoDMRs did not display significant GO enrichment.

Furthermore, 89% of CGIs overlapped with a hyperDMR, and 39% of CGIs overlapped with a hypoDMR, which were reduced to 78% and 15%, respectively, when only DMRs shared between patients were considered. In addition, 12% of genes ( $n = 5193$ ) contained a hyperDMR in the promoter, and 7% of genes ( $n = 3075$ ) contained a hypoDMR, with only 5% ( $n = 2171$ ) and 1% ( $n = 465$ ) of genes overlapping with shared DMRs, respectively. The results by gene biotype are presented in Table S2. Genes with a DMR in the promoter included 48 genes previously shown to be altered (amplified, mutated, methylated, fused, *etc.*) in NSCLC [27 recurrent (containing a DMR in more than one patient)], 188 genes from the Cancer Gene Census database of Catalogue Of Somatic Mutations In Cancer (COSMIC) [including 39 oncogenes, 25 tumor suppressor genes (TSGs), and 15 genes annotated as both] (104 recurrent), 164 epigenetic regulators from the EpiFactors database (78 recurrent), 43 cancer testis antigens (CTAs; 25 recurrent), and 29 of the top 100 most highly expressed genes in lung (16 recurrent).

Although most genes with a promoter methylation change were unique to an individual, genes were found to contain a hypoDMR in up to six patients and a hyperDMR in up to nine (Figure 3B). When considering smoking status, both smokers ( $n = 4$ ) and never-smokers ( $n = 2$ ) displayed recurrent hyperDMRs (679 and 273, respectively) and recurrent hypoDMRs (79 and 99, respectively) (Figure 3C). Interestingly, despite having only two never-smokers, the number of recurrent hypoDMRs was greater than the number recurrent in the four smokers, suggesting perhaps that the never-smoker hypoDMRs might be more relevant to tumor formation. Classifying genes with altered promoter methylation according to whether the alteration was seen in smokers, never-smokers, or both, revealed that the majority of recurrently altered genes appeared in both smokers and never-smokers (Figure 3D). However, 65/99 genes with recurrent hypoDMRs and 46/273 genes with recurrent hyperDMRs in never-smokers were not seen in smokers, while 34/79 genes with recurrent hypoDMRs and 318/679 genes with recurrent hyperDMRs in smokers were not in never-smokers. Genes with recurrent never-smoker-specific hypoDMRs in the promoter include *ASPSCR1*, a known fusion oncogene whose methylation is linked to prenatal smoking and reduced lung function [30]; the DNA topoisomerase *TOP2A*, a pan-cancer up-regulated gene [31,32]; and *DPP9* and *USP39*, which have been implicated in lung cancer [33,34]. Genes with recurrent never-smoker-specific hyperDMRs in the promoter include *SMYD2*, encoding a lysine methyltransferase that targets EML4-ALK fusion proteins [35], and *CYP1B1*. CYP1B1 converts benzo[*a*]pyrene from tobacco smoke into its more carcinogenic form and is up-regulated in smokers compared to never-smokers [36]. In many cases, the DMR was one of many in the gene promoter, although no transcript contained both a never-smoker-specific DMR and a smoker-specific DMR.

Genes associated with recurrent promoter hypomethylation, regardless of smoking status, include *FAM83A*, a lung



Brain pathway genes with a promoter hyperDMR (orange) in at least 5 patients

cancer biomarker and potential oncogene [37], whose expression in TCGA LUAD samples was almost 100-fold over that found in non-malignant lung samples (Wilcoxon test,  $P < 0.001$ ) (Figure 3E). The *FAM83A* promoter was hypomethylated in six patients, and several isoforms also lost methylation in TCGA (Wilcoxon test,  $P < 0.001$ ). *SEPT9* exhibited promoter hypomethylation in five patients and hypermethylation in three patients. Although overall gene expression increased significantly in the TCGA LUAD samples (Figure 3E), individual transcripts could be up-regulated or down-regulated (Figure S13A and B). Other genes containing recurrently hypomethylated promoters included *TNFRSF10A* (in five patients), encoding a TNF-related apoptosis-inducing ligand receptor that induces apoptosis [38], and *MET* (in three patients), a known oncogene implicated in NSCLC. In addition, *MUC5B*, which is involved in goblet cell mucus production, is down-regulated in mucinous adenocarcinoma [39], and was previously shown to be associated with LUAD cancer-specific DNA methylation changes [40]. In our study, *MUC5B* displayed promoter hypomethylation in four patients.

Genes with promoter hypermethylation in multiple patients included those within the *HOXA* cluster, which is densely hypermethylated in NSCLC (see below) [41,42]. Additionally, the promoter of *PCDH7*, a protocadherin gene with an oncogenic function in lung cancer [43], was hypermethylated in nine patients. Its expression level significantly increased in LUAD samples (Wilcoxon test,  $P < 0.001$ ) (Figure 3E), although like many genes, its expression had a much larger range than in normal lung samples. The promoter of *NKX2-1*, a lung cancer biomarker involved in lung development [44], was hypermethylated in five patients (Figure S14), as was that of *SOX17*. Several genes with recurrent promoter hypermethylation and/or nearby intergenic hyperDMRs were also previously shown to be associated with LUAD or LUSC cancer-specific hypermethylation changes, including *HYAL2* (in eight patients), *AQP1* (in seven patients), *XRCC3* (in five patients), *RARA* (in four patients), *SPTBN1* (in three patients), *EPAS1* (in three patients), *CD34* (in two patients), and *CLU* (in two patients). Other notable genes included *SYT10* and *KCNK1* (in eight patients each).

In many cases, the DMR was over the promoter of a shorter or non-coding isoform of the gene. However, we also identified potential instances of promoter switching, in which one transcript promoter of a gene became hypermethylated, while another became hypomethylated (Figure S15A–F). For both *ADCY2* and *ASPG*, the promoters of the longest protein-coding transcripts (*ADCY2*: ENST00000338316.4; *ASPG*: ENST00000551177.1 and ENST00000455920.2)

became hypermethylated in five patients, and the promoters of shorter isoforms became hypomethylated in a subset. The promoter methylation level of the longest transcripts also increased significantly in TCGA LUAD samples (Figure S15A and D), and the transcript expression levels dropped (Wilcoxon test,  $P < 0.001$ ) (Figure S15C and F). In both cases, the overall gene expression level dropped as well (Wilcoxon test,  $P < 0.001$ ) (Figure S15B and E); however, because the shorter isoforms were not captured by TCGA, we were unable to determine whether their expression increased.

In addition to individual genes with highly recurrent DMRs in the promoter, we looked at pathway-level effects using the Database for Annotation, Visualization and Integrated Discovery (DAVID). Genes with hypoDMRs in their promoters in multiple patients were enriched in spleen-specific expression ( $Q < 0.05$ ) (Figure S15G and H), while genes with hyperDMRs in their promoters were enriched for brain-specific expression (Figure 3F and G). Both sets were enriched on chromosome 19 and in alternative splicing functions. Those overlapping hyperDMRs were also enriched in several functions, including cell adhesion and terms related to homeobox genes and development, reflecting the abundance of hyperDMRs over *HOX* clusters.

We also determined whether DMRs overlapping with gene promoters encode binding motifs for transcription factors (TFs) dysregulated in NSCLC, which could provide a coordinated method for regulating several genes at once. Based on HOMER known motif analysis, shared promoter hypoDMRs were enriched in binding motifs for TFs involved in signaling and cell proliferation (Figure S15I). Additionally, 71% of the DMRs encode a binding site for NKX2-1 (up from the background observed value of 64%). In our dataset and in TCGA, promoters of the gene were recurrently hypermethylated, but its expression was not significantly different between normal lung samples and primary tumors. Promoter hyperDMRs were enriched in binding motifs for TFs involved in embryonic development, as well as hypoxia and angiogenesis (Figure S15J).

DMRs in intronic or intergenic regions may also serve as alternative promoters or enhancers that become activated in NSCLC. In contrast to promoters, they are also less likely to have been captured by TCGA or other large studies that relied on probe-based methylation profiling methods. In this study, we identified several recurrent intergenic DMRs, including chr13:53775000–53775500, a bivalent enhancer that was hypermethylated in 8 of the 10 patients and is 49 kb from the nearest gene; and chr6:158182500–158183000, which was hypomethylated in 6 of the 10 patients and is 61 kb from the nearest gene.

### Figure 3 DMRs within gene promoters

**A.** LOR enrichment of DMRs over genic features, intergenic regions, CGIs, and repeats compared to the background distribution of 500-bp bins containing CpGs, by DMR direction. **B.–D.** Number of patient-matched tumor vs. normal comparisons in which genes contain a DMR in their promoter(s), stratified by DMR direction (B), by DMR direction and smoking status (C), and by DMR direction and unique smoking status (classifying DMRs as identified in smokers, never-smokers, or both) (D). **E.** Expression of selected genes in TCGA LUAD samples (red) and matched normal lung samples (blue). A pseudocount of 0.00001 was added to each value. Lines represent median values. **F.** Significantly enriched gene sets among genes with a hyperDMR in the promoter in at least two patients, as determined by DAVID (top 20 terms by corrected  $P$ , out of 160 terms). Only terms with a Benjamini-corrected  $P < 0.05$  are included. Terms are labeled with the number of selected genes and ordered by corrected  $P$  within each category. **G.** Indication of DAVID Brain UP\_TISSUE Pathway genes with a hyperDMR in the promoter in at least five patients, stratified by patient. LOR, log odds ratio; PMD, partially methylated domain; DAVID, the Database for Annotation, Visualization, and Integrated Discovery.



There were also recurrent intronic DMRs, including the hyperDMR chr14:37136000–37136500, a bivalent enhancer within a *PAX9* intron (in 8 of the 10 patients), and the hypoDMR chr2:240169000–240169500 within an *HDAC4* intron (in 6 of the 10 patients).

Like hypoDMRs overlapping with promoters, shared intergenic DMRs were also enriched for AP-1 complex binding motifs, as well as other cell signaling and proliferation pathways (Figure S16A and B). Several of the TFs for which shared intergenic hypoDMRs have enriched binding motifs have altered expression levels in TCGA LUAD data, although half have lower expression in tumors, including *JUNB* and *ATF3* (Figure S16C).

Additionally, we determined whether intergenic-exclusive DMRs were enriched near genes with particular biological functions using GREAT. Shared hypoDMRs were significantly enriched near genes with the function “glucose import” ( $Q < 0.05$ , seven DMRs and seven genes). In contrast, shared hyperDMRs were enriched near genes with functions involved in morphogenesis and transcription (Figure S16D).

Finally, we identified small RNAs whose gene bodies overlapped with patient-matched DMRs. In total, 72 small RNA genes overlapped with a hypoDMR and 82 small RNA genes overlapped with a hyperDMR in at least one patient, although only 6 and 26 did so in multiple patients, respectively. *MIR663A* (chr20:26188821–26188914), which belongs to a processed transcript gene, was hypermethylated in six patients. The hypermethylated region overlapped with the promoter of the other isoforms and was annotated as a bivalent promoter; no other genes were in the vicinity. *MIR663A* is a known NSCLC tumor suppressor that acts through a variety of downstream targets, including TGF $\beta$ , p53, p21, and JunD [45–47], and its predicted targets include ACSL3, TGFB1, and HOXC10. *MIR487A* (chr14:101518782–101518862), which promotes tumor growth and metastasis in other cancers [48], was hypomethylated in four patients. For both, the expression level in LUAD samples was higher than that in normal lung samples (Wilcoxon test,  $P < 0.001$ ) (Figure S17). Additionally, the small nuclear RNA (snRNA) *RNVU1-8* (chr1:146551294–146551419), an alternative snRNA involved in mRNA processing [49], was hypermethylated in five patients, while the vault RNA *VTRNA1-2* (chr5:140098509–140098598), which may contribute to multidrug resistance in cancer cell lines [50], was hypermethylated in two patients.

### The majority of DMRs overlap with active chromatin states in tissues profiled by the Roadmap Epigenomics Project

Next, to better understand which regions may be susceptible to changes in DNA methylation in NSCLC, we determined the epigenetic state of each DMR in adult lung and other normal human tissues profiled by the Roadmap Epigenomics Project. For each of the 127 consolidated epigenomes, the Roadmap Epigenomics Project assigned a composite epigenetic state generated from five core histone modifications (15-state model) using ChromHMM. In addition, 98 epigenomes were also annotated with an 18-state model that includes H3K27ac.

We looked first at the epigenetic state of the DMRs in adult lung (epigenome E096). Compared to all potential regions, hypoDMRs were enriched in the enhancer states, ZNF/Rpts state, and heterochromatin state (Figure 4A). Although

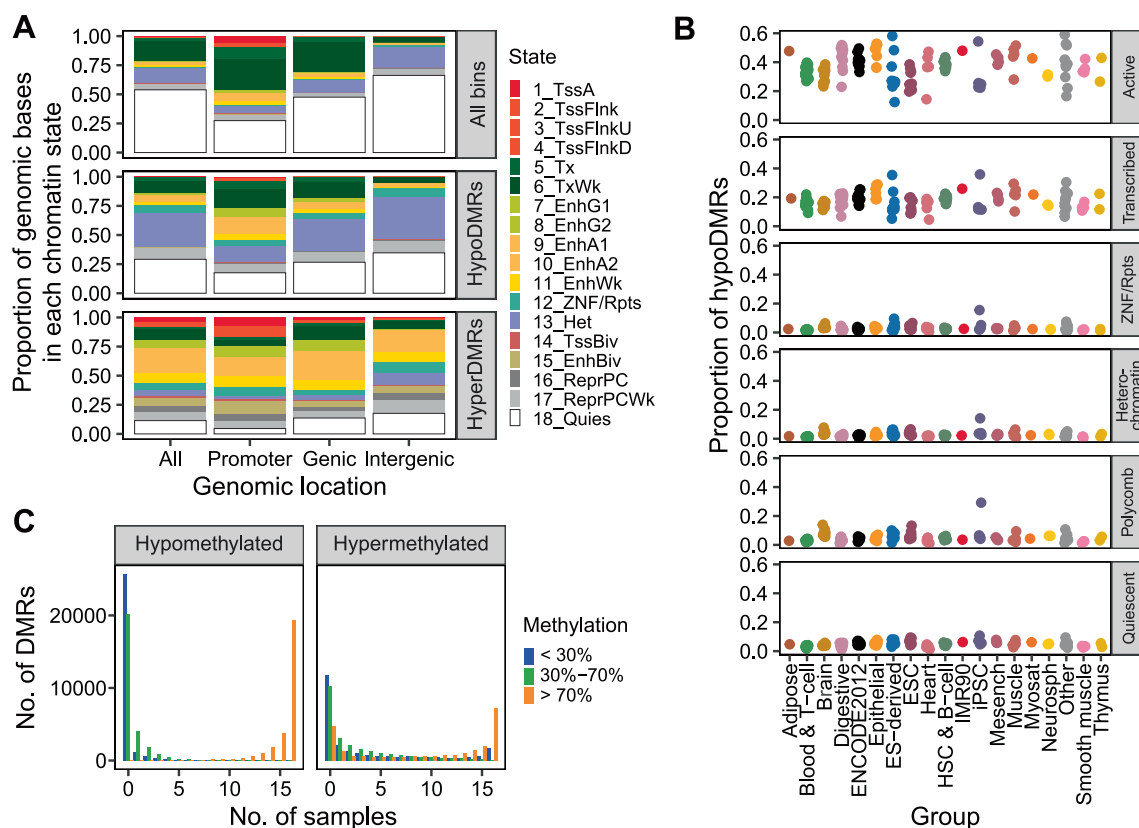
hypoDMRs were also enriched in polycomb-repressed states, hyperDMRs were enriched > 50-fold for the bivalent promoter and enhancer states (14\_TssBiv and 15\_EnhBiv) and > 25-fold for the polycomb-repressed state (16\_ReprPC), as well as active promoter and enhancer states. In contrast, both hypoDMRs and hyperDMRs were strongly depleted in the 18\_Quies quiescent state (29% and 11% of all DMR bases, respectively, vs. 54% of potential DMR bases), which represents an absence of histone modification chromatin immunoprecipitation sequencing signal. The enrichments held true even when DMRs were split by feature overlap and/or the number of patients in which they were found, although hyperDMRs were more likely to overlap with polycomb-repressed regions the more frequently they occurred (Figure S18A). Enrichment profiles were comparable for DMRs from smoker patients and those from never-smokers (Figure S19A). However, hyperDMRs exclusive to smokers demonstrated increasing enrichment over active enhancers the more they were shared across patients, in contrast to hyperDMRs observed in never-smokers (Figure S19B).

Overall, only 14% of hypoDMRs and 51% of hyperDMRs were in an active regulatory ChromHMM state in normal lung samples, although shared hypoDMRs were more likely to be in those states. Repressed genomic regions that became activated in NSCLC may have been poised to do so because they were regulatory elements in another tissue. Indeed, 65% of hypoDMRs and 94% of hyperDMRs were in an active regulatory state in a Roadmap sample besides adult lung (15-state model, excluding cancer cell lines), including > 50% of hypoDMRs in each repressed state in E096 (Figure S18B).

The proportion of DMRs in an active regulatory state in each sample varied by Roadmap sample group (Kruskal–Wallis test,  $P < 0.01$ , by DMR direction and composite ChromHMM state in E096). Interestingly, hypoDMRs in the heterochromatin, polycomb, or ZNF/Rpts state in adult lung were most likely to be in an active regulatory state in E022, an induced pluripotent stem cell (iPSC) sample that did not have a higher overall proportion of bases in those states (Figure 4B). They were also more likely to be in active states in brain, embryonic stem cells (ESCs), muscle, and other organs, and this was observed in hypoDMRs regardless of smoking status (Figure S19C). The three other normal lung samples profiled by the Roadmap Epigenomics Project — fetal lung (E088), normal human lung fibroblast primary cells (E128), and the IMR90 fetal lung fibroblast cell line (E017) — did not stand out in terms of the proportion of hypoDMRs in active regulatory states in those samples.

Additionally, we compared the epigenetic state of the DMRs in the A549 lung carcinoma cell line (E114) to the state in adult lung to determine whether changes in CpG methylation were matched by histone modification alterations. Both hypoDMRs and hyperDMRs were enriched relative to genomic background in active regulatory and polycomb-repressed states in A549 (Figure S18C). Regardless of epigenetic state in adult lung, hyperDMRs were far more likely to be in polycomb-repressed states in A549.

Finally, we found the average CpG methylation level of each DMR in 16 Roadmap samples profiled with methylCRF, none of which were lung samples. In general, regions that became hypomethylated in NSCLC were highly methylated in most or all other samples (see Materials and methods for complete listing of samples), suggesting that they lost regula-



**Figure 4** Epigenetic state of DMRs in Roadmap tissues

**A.** Proportion of Roadmap sample E096 (adult lung) genomic bases in each 18-state ChromHMM state, overall and overlapping hypoDMRs or hyperDMRs, split by exclusive feature overlap. Genic DMRs overlap with genes but not promoters, and intergenic DMRs do not overlap with genes or promoters. The overall states in E096 are restricted to regions overlapping with 500-bp bins that contain CpGs. 500-bp bins: promoter  $n = 453,670$ , genic  $n = 2,548,316$ , intergenic  $n = 2,267,290$ . HypoDMRs: promoter  $n = 4663$ , genic  $n = 13,730$ , intergenic  $n = 14,472$ . HyperDMRs: promoter  $n = 12,628$ , genic  $n = 14,670$ , intergenic  $n = 7128$ . **B.** The proportion of hypoDMRs in an active 15-state ChromHMM state in each Roadmap sample, by sample group (columns) and its 18-state ChromHMM state in E096 (rows) (see Materials and methods for composite state definitions). **C.** Distribution of the number of Roadmap samples in which each DMR has each average CpG methylation level, as assigned by methylCRF, split by DMR direction in the patient-matched NSCLC samples. ESC, embryonic stem cell; HSC, hematopoietic stem cell; iPSC, induced pluripotent stem cell.

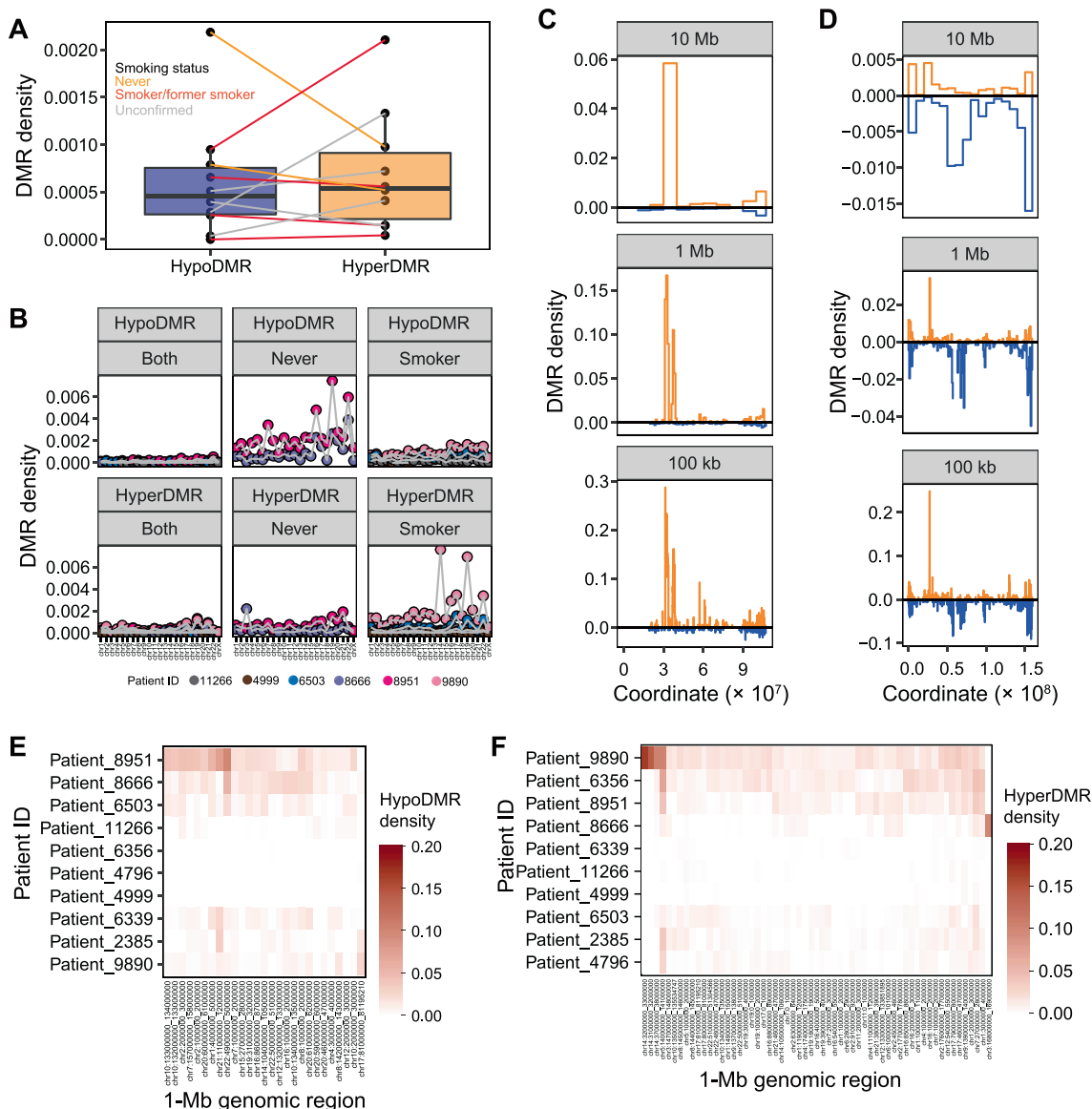
tion in NSCLC (Figure 4C). Interestingly, hypoDMRs were most likely to have lower methylation in epithelial samples (Figure S18D). In contrast, regions that became hypermethylated in NSCLC had a more variable profile and were hypomethylated or intermediately methylated in at least some other samples.

#### DMR density is non-random across the genome

As noted above, DMRs were not evenly distributed across the genome. Although the median DMR density per patient was 0.05% for both DMR directions ( $\sim 1$  DMR per 2000 bins) (Figure 5A), hyperDMRs exclusive to never-smokers were less dense than hypoDMRs, whereas the ratio of hyperDMR to hypoDMR density was relatively comparable for smoker-exclusive DMRs (Figure S20A). DMR density also varied heavily by chromosome for some patients. For example, chromosome 19 had unusually high DMR density for many patients, reflecting its high gene density and unique epigenetic profile (Figure S20B). However, when DMRs were separated according to smoking status (DMRs unique to smokers,

unique to never-smokers, or found in both), the hypoDMRs specific to never-smokers displayed high density across both never-smokers, particularly over chromosomes 16, 19, and 22 (Figure 5B). There were also instances of chromosomal outliers unique to individual patients, such as chromosome 14 for Patient\_9890 and chromosome 7 for Patient\_8951. These may represent instances of genomic rearrangement or loss of topologically associating domain (TAD) boundaries during NSCLC transformation that permitted chromatin spreading, although in both cases, the MRE-seq and MeDIP-seq read densities over the chromosomes did not suggest a large-scale copy number change relative to other tumors (Figure S1).

To better understand the distribution of DMRs, we profiled their density at three resolutions ranging from 100 kb to 10 Mb and identified windows with unusually high DMR densities (Figure 5C and D). For example, the 10-Mb window, chr14:30000000–40000000, had a hyperDMR density of 5.8% in Patient\_9890, representing 81% ( $n = 1075$ ) of the hyperDMRs on chromosome 14 but only 11% of potential DMR locations. Most of the DMRs were within two  $\sim 2$ -Mb-sized regions, with the density over the 1-Mb window, chr14:32000000–



**Figure 5 DMR density across patients, regions, and smoking status**

**A.** Genome-wide DMR density per patient (number of DMRs vs. number of 500-bp bins overlapping with CpGs), by DMR direction, with lines colored by smoking status of patient. **B.** DMR density per patient and chromosome, by DMR direction, for DMRs exclusive to smokers, exclusive to never-smokers, and identified in both. **C.** and **D.** DMR density (number of DMRs vs. 500-bp bins overlapping with CpGs) at three levels of resolution on chromosome 7 of Patient\_9890 (C) and on chromosome 7 of Patient\_8951 (D). **E.** and **F.** Mean hypoDMR (E) and hyperDMR (F) densities across 1-Mb windows in which the DMR density is > 1% in more than one patient or is one of the top 15 densest regions across samples, by patient.

33000000, reaching 16.7%. Because many patient-specific DMRs were within high-density regions — for instance, 13% of hyperDMRs specific to Patient\_9890 were within 1-Mb windows with a DMR density of > 10% (Figure S21) — most analyses focused on shared DMRs.

Although many of the high-density windows were specific to a single patient, others were recurrent. At a resolution of 1 Mb (approximately the length of a TAD), 82 windows had a DMR density of > 1% in multiple patients (Figure 5E and F, Figure S20C). This included chr5:140000000–141000000, which was densely hypermethylated in seven patients and is

centered over a protocadherin gene cluster; chr7:27000000–28000000 over the *HOXA* cluster (> 1% hyperDMR density, in six patients); and chr17:46000000–47000000 over the *HOXB* cluster (> 1% hyperDMR density, in four patients). In contrast, the only protein-coding gene within chr1:4000000–5000000 (> 1% hypoDMR density, in four patients) was *AJAPI*.

To further explore which regions of the genome may be predisposed to large-scale methylation alterations in NSCLC, we compared DMR density to features such as gene, CpG, and repeat density, and the epigenetic state in normal lung. Many

of these features were highly correlated (Figure S22A), with the first several PCs separating (1) the proportion of the window in the quiescent state (18\_Quies) in normal lung (E096) from gene density and active states, (2) active states from repressed states, and (3) polycomb-repressed states from the heterochromatin and ZNF/Rpts states and repeat density (Figure S22B–E; File S1).

For both hypoDMRs and hyperDMRs, DMR density in individual windows, the mean density across patients, and the number of patients in which the DMR density was > 1% were negatively correlated with the proportion of the windows in the quiescent state (Figure S22F and G). HypoDMR density was most highly correlated with the heterochromatin, ZNF/Rpts, and polycomb-repressed states as well as CpG density. HyperDMR density was most highly correlated with gene and CpG densities and the polycomb-repressed states, as well as active genic and enhancer states.

Interestingly, DMR density in both directions was also anticorrelated with the starting methylation level in the normal lung sample from that patient, as estimated by methylCRF. If changes in CpG methylation were stochastic, the starting methylation level should dictate the direction of methylation change; however, a higher normal lung methylation level led to higher DMR density in both directions, likely because it is linked to gene and CpG densities. Together, these results suggest that (1) changes in methylation level during primary NSCLC transformation were enriched in regions with higher gene and CpG densities, (2) hypoDMRs were skewed toward heterochromatin-repressed regions, (3) hyperDMRs were enriched over active regions, and (4) both were enriched over polycomb-repressed regions.

#### Many hyperDMRs are exclusive to adenocarcinomas and enriched for nervous system development processes

Next, we identified DMRs associated with specific patient and/or tumor characteristics.

Although *TP53* mutations were previously shown to be associated with lung cancer grade [4], few patient-matched DMRs were exclusive to low-stage (1/1A) or high-stage (3A/3B) tumors (10 and 9 DMRs, respectively; see Materials and methods). However, 99 hypoDMRs and 18 hyperDMRs were exclusive to acinar cell adenocarcinoma ( $n = 2$  patients), and 37 hypoDMRs and 389 hyperDMRs were exclusive to adenocarcinoma ( $n = 2$  patients). For categories including Patient\_4999, DMRs were counted as exclusive if they were present in all other members of that category. Although many of the exclusive DMRs may be passenger events, the number of exclusive DMRs for adenocarcinoma tumors was on the high end of a distribution generated with shuffled patient clinicopathologic data, suggesting that many of them are truly biology-specific (Figure 6A).

The adenocarcinoma-specific hyperDMRs were significantly enriched for several GO biological processes related to nervous system development by GREAT (Figure 6B). A small set of genes were included in many of the terms, including *NKX6-2*, *PAX6*, *LHX3*, *SOX1*, *ISL1*, and *HOXD10*. 44% ( $n = 186$ ) of the adenocarcinoma-specific DMRs overlapped with promoters. These included *ZSCAN30* and *ZIK1*, whose promoters became hypermethylated only in adenocarcinoma.

#### The majority of DMRs overlap with repeats, demonstrating enrichment over classes and subfamilies by direction of methylation change

Finally, we looked at methylation changes over repeats during NSCLC transformation. This is a relatively understudied aspect of cancer biology, especially considering that repeats comprise half of the human genome and can serve as alternative promoters and enhancers. Utilizing MeDIP-seq and MRE-seq data allowed us to better explore this component of the genome, as 53% of the CpGs profiled by methylCRF overlapped with a repeat, compared to 16% of those profiled by TCGA.

As noted above (Figure S11A), 46% of hyperDMRs ( $n = 11,239$ ) and 73% of hypoDMRs ( $n = 20,685$ ) identified in this study overlapped with a repeat. Across the 10 patients, this corresponded to 50,699 repeats that overlapped with a patient-matched DMR, including 35,414 overlapping with hypoDMRs and 15,652 overlapping with hyperDMRs. In addition, 12% of the repeats overlapped with a hypoDMR in multiple patients (maximum 6) and 22% of the repeats overlapped with a hyperDMR in multiple patients (maximum 8) (Figure 7A). This included a DMR (chr10:67554500–67555000) far from any gene that overlapped with adjacent L1M5, LTR24C, and AluY elements and was hypomethylated in five patients.

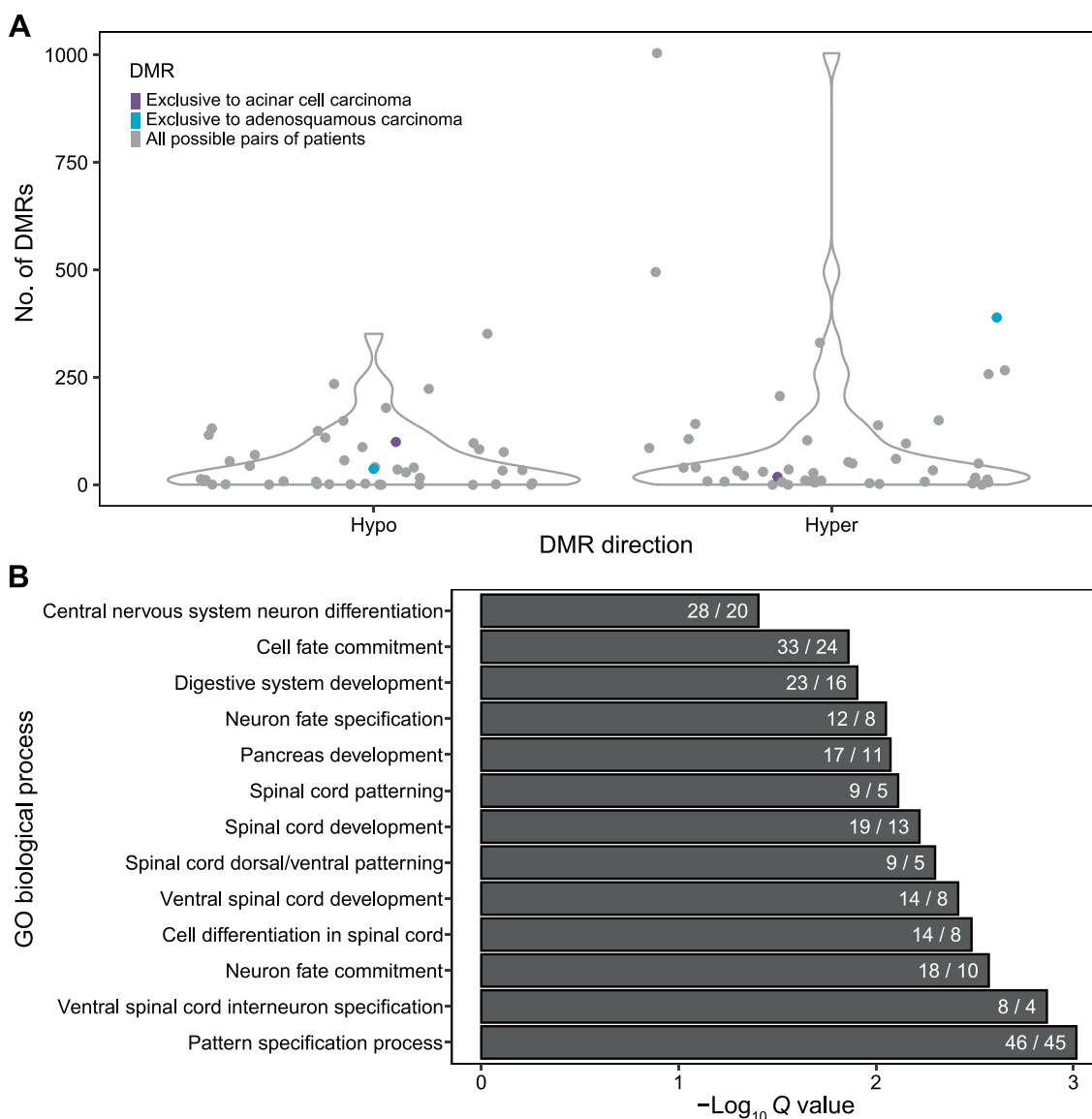
In a single patient, the number of repeats overlapping with DMRs ranged from 8 to 15,213 for hypoDMRs (mean 4040) and from 109 to 6978 for hyperDMRs (mean 2094) (Figure 7B), reflecting the landscape of CpG methylation changes in those patients.

Repeats were divided into several classes of retrotransposons, DNA transposons, and other repetitive elements. When compared to the number of repeats per class, short interspersed nuclear element (SINE) and long terminal repeat (LTR) elements were enriched among those overlapping with shared hypoDMRs (Figure 7C). In contrast, low complexity regions and simple repeats were enriched among those overlapping with hyperDMRs, and DNA transposons and long interspersed elements (LINEs) were depleted from both sets. These patterns held true when all patient-matched DMRs were considered (Figure S23A).

Next, we looked at CpG methylation changes over repeat subfamilies during NSCLC transformation. Several of the youngest TE subfamilies had the highest overall methylation in normal lung samples, including SVA and AluY subfamilies and LTR12C (Pearson correlation,  $P > 0.05$ , median Jukes–Cantor distance vs. median methylation level) (Figure S23B). Additionally, TE subfamilies with higher methylation in normal lung exhibited less variation in methylation level between normal samples (Pearson correlation coefficient =  $-0.53$ ,  $P < 0.001$ ) (Figure S23C), reflecting the importance of DNA methylation as a repressive tool. Younger TE subfamilies lost more CpG methylation on average in NSCLC (Pearson correlation coefficient =  $-0.30$ ,  $P < 0.001$ ) (Figure S23D), although interestingly, this was not primarily due to their higher starting methylation level, as there was little correlation between the median methylation level in normal samples and the change between normal lung samples and tumors (Pearson correlation coefficient =  $0.12$ ,  $P < 0.001$ ).

As expected, the variation in repeat subfamily methylation level across samples increased in tumors (Figure S23E), although it was higher for those with greater variation in nor-





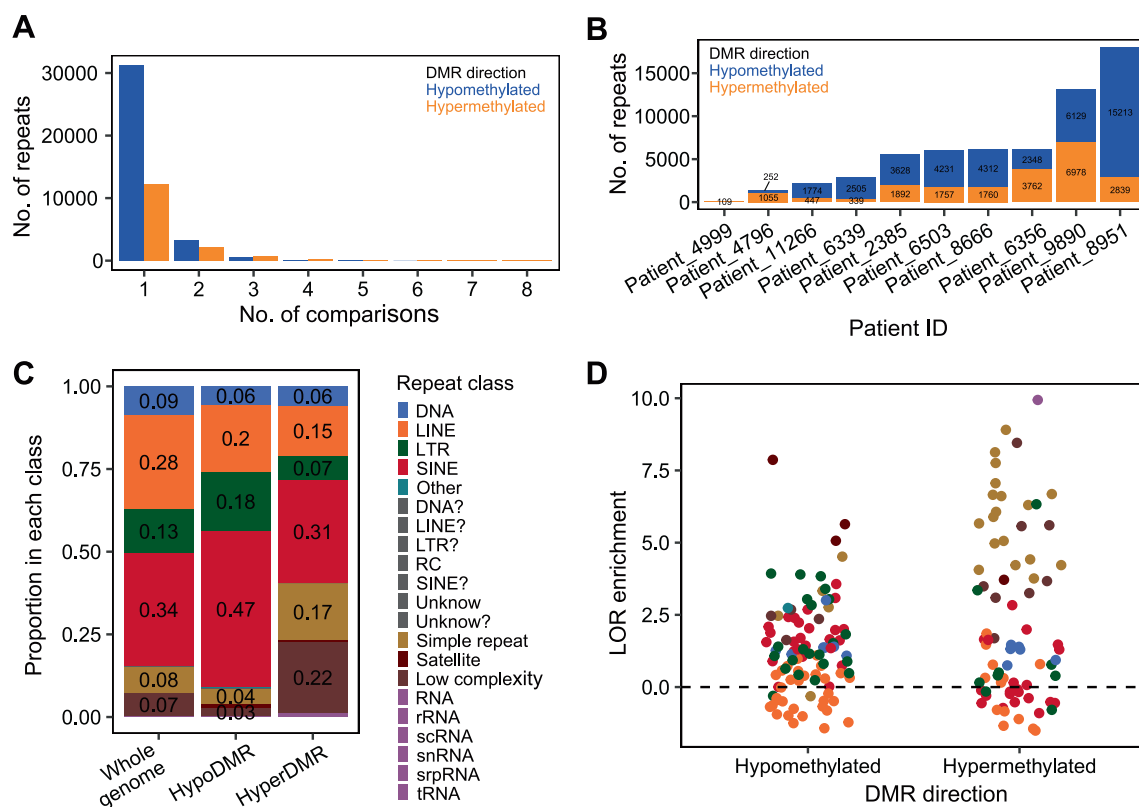
**Figure 6 Tumor histological subtype-exclusive DMRs**

**A.** Number of DMRs exclusive to acinar cell carcinoma (purple) or adenosquamous carcinoma (turquoise) using true tumor subtype or all possible pairs of patients (gray), by DMR direction. **B.** Top 20 significant GO biological processes for hyperDMRs specific to adenosquamous carcinoma, as identified by GREAT (see Materials and methods). Terms are ordered by FDR-corrected binomial  $Q$  and are labeled by the number of DMRs over the number of genes involved. FDR, false discovery rate.

mal lung samples (Pearson correlation coefficient = 0.75,  $P < 0.001$ ) (Figure S23F). Interestingly, many of the subfamilies with the largest median change in methylation between normal lung and tumor samples exhibited a bimodal distribution in tumors, with the same subfamily exhibiting increased methylation in some tumors and decreased methylation in others (Figure S23G).

In addition, we calculated the enrichment of DMRs over repeat subfamilies compared to their length to determine which were most prone to concentrated DNA methylation alterations. Most repeat subfamilies overlapped with DMRs: 74% of all subfamilies and 85% of TE subfamilies overlapped with hypoDMRs (1032 of 1397 and 824 of 968, respectively), and 63% and 69% overlapped with hyperDMRs (877 of 1397 and 665 of 968, respectively).

When considering all DMRs, the most highly enriched subfamilies in both directions were transfer RNAs (tRNAs), low complexity regions, satellite repeats, and simple repeats (Figure S24A). However, when restricted to DMRs shared between patients, TE subfamilies were better represented among those overlapping with hypoDMRs, particularly ERV1 subfamilies (Figure 7D). Highly enriched subfamilies included LTR1, LTR1D, LTR12C, LTR17, MER45A, MER50, MER52A, MER52D, AluY, AluYc, AluYk4, and SVA\_F. 11% of LTR12C elements ( $n = 278$ ) overlapped with a hypoDMR in at least one patient, and 53 LTR12C elements were hypomethylated in multiple patients [Log odds ratio (LOR) = 3.04,  $n = 45$  DMRs; maximum 3 patients,  $n = 8$ ]. This included the smoker-specific DMR in the *ARRB2* promoter mentioned above. Similarly, 10% of MER52D elements



**Figure 7 Repeat overlap with patient-matched DMRs**

**A.** Number of comparisons in which each repeat overlaps with a DMR, by DMR direction. **B.** Number of repeats overlapping with DMRs per patient. Bars are labeled with the number of repeats. Counts less than 100 are not shown. **C.** Proportion of repeats in each class across the genome and overlapping with hypoDMRs or hyperDMRs shared between patients. Proportions less than 0.02 are not shown. **D.** LOR enrichment of shared DMRs over repeat subfamilies, colored by repeat class and stratified by DMR direction. Only subfamilies overlapping with more than 5 DMRs across all patients are shown. Dashed line represents no enrichment or depletion. LINE, long interspersed element; LTR, long terminal repeat; SINE, short interspersed nuclear element; RC, rolling circle; rRNA, ribosomal RNA; scRNA, small cytoplasmic RNA; snRNA, small nuclear RNA; srpRNA, signal recognition particle RNA; tRNA, transfer RNA.

( $n = 57$ ) and 9% of MER52A elements ( $n = 149$ ) overlapped with a hypoDMR in at least one patient.

In contrast, the LTR subfamily MER57E3 was enriched in shared hyperDMRs (LOR = 6.33,  $n = 7$  DMRs). Of the eight MER57E3 elements that overlapped with DMRs in multiple comparisons, most were within the first intron of a ZNF, pseudo-, or non-coding gene; they were also often adjacent to a MER21C element. MER57E3 had the lowest methylation level of any TE subfamily in normal lung samples (median 49.8% methylation), likely due to its overlap with promoter elements [51].

There was a moderate correlation between the CpG density of TE subfamilies and their enrichment for hypoDMRs (Spearman's  $\rho = 0.47$ ,  $P < 0.001$ ), which was less true for hyperDMRs (Spearman's  $\rho = 0.17$ ,  $P < 0.001$ ) (Figure S24B). However, although several of the most enriched subfamilies were relatively young, there was little correlation between TE subfamily age and DMR enrichment (Spearman's  $\rho < 0.2$  in both directions). There was also little correlation between DMR enrichment and overall changes in CpG methylation level across the subfamily (Pearson correlation coefficient  $< 0.25$  for hyperDMRs,  $P > 0.1$  for hypoDMRs).

A recent study used TCGA data to identify TEs that spliced into oncogenes to create alternative isoforms in cancer [15].

Four of the TEs identified in that study overlapped with patient-matched hypoDMRs here, although only in one patient each. They included an intergenic AluSp element (chr19:15439585–15439882) that spliced into *BRD4*, as well as an intronic AluSp element (chr20:490814–491118) that formed an alternative *CSNK2A1* transcript in hundreds of TCGA LUAD and LUSC samples.

## Discussion

In this study, we used two complementary, genome-wide DNA methylation profiling techniques to interrogate DNA methylation changes in primary NSCLC and to detect local DMRs in comparison to patient-matched non-malignant lung tissue. Methylation changes in primary NSCLC were highly heterogeneous. As expected, tumors lost methylation over intergenic regions and repeats and gained methylation over promoters. *HOX* and protocadherin clusters, as well as several lung biomarkers and oncogenes, were recurrently targeted by promoter methylation changes, as were small RNAs. However, we also identified recurrent, intergenic DMRs, which may represent distal regulatory elements affected in NSCLC. These regions could potentially serve as biomarkers in cell-free DNA assays, for example, for early lung cancer detection.

Many of the genes identified in [20] as frequently methylated in the LUAD CGI methylator phenotype-high cluster, which were enriched for WNT pathway genes, were also hypermethylated in our study. As noted, the promoters of *HOXA9*, *SOX17*, and *GATA2* overlapped with hyperDMRs in seven, five, and four patients in our study, respectively. *CDKN2A*, which is frequently methylated in LUAD and is inactivated by epigenetic silencing in 21% of LUSC [19], overlapped with hyperDMRs in three patients. *STK11*, a frequently mutated tumor suppressor gene, overlapped with hyperDMRs in five patients, but only over the promoter of a shorter isoform. In contrast, the most common driver genes in NSCLC, *EGFR*, *KRAS*, and *TP53*, rarely or never overlapped with DMRs. Interestingly, one of the most commonly methylated genes in this study, *SEPT9*, is the basis for an FDA-approved, methylation-based colorectal cancer detection test (Epi proColon) and is being explored as a biomarker for lung cancer [52,53]. Promoter methylation of *SOX17* also showed promise as a lung cancer biomarker [54].

Pathways related to cell signaling, cell cycle and proliferation, adhesion, chromatin and RNA splicing factors, and the oxidative stress response are commonly altered in LUAD, with 76% of tumors exhibiting RTK/RAS/RAF pathway activation [4,5,20]. In our study, similar pathways were enriched among genes with DMRs in their promoters. In line with previous studies, we also demonstrated recurrent gain of methylation over regions marked with polycomb repressive marks, which were hypomethylated in most tissues [9]. Previous research has also indicated that enhancers lost in cancer targeted cell fate-specifying genes and were lineage-specific, while those activated in cancer were more universal and targeted growth-related genes [24,55]. Here, recurrent hypoDMRs, including those in intergenic regions, were enriched for binding motifs for the AP-1 complex, as well as the lung development TF NKX2-1, which is both lineage-specific and related to development.

Additionally, we identified DMRs exclusive to both tumor subtype and smoking status. Although underpowered to reach significance due to small sample sizes, we did observe that smoker patients generally exhibited fewer DMRs than those seen in never-smokers, with the exception of Patient\_9890. In addition, never-smokers tended to have a higher percentage of hypoDMRs than smokers, and more frequently displayed recurrent promoter hypomethylation despite fewer samples ( $n = 2$  never-smokers,  $n = 4$  smokers), suggesting that loss of methylation may promote cancer development in never-smokers. In contrast, smokers more frequently displayed recurrent promoter hyperDMRs than never-smokers. These observations are consistent with the hypothesis that smoking-induced DNA damage recruits DNMT1 and leads to local increased methylation at repaired sites [56], although additional studies are needed to validate these trends due to small sample sizes.

Alexandrov et al. [7] identified only 434 differently methylated CpGs based on smoking status, most of which were in genes with no known cancer function. Similarly, although Freeman et al. [57] identified 263 CpGs significantly associated with smoking in the TCGA LUAD and LUSC methylation datasets, only five replicated in an independent dataset. Few genes identified in earlier studies as preferentially mutated or silenced via promoter methylation based on smoking status overlapped with DMRs in our study, including *p16*, *APC*,

and *MLH1/2* [6]; *MICAL3* [5]; *CST6*, *EMILIN2*, *LAYN*, and *MARVELD3* [58]; and *RASSF1*, *MGMT*, *RARB*, *DAPK*, *WIFI1*, and *FHIT* [59]. *CPEB1*, which was identified as specific to smokers in [58], and *CDKN2A*, identified in a meta-analysis of 97 studies by [59], overlapped with hyperDMRs here, although none were smoking status-specific. However, we identified DMRs overlapping with additional genes known to be involved in NSCLC or smoking, such as *CYP1B1*.

Finally, we identified repeat subfamilies and individual repeats that were frequently affected by DNA methylation changes in NSCLC. Promoters overlapping with repeats are often up-regulated in cancer [32]. In our study, we observed a strong enrichment of hypoDMRs over ERV1 subfamilies. Although all major TE classes contain a sense-directional promoter [60], L1 elements are frequently 5' truncated, while LTR repeats retain their regulatory elements after recombination removes the intervening viral genome [13], potentially explaining this phenomenon. We also observed hypomethylation over several young subfamilies, including SVA\_F and several AluY subfamilies. Finally, we identified examples of repeats that have been shown to splice into known oncogenes in TCGA cancer samples to form alternative transcripts [15].

Here, we expand our knowledge of the epigenetic changes that occur during primary NSCLC transformation across LUAD subtypes and smoking status, potentially informing treatment strategies such as epigenetic therapy and immunotherapy. Future studies aimed at whole methylome profiling of patient-matched normal and lung tumor samples of smokers and never-smokers should be conducted to further assess the frequency of DMR recurrence and specificity of methylation alterations for specific cohorts of patients.

## Materials and methods

### Samples and clinical data

Snap-frozen primary NSCLC tumor specimens ( $n = 17$ ) and patient-matched non-malignant tissue ( $n = 10$ ) along with de-identified pathology and demographic data were obtained from the Siteman Cancer Center Tissue Procurement Core. All specimens were previously collected from surgical remnants with patient consent under a protocol approved by the Institutional Review Board (IRB) (Approval No. 201305031). All frozen tissue specimens were sectioned and stained with hematoxylin and eosin to confirm histopathology and estimate tumor cell purity in each malignant tissue specimen. Serial frozen tissue sections were used for genomic DNA isolation using spin-column based purification (Catalog No. 56304, Qiagen, Germantown, MD). Quantification of all DNA samples was performed by nanodrop spectrophotometry prior to methylation analyses using a GE NanoVue Portable Spectrophotometer (Catalog No. 80-2140-21, GE HealthCare, Chicago, IL).

Patient demographics and pathology data, including smoking status, pathologic diagnosis, stage, and tumor purity, were provided by the Tissue Procurement Core. Smoking status and patient sex were confirmed via chart review by an "honest data broker". For analyses between never-smokers and smokers, former smokers were included in the latter category. Samples listed as "Acinar adenocarcinoma", "Acinar cell carcinoma", and "Acinic cell adenocarcinoma" were combined into a single

category, “Acinar cell carcinoma”, for all analyses. Specimens for which demographic or pathology data could not be obtained were excluded from the relevant statistical tests (Table S1).

## Features

GENCODE (v19) comprehensive gene features and CGIs (cpgIslandExtUnmasked) were downloaded from the University of California Santa Cruz (UCSC) Table Browser. Promoters were defined as 2000 bp upstream and 500 bp downstream of GENCODE transcription start sites. A lookup table associating each Ensembl transcript ID with gene ID, name, and biotype was downloaded from the UCSC Table Browser. Chromosomes and promoters overlapping with CGIs were added to each transcript. The National Center of Biotechnology Information (NCBI) gene alias table (Homo\_sapiens\_gene\_info.gz) was downloaded from NCBI. The hg19 RepeatMasker rmsk file was used to identify repeats.

## MRE-seq and MeDIP-seq

MRE-seq and MeDIP-seq were performed on all samples using the procedures described in [61]. MRE-seq was performed with four enzymes: *HpaII*, *SsiI*, *Hin6I*, and *HpyCH4IV* (Catalog Nos. R0171S, R0551S, R0124S, and R0619S, New England Biolabs, Ipswich, MA). A base offset of 3 was used during sequencing.

Adapter trimming, read alignment to the hg19 genome with BWA, and methylQA post-processing were also performed as described in [61]. methylQA produced a bed file of unique alignments (mapping quality  $\geq 10$ ), virtually extended, for each MeDIP-seq alignment file. For MRE-seq, it removed reads that did not map to MRE cut sites. Chromosomes 1–22, X, Y, and M were included.

MRE-seq and MeDIP-seq read counts per 500-bp bin were generated using M&M. MeDIP-seq read counts were normalized to reads per kilobase per million mapped reads (RPKM) using the total number of input reads. For MRE-seq, bins were restricted to those containing an MRE cut site, and read counts were normalized to RPKM using the total number of reads mapped.

PCA was performed on all samples using the MeDIP-seq and MRE-seq RPKM per 500-bp bin as features with the *prcomp()* function. Matrices were scaled and centered, and only bins with variation across samples were included. The variance explained by each PC was calculated from the standard deviation.

## methylCRF

methylCRF was performed on the 27 primary tumor and normal lung samples to estimate the methylation level at each CpG using MRE-seq and MeDIP-seq data. The methylCRF package was downloaded and installed from <http://methylcrf.wustl.edu/>. methylCRF was run using Perl (v5.14.4), R (v3.3.0), and SAM tools (v1.3.1).

MRE normalization was performed on aligned MRE-seq reads as recommended, using a quality score of 10 and a base offset of 3 for *sam2bed.pl* and the 4-enzyme MRE fragment file as input for *MRE\_norm.pl*.

methylCRF was performed on normalized MRE-seq read counts and MeDIP-seq extended alignments in bed format using the HIES model-specific files, the hg19 genome-specific files, the 4-enzyme virtual digest file, and a gap size of 750 bp. Methylation estimates were assigned for 28,085,255 CpGs across chromosomes 1–22, X, Y, and M (chromosome M always misses methylation value). Chromosomes 1–22, X, and Y were used for all methylCRF analyses.

## Methylation distribution in 1-kb windows

Non-overlapping 1-kb windows were generated using *bedtools makewindows* across chromosomes 1–22, X, Y, and M. Windows were intersected with CpG methylation level estimates generated by methylCRF using *bedtools intersect*, and the mean methylation over all CpGs within each window was calculated for all samples. Windows that did not overlap with a CpG were excluded from analyses.

PCA was performed on genomic windows with variation between samples using the *prcomp()* function. Matrices were scaled and centered. The variance explained by each PC was calculated from the standard deviation. Group centroids were identified by calculating the mean along each PC by malignant status. The distance to the group centroid for normal and tumor samples was calculated using Euclidean distance over all PCs.

A Pearson correlation distance matrix between all samples across the 1-kb windows was generated.

## Average methylation over genomic features

The average methylation level over each feature was calculated by first internally collapsing features to reduce overlapping bases using *bedtools merge*, and then intersecting the merged features with the CpG methylation level estimates generated by methylCRF using *bedtools intersect*. Then, the mean methylation level over all CpGs overlapping with the feature was calculated for all samples.

## MRE cut site saturation

The number of sampled restriction enzyme cut sites and the number of cut site-filtered MRE-seq reads per sample were obtained from the methylQA reports.

Cut sites specific to normal and tumor samples were identified by comparing the CpGs included in the MRE-seq CpG bedGraph files output by methylQA. The number of normal and tumor samples in which each CpG was sampled was counted. Cut sites overlapping with each genomic feature were identified using *bedtools intersect*, and the distribution was compared to that of all possible restriction enzyme sites in the genome ( $n = 10,214,062$ ).

## Comparison to previously published methylCRF data

CpG methylation level estimates generated by methylCRF for 16 samples profiled in [62] [Gene Expression Omnibus (GEO): GSE86505] were downloaded from [https://wangftp.wustl.edu/~jli/final\\_hub/methylCRF\\_score](https://wangftp.wustl.edu/~jli/final_hub/methylCRF_score), and cancer types were extracted from sample names.



methylCRF was performed on three samples profiled by the Roadmap Epigenomics Project using MeDIP-seq and MRE-seq (GEO: GSM613862, GSM613842, GSM669614, GSM669604, GSM707021, and GSM707017). methylQA was performed on the MeDIP-seq read alignments as recommended (<http://methylqa.sourceforge.net/tutorial.php>) after adding the “chr” prefix to each chromosome in the header with *samtools reheader*. MRE normalization was performed using the 3-enzyme fragment file with a base offset of 3, and methylCRF was performed using the 3-enzyme virtual digest file. The average methylation in 1-kb windows across the genome was calculated as above, excluding windows that did not overlap with a CpG.

To compare to the samples profiled in this study, which were processed with four restriction enzymes, methylCRF was performed on two samples (P4796\_N\_UC and P4796\_T\_UC) considering only three enzymes (excluding *Hpy-CH4IV*). MRE normalization was repeated using the 3-enzyme fragment file, and methylCRF was performed using the 3-enzyme virtual digest file. The average methylation in 1-kb windows across the genome was calculated as above, excluding windows that did not overlap with a CpG. PCA was performed as above on all primary NSCLC samples, the three Roadmap samples, and the two re-processed primary NSCLC samples simultaneously.

## M&M

M&M (R package methylMnM) was used to identify DMRs between samples using MRE-seq and MeDIP-seq data. First, the functions *countcpgbin* and *countMREcpgbin* were used to count the number of CpGs and restriction enzyme cut sites in 500-bp bins across the genome (chromosomes 1–22, X, Y, and M;  $n = 6,166,049$  bins). Cut sites for the four enzymes used in MRE-seq were considered, and a blacklist file was provided (hg19 DAC Blacklisted Regions, e.g., *wgEncodeDacMapabilityConsensusExcludable* from the UCSC Genome Browser).

Then, 500-bp DMRs were identified using the read alignment bed files output by methylQA for MRE-seq and MeDIP-seq from two samples. The functions *countMeDIPbin* and *countMREbin* were used to produce normalized read counts per bin for each sample. Default values were used for *MnM.test*, *MnM.qvalue*, and *MnM.selectDMR*, and no polymerase chain reaction threshold was set.

DMRs were identified between patient-matched normal lung and tumor samples from the same patient ( $n = 10$ ), between all pairs of normal samples ( $n = 45$ ), and between all normal–tumor sample pairs from different patients ( $n = 170$ ).

### DMR $Q$ value threshold

The DMR  $Q$  value threshold was selected from among four potential thresholds ( $1E-2$  to  $1E-5$ ). For each threshold, the false positive ratio was calculated for patient-matched normal vs. tumor comparisons as the mean number of DMRs identified between the given normal and all other normal samples divided by the number of patient-matched DMRs. The proportion of DMRs shared between patient-matched comparisons was also calculated for each threshold.

Additionally, all DMRs between patient-matched normal and tumor samples were intersected with the CpG methylation levels generated by methylCRF for each sample in the pair, and the mean methylation level across all CpGs within each DMR was calculated for each sample. The change in mean methylation level between patient-matched normal and tumor samples was calculated for each DMR identified in that pair. The proportion of DMRs with  $> 10\%$  methylation change in either direction was calculated for each  $Q$  value threshold.

DMRs with a  $Q$  value  $< 0.001$  from all comparisons were combined and assigned a unique ID by position. First, adjacent DMRs were merged and assigned a number based on their location in the genome. Then, individual 500-bp DMRs were assigned a second number representing their position within the merged block. IDs were consistent across sample comparisons. There were 1,817,809 instances of 245,723 unique DMRs across all comparisons.

Because all patients with patient-matched normal and tumor samples were female, DMRs on chromosome Y ( $n = 1112$  instances of 80 unique DMRs, 0.03% of all DMRs) were excluded from downstream analyses.

### Feature overlap

To determine the genomic context of DMRs identified in this study, all unique DMRs were intersected with GENCODE genic features, intergenic regions, CGIs, and repeats using *bedtools intersect*. As background, genome-wide 500-bp bins (M&M input) were also intersected with the features. Only bins on chromosomes 1–22 and X that overlapped with a CpG (e.g., potential DMRs) and did not overlap with a blacklisted CpG were considered ( $n = 5,269,276$ ). The proportion of DMRs and 500-bp bins overlapping with each feature category was calculated, and the LOR enrichment of DMR overlap vs. bin overlap was calculated for each feature.

Intersections of unique DMRs with individual CGIs were identified by CGI ID ( $n = 524$ , chromosomes 1–22 and X). The proportion of unique CGIs overlapping with a DMR (any or shared between multiple patient-matched comparisons) was calculated.

The number of GENCODE genes whose promoter(s) overlapped with a patient-matched DMR (any or shared by multiple patients) was calculated and compared to the number of genes whose promoter(s) overlapped with a 500-bp bin containing a CpG (e.g., potential DMRs) by gene biotype. Greater than 99.9% of transcripts and genes contained a potential DMR in their promoter(s) (chromosomes 1–22 and X), including  $> 99\%$  of all gene and transcript biotypes except mitochondrial ribosomal RNA (rRNA) and tRNA. Unique transcripts and genes were identified by IDs, not gene names.

In addition, exclusive feature overlap categories were assigned to each DMR: promoter (DMRs overlapping with a promoter), genic (DMRs overlapping with a gene but not a promoter), and intergenic (intergenic DMRs overlapping with neither a promoter nor a gene).

For each GENCODE transcript and gene, the number of unique patient-matched DMRs overlapping with the promoter(s), the number of patient comparisons in which the promoter(s) overlapped with a DMR, and the total number of DMR overlaps across comparisons were calculated by direction of methylation change.

For each DMR, the distance to the closest gene on either strand was found using *bedtools closest*, reporting all ties.

### Gene sets

Several gene lists were used to identify genes previously recognized as involved in cancer. A list of genes previously identified as altered in LUAD and LUSC was compiled from nine publications [4–6,8,19,20,58,59,63]. The Cancer Gene Census (CGC) database (v79) was downloaded from COSMIC. The EpiFactors database (v1.7.3) was downloaded from the FANTOM5 consortium. CTAs (gene names plus alternative names) were obtained from the CTdatabase website (<http://www.cta.lncc.br/>; accessed 2/13/2019). Additionally, genes listed as “cancer/testis antigen” or “cancer testis antigen” in the NCBI gene alias table were also included. The top 100 most highly expressed genes in lung were obtained from the Genotype-Tissue Expression Project (GTEx) data portal, excluding mitochondrial genes (<https://gtexportal.org/home/eqtls/tissue?tissueName=Lung>).

Genes included in sets of interest were identified by gene names and linked to gene IDs using the GENCODE lookup table. For gene names not included in the GENCODE table, where possible, the name was identified among synonyms in the NCBI gene alias table and linked to the canonical name. For GTEx genes, the gene ID (without version suffix) provided by GTEx was used to restrict gene names to the correct Ensembl ID in the lookup table. In the one case where the ID provided by GTEx did not match any IDs for the gene name in the GENCODE lookup table (“TXNIP”), the ID from the lookup table was used instead. Four genes previously identified in NSCLC, 7 Cancer Gene Census genes, 5 EpiFactors genes, 51 CTAs, and 1 GTEx gene were not present in either the GENCODE lookup table or the NCBI gene alias table.

GENCODE IDs were used to link genes between gene sets and link genes and transcripts with gene set membership. The CGC “Role in Cancer” field was used to identify TSGs and oncogenes. The EpiFactors “Function” field was used to identify gene function.

### Alternative promoters

To identify potential instances of promoter switching in primary NSCLC, all genes with multiple promoters overlapping with DMRs and at least one hypoDMR and one hyperDMR in the same patient-matched comparison were identified. Genes exhibiting that combination in multiple patient-matched comparisons were selected for further investigation.

### Identification of small RNAs overlapping with patient-matched DMRs

MicroRNAs (miRNAs) and small RNAs whose gene bodies overlapped with patient-matched DMRs (transcript biotype: “miRNA”, “misc\_RNA”, “snRNA”, “snoRNA”, “rRNA”, “Mt\_rRNA”, and “Mt\_rRNA”) were identified. Predicted miRNA targets were obtained from miRDB information on the miRBase website.

### DAVID

DAVID was performed on lists of genes with patient-matched hypoDMRs or hyperDMRs in their promoter(s), scaled by the number of comparisons in which the gene overlapped with a DMR. Genes were looked up using “official gene symbol”. The categories OMIM\_DISEASE, COG\_ONTOLOGY, UP\_KEYWORDS, CHROMOSOME, BBID, BIOCARTA, KEGG\_PATHWAY, and UP\_TISSUE were tested. To accommodate the list of genes with hyperDMRs in the promoter in any patient, the list was first split into six smaller lists, and gene IDs that were not mappable by DAVID were excluded. The remaining gene names were combined into a single list and analyzed as before.

Terms with a Benjamini-corrected  $P < 0.05$  were considered significant.

### Roadmap Epigenomics Project epigenomes

Clinicopathologic data for the Roadmap Epigenomics Project consolidated epigenomes were obtained from the Roadmap Epigenomics Project data portal ([https://egg2.wustl.edu/roadmap/web\\_portal/meta.html](https://egg2.wustl.edu/roadmap/web_portal/meta.html), Consolidated\_EpigenomeIDs\_summary\_Table). Cancer cell lines were those with “leukemia” or “carcinoma” in the epigenome name ( $n = 5$ ), which were excluded from analyses unless otherwise stated. Group colors are those assigned in the Roadmap publication [64].

Roadmap Epigenomics Project ChromHMM annotations were downloaded from the Roadmap Epigenomics data portal, including 127 samples annotated with the 15-state model and 98 samples annotated with the 18-state model (<http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz> and [https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core\\_K27ac/jointModel/final/all.mnemonics.bedFiles.tgz](https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/all.mnemonics.bedFiles.tgz)). DMRs were intersected with the annotations, and the length of each DMR annotated with each ChromHMM state in each sample was calculated.

As background, the proportion of samples E096 and E114 (Lung and A549 EtOH 0.02pct lung carcinoma) in each 18-state ChromHMM state over 500-bp bins containing a CpG (M&M input, chromosomes 1–22 and X) was calculated. For E096, 500-bp bins were also split into promoter-overlapping, genic-exclusive (do not overlap with promoters), and intergenic-exclusive categories (do not overlap with promoters or genes). The proportion of all samples in each 15-state ChromHMM state over 500-bp bins containing a CpG (M&M input, chromosomes 1–22 and X) was also calculated.

Colors used for ChromHMM states are those assigned in [64] for the 15-state and 18-state models. Active regulatory states were states 1–3 and 6–7 for the 15-state model and states 1–4 and 7–11 for the 18-state model. Additionally, other 18-state ChromHMM states were assigned to composite states: transcribed (states 5 and 6) and polycomb (states 14–17).

methylCRF CpG methylation level estimates generated for 16 epigenomes by the Roadmap Epigenomics Project were downloaded from the Roadmap Epigenomics data portal and reformatted into bed format (Sample IDs: E003, E027, E028, E037, E038, E047, E053, E054, E055, E056, E057, E058, E059, E061, E081, E082). DMRs were intersected with

the methylation levels using *bedtools intersect*. Then, the mean methylation level over all CpGs overlapping with the DMR was calculated for all epigenomes. Methylation levels were divided into three categories: hypomethylated (< 30%), intermediately methylated (30%–70%), and hypermethylated (> 70%). Eight DMRs did not overlap with any CpG in the Roadmap files.

## HOMER

Enrichment analysis of known motifs was performed using HOMER *findMotifsGenome.pl* on each DMR set (hg19; parameters: *-size given -nomotif*). DMR sets included hypoDMRs and hyperDMRs in exclusive promoter, genic, and intergenic categories, scaled by the number of patient-matched comparisons in which they were found. The knownResults.txt output files were used in downstream analyses. Only motifs with a Benjamini-corrected  $P < 0.05$  were considered.

The TF name was extracted from the “Motif Name” field in the knownResults.txt file. TFs present in TCGA gene expression data were identified using case-insensitive exact name matches.

## GREAT

GREAT analysis (v4.0.4) was performed on each DMR set using the basal plus extension model (5000 bp upstream, 1000 bp downstream, 1,000,000 bp max extension), including curated regulatory regions. All significantly enriched GO biological processes with the following thresholds were downloaded: false discovery rate (FDR)  $P < 0.05$  by hypergeometric and binomial tests, at least one observed gene hit, and minimum 2-region-based fold enrichment (default). DMR sets included hypoDMRs and hyperDMRs in the intergenic-exclusive category, scaled by the number of patient-matched comparisons in which they were found.

## DMR density in windows

The average genome-wide DMR density for each sample comparison and direction of methylation change was calculated as the proportion of 500-bp bins overlapping with a CpG that were called as DMRs. The average DMR density across each chromosome was calculated in the same manner.

Non-overlapping windows of size 10 Mb, 1 Mb, 100 kb, and 10 kb were created using *bedtools makewindows* across the genome, excluding chromosome Y and chromosome M. The numbers of hypoDMRs and hyperDMRs per window were calculated for each patient-matched comparison. The DMR density per window was calculated as the proportion of 500-bp bins overlapping with a CpG called as DMRs, for each direction of methylation change.

The proportion of DMRs exclusive to a patient-matched comparison that fell within high-density windows was calculated for each comparison and direction of methylation change.

The CpG and repeat density for each window were calculated by normalizing by the length of the window. GENCODE transcripts overlapping with each window were also identified, and the density of genes, protein-coding genes, transcripts, and

protein-coding transcripts were calculated using the biotype information from GENCODE.

All windows were intersected with the CpG methylation level estimates generated by methylCRF using *bedtools intersect*. Then, the mean methylation level over all CpGs overlapping with the window was calculated for all samples. For comparisons to DMR density, the average methylCRF level was restricted to the normal sample for that comparison.

In addition, the proportion of each window annotated as each 18-state ChromHMM state in Roadmap sample E096 was calculated.

PCA was performed on all 1-Mb windows using gene, CpG, and repeat density and the proportion of the window in each ChromHMM state as features with the *prcomp()* function. Matrices were scaled and centered prior to transformation. Variance explained was calculated for each PC. The Pearson correlation of each variable with the first three PCs was calculated.

The Pearson correlation of all 1-Mb window features with DMR density in individual comparisons, the number of comparisons in which the window had > 1% DMR density, and the mean DMR density across comparisons were calculated.

## Clinicopathologic data-specific DMRs

Patient-matched DMRs specific to patient clinicopathologic data categories were identified using frequency information. Never-smoker-specific DMRs were those found between the normal lung *vs.* tumor samples of both never-smokers and not present in smokers, while smoker-specific DMRs were found in 3–4 smokers and no never-smokers, ignoring presence in patients with unconfirmed smoking status. Low-stage DMRs were those found in 4–5 patients with tumor stage 1/1A but not those with tumor stage 3A/3B, and high-stage DMRs were those found in at least three patients with tumor stage 3A/3B but not those with tumor stage 1/1A, ignoring presence in patients with unconfirmed tumor stage. Tumor subtype-specific DMRs were those found in all patients with that tumor subtype and in no other.

To determine whether the number of smoking status-specific DMRs was likely to be identified by chance, the smoking status of the six patients with confirmed status was shuffled to create all possible permutations with two never-smokers and four smokers ( $n = 15$ ). Then, the number of smoking status-specific DMRs was recalculated for each permutation. For subtype-exclusive DMRs, the number of exclusive DMRs was recalculated for all possible pairs of patients ( $n = 45$ ) and compared to the true number for acinar/acinar cell adenocarcinoma and adenosquamous carcinoma.

GREAT analysis was performed on patient-matched DMRs exclusive to never-smokers, smokers, adenosquamous carcinoma, and acinar cell/acinar cell adenocarcinoma by direction of methylation change using the same parameters as above. GREAT analysis was also performed on DMRs between never-smoker and smoker normal samples that were not found between the never-smoker samples. The hypoDMRs exclusive to each category did not return any significant GO biological processes.

The proportion of smoking status-specific DMRs that fell within high-density 1-Mb windows of the same direction of methylation change in each patient was calculated.



To determine whether smoking was associated with DNA methylation alterations in normal lung samples, normal *vs.* normal sample DMRs were assigned to never–never, smoker–smoker, or never–smoker comparison categories ( $n = 1214$ ). DMRs found between never-smoker and smoker normal samples, but not between the never-smoker samples, were identified. Additionally, for each normal sample, the number of DMRs found between the sample and never-smokers and the sample and smokers was calculated and tested for significant differences using Wilcoxon tests. Example DMRs ( $n = 28$ ) were identified as those found in three or more never-smoker *vs.* smoker comparisons, including both never-smoker samples, and not between the two never-smoker samples.

## TCGA

TCGA LUAD CpG methylation and gene, isoform, and miRNA expression data were downloaded from the TCGA legacy data portal using the `gdc-client` and manifests generated through the Genomic Data Commons legacy data portal (<https://portal.gdc.cancer.gov/legacy-archive/>).

Clinicopathologic data for all downloaded LUAD files ( $n = 2214$ ) was obtained from the legacy Application Programming Interface endpoint using the file ID, including file name, data category, data type, sample ID, sample type, and case ID. A summary of the number of samples and files by data type per case is provided (Table S3). All samples with gene expression data also have isoform expression data and *vice versa*.

All 450K array methylation level files were downloaded (data type: “Methylation beta value”; platform: “Illumina Human Methylation 450”;  $n = 507$ ). The average CpG methylation level over DMRs identified in this study and GENCODE transcript promoters was calculated in all TCGA LUAD and matched normal samples. First, promoters or unique DMRs with  $Q < 0.001$  were intersected with the 450K array CpG methylation levels from all LUAD samples. Then, the mean methylation level over all CpGs within each DMR/promoter in each TCGA sample was calculated. DMRs or promoters without methylation data in any TCGA sample were excluded from analysis.

The mean methylation level over each patient-matched DMR was averaged across TCGA samples by sample type (matched normal lung or primary LUAD) and by DMR direction of methylation change, excluding samples in which the DMR was missing methylation data. The mean methylation level over patient-matched DMRs identified in smokers and never-smokers (either at all or recurrently) was also calculated across TCGA samples, by sample type, DMR direction of methylation change, and TCGA smoker status designation, again excluding samples in which the DMR was missing methylation data. TCGA samples were classified as either smokers (having a value in the metadata category “cigarettes per day” or “years smoked”) or never-smokers (empty values in both metadata categories).

To compare 450K array CpG coverage to methylCRF CpG coverage, TCGA CpG positions were extended 1 bp upstream, and CpGs without a chromosome were excluded ( $n = 65$ ). Of the remaining 485,512 CpGs, 0.8% did not overlap with a methylCRF CpG ( $n = 3868$ ). 98.3% of methylCRF CpGs ( $n = 27,603,511$ ) did not overlap with a TCGA CpG.

All normalized gene and isoform expression files were downloaded (experimental strategy: “RNA-seq”; data type: “Gene expression quantification” or “Isoform expression quantification”; filtered to files ending in “rsem.genes.normalized\_results” or “rsem.isoforms.normalized\_results”;  $n = 576$  each).

Gene symbols for gene and isoform expression levels in TCGA were extracted from the gene IDs in the gene expression files and the TCGA hg19 `gaf` file, restricted to “transcript” entries. Mappings between UCSC knownGene isoforms and Ensembl transcripts (`hg19.knownToEnsembl`) were obtained from the UCSC Table Browser, and isoform IDs used in TCGA were connected to Ensembl transcripts using abbreviated isoform IDs without version suffixes.

For selected genes, average promoter CpG methylation, gene expression, and isoform expression levels were plotted for all TCGA samples, comparing LUAD primary tumors and matched normal lung samples. All GENCODE transcripts associated with the gene name were included in the promoter methylation analyses. All TCGA genes and isoforms associated with the gene name were included in the expression analyses. The difference in median promoter methylation level or the fold change in median expression level was calculated between primary tumors and matched normal lung samples, and Wilcoxon tests were used to test for significant differences for each gene, transcript, or isoform. Where possible, isoform IDs were linked to GENCODE transcript IDs. Because some promoters did not overlap with CpGs and some GENCODE gene names were not included in the TCGA expression data, all analyses could not be performed for all genes.

All miRNA files processed with miRBase (v20) were downloaded (experimental strategy: “miRNA-Seq”; data type: “miRNA gene quantification”; filtered to “mirbase20”;  $n = 555$  files). miRNAs were linked to gene symbols using the TCGA hg19 `gaf` file and the NCBI gene alias table. The TCGA hg19 `gaf` file was restricted to “pre-miRNA” entries, and gene names and miRNA IDs were extracted from their respective columns. Additionally, genes with synonyms containing “has-” were extracted from the NCBI gene alias table along with the associated gene names. Both sources were used to link miRNA IDs assigned by miRBase with gene names in the TCGA data.

The miRNA expression levels of specific genes in TCGA LUAD and normal lung samples were plotted and tested for significant differences between the sample types using a Wilcoxon test. All miRNA IDs associated with the gene name were included.

## Repeats

TEs were considered those with class: “DNA”, “DNA?”, “LINE”, “LINE?”, “LTR”, “LTR?”, “Other”, “RC”, “SINE”, “SINE?”, “Unknown”, and “Unknown?”. Only repeats on chromosomes 1–22 and X were considered.

The Jukes–Cantor evolutionary distance for each repeat was calculated from the substitution rate compared to the Repbase subfamily consensus sequence. The total length of bases overlapping with each subfamily, the number of elements, the median Jukes–Cantor distance, and the number of unique methylCRF CpGs overlapping with each subfamily were calculated.



The average methylation over each subfamily in each sample was calculated by identifying all unique CpGs that overlapped with the subfamily, then calculating the mean methylation level assigned by methylCRF over all CpGs in each sample. Seventeen subfamilies did not overlap with any CpG. For each subfamily, a Wilcoxon test was performed on the mean methylation level in normal vs. tumor samples. Only subfamilies with at least 10 CpGs were considered for downstream analyses.

The LOR enrichment for each subfamily and DMR direction was calculated from the proportion of DMRs overlapping with the subfamily vs. the proportion of total repeat length represented by the subfamily. Enrichment was calculated for all DMRs and those found in > 1 patient.

For all repeats overlapping with a patient-matched DMR, the number of unique DMRs overlapping with the repeat, the total number of DMRs overlapping with the repeat across comparisons, and the number of comparisons in which the repeat overlapped with a DMR were calculated by direction of methylation change. The closest gene on each strand was identified using *bedtools closest*, including all tied genes, ignoring genes upstream of the repeat on the same strand and downstream of the repeat on the opposite strand. TEs overlapping with hypoDMRs present in > 1 comparison upstream of a gene on the sense strand (or downstream on the opposite strand, in the case of LINE elements) were identified.

A list of TEs that splice into oncogenes to form alternative isoforms in TCGA cancer samples was downloaded from [15] (Table S2). The TE coordinates were lifted over to hg19 from hg38 using LiftOver, and TEs that overlapped with hypoDMRs in this study were identified.

Repeat class colors were those assigned on the WashU Epigenome Browser RepeatMasker track.

## Software

The following software packages were used: *vegan* (v2.5.6), *ggplot* (v2 3.2.1), *reshape* (v2 1.4.3), *plyr* (v1.8.4), *grid* (v3.5.1), *gridExtra* (v2.3), *RColorBrewer* (v1.1.2), *knitr* (v1.26), *readxl* (v1.3.1), *UpSetR* (v1.4.0), and *combinat* (v0.0.8).

## Ethical statement

An IRB-approved waiver of informed consent (Approval No. 201305031) was obtained for all patient samples analyzed in the present study.

## Code availability

Custom scripts generated for analyzing data in the present study have been made publicly available at <https://github.com/jaflynn5/DNA-Methylation-Changes-in-NSCLC-by-Smoking-Status>.

## Data availability

The MeDIP-seq, MRE-seq, M&M, and methylCRF data generated in this study are publicly available at a public WashU Epigenome Browser datahub ([https://epigenomegateway.wustl.edu/browser/?genome=hg19&hub=https://wangftp.wustl.edu/~epehrsson/NSCLC\\_primary/NSCLC\\_primary](https://epigenomegateway.wustl.edu/browser/?genome=hg19&hub=https://wangftp.wustl.edu/~epehrsson/NSCLC_primary/NSCLC_primary)).

Tracks are linked to sample and patient clinicopathologic data. In addition, all sequencing data generated as a part of this study have been deposited in GEO (GEO: GSE210957).

## Competing interests

Siddhartha Devarakonda served on the advisory board for AstraZeneca, Merus, Jazz Pharmaceuticals, and Genentech. All the other authors have declared no competing interests.

## CRedit authorship contribution statement

**Jennifer A. Karlow:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Erica C. Pehrsson:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Xiaoyun Xing:** Methodology, Investigation, Writing – original draft. **Mark Watson:** Conceptualization, Investigation, Data curation, Resources, Writing – original draft, Supervision, Funding acquisition. **Siddhartha Devarakonda:** Investigation, Data curation, Resources, Writing – original draft. **Ramaswamy Govindan:** Conceptualization, Investigation, Resources, Writing – original draft, Supervision, Funding acquisition. **Ting Wang:** Conceptualization, Validation, Formal analysis, Investigation, Data curation, Resources, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

## Acknowledgments

Jennifer A. Karlow is supported in part by the Siteman Cancer Center Precision Medicine Pathway (Grant No. T32CA113275). Erica C. Pehrsson is supported by a Postdoctoral Fellowship from the American Cancer Society (Grant No. PF-17-201-01). Erica C. Pehrsson, Jennifer A. Karlow, and Ting Wang are supported by the National Institutes of Health [Grant Nos. R01HG007354 (National Human Genome Research Institute), R01HG007175 (National Human Genome Research Institute), R01ES024992 (National Institute of Environmental Health Sciences), U01CA200060 (National Cancer Institute), U24ES026699 (National Institute of Environmental Health Sciences), and U01HG009391 (National Human Genome Research Institute)] and the American Cancer Society (Grant No. RSG-14-049-01-DMC).

## Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gpb.2023.03.006>.

## ORCID

ORCID 0000-0001-5041-2669 (Jennifer A. Karlow)  
ORCID 0000-0002-2968-6179 (Erica C. Pehrsson)  
ORCID 0000-0002-1045-1775 (Xiaoyun Xing)  
ORCID 0000-0002-3935-9980 (Mark Watson)

ORCID 0000-0003-2693-3848 (Siddhartha Devarakonda)

ORCID 0000-0002-6964-9612 (Ramaswamy Govindan)

ORCID 0000-0002-6800-242X (Ting Wang)

## References

- [1] National Cancer Institute. Cancer stat facts: lung and bronchus cancer. [Internet]. Bethesda, USA: National Cancer Institute; 2022. <https://seer.cancer.gov/statfacts/html/lungb.html>.
- [2] Nawaz K, Webster RM. The non-small-cell lung cancer drug market. *Nat Rev Drug Discov* 2016;15:229–30.
- [3] Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. *Science* 2015;348:69–74.
- [4] Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature* 2018;553:446–54.
- [5] Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic landscape of non-small cell lung cancer in smokers and never smokers. *Cell* 2012;150:1121–34.
- [6] Subramanian J, Govindan R. Molecular profile of lung cancer in never smokers. *EJC Suppl* 2013;11:248–53.
- [7] Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. *Science* 2016;354:618–22.
- [8] Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat Genet* 2016;48:607–16.
- [9] Baylin SB, Jones PA. Epigenetic determinants of cancer. *Cold Spring Harb Perspect Biol* 2016;8:a019505.
- [10] Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* 2016;351:1454–8.
- [11] Lamprecht B, Bonifer C, Mathas S. Repeat-element driven activation of proto-oncogenes in human malignancies. *Cell Cycle* 2010;9:4276–81.
- [12] Wolff EM, Byun HM, Han HF, Sharma S, Nichols PW, Siegmund KD, et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet* 2010;6:e1000917.
- [13] Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. *Mob DNA* 2016;7:24.
- [14] Brocks D, Schmidt CR, Daskalakis M, Sik Jang H, Shah NM, Li D, et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat Genet* 2017;49:1052–60.
- [15] Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, et al. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet* 2019;51:611–7.
- [16] Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* 2015;162:974–86.
- [17] Roulois D, Yau HL, Singhanian R, Wang Y, Danesh A, Shen SY, et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* 2015;162:961–73.
- [18] Wrangle J, Wang W, Koch A, Easwaran H, Mohammad HP, Vendetti F, et al. Alterations of immune response of non-small cell lung cancer with azacytidine. *Oncotarget* 2013;4:2067–79.
- [19] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;489:519–25.
- [20] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 2014;511:543–50.
- [21] Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res* 2013;23:1541–53.
- [22] Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res* 2013;23:1522–40.
- [23] Karlow JA, Devarakonda S, Xing X, Jang HS, Govindan R, Watson MA, et al. Developmental pathways are epigenetically reprogrammed during lung cancer brain metastasis. *Cancer Res* 2022;82:2692–703.
- [24] Karlow JA, Miao B, Xing X, Wang T, Zhang B. Common DNA methylation dynamics in endometrial adenocarcinoma and glioblastoma suggest universal epigenomic alterations in tumorigenesis. *Commun Biol* 2021;4:607.
- [25] Weissferdt A, Moran CA. Reclassification of early stage pulmonary adenocarcinoma and its consequences. *J Thorac Dis* 2014;6:S581–8.
- [26] Yang J, Wang S, Yang Z, Hodgkinson CA, Iarikova P, Ma JZ, et al. The contribution of rare and common variants in 30 genes to risk nicotine dependence. *Mol Psychiatry* 2015;20:1467–78.
- [27] Raghuvanshi SK, Nasser MW, Chen X, Strieter RM, Richardson RM. Depletion of  $\beta$ -arrestin-2 promotes tumor growth and angiogenesis in a murine model of lung cancer. *J Immunol* 2008;180:5699–706.
- [28] Furukawa C, Daigo Y, Ishikawa N, Kato T, Ito T, Tsuchiya E, et al. Plakophilin 3 oncogene as prognostic marker and therapeutic target for lung cancer. *Cancer Res* 2005;65:7102–10.
- [29] Chhabra D, Sharma S, Kho AT, Gaedigk R, Vyhldal CA, Leeder SJ, et al. Fetal lung and placental methylation is associated with in utero nicotine exposure. *Epigenetics* 2014;9:1473–84.
- [30] Richardson TG, Richmond RC, North TL, Hemani G, Smith GD, Sharp GC, et al. An integrative approach to detect epigenetic mechanisms that putatively mediate the influence of lifestyle exposures on disease susceptibility. *Int J Epidemiol* 2019;48:887–98.
- [31] Han Y, Li G, Su C, Ren H, Chu X, Zhao Q, et al. Exploratory study on the correlation between 14 lung cancer-related gene expression and specific clinical characteristics of NSCLC patients. *Mol Clin Oncol* 2013;1:887–93.
- [32] Kaczkowski B, Tanaka Y, Kawaji H, Sandelin A, Andersson R, Itoh M, et al. Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers. *Cancer Res* 2016;76:216–26.
- [33] Tang Z, Li J, Shen Q, Feng J, Liu H, Wang W, et al. Contribution of upregulated dipeptidyl peptidase 9 (DPP9) in promoting tumorigenicity, metastasis and the prediction of poor prognosis in non-small cell lung cancer (NSCLC). *Int J Cancer* 2017;140:1620–32.
- [34] Fraile JM, Manchado E, Lujambio A, Quesada V, Campos-Iglesias D, Webb TR, et al. USP39 deubiquitinase is essential for KRAS oncogene-driven cancer. *J Biol Chem* 2017;292:4164–75.
- [35] Wang R, Deng X, Yoshioka Y, Vougiouklakis T, Park JH, Suzuki T, et al. Effects of SMYD2-mediated EML4-ALK methylation on the signaling pathway and growth in non-small-cell lung cancer cells. *Cancer Sci* 2017;108:1203–9.
- [36] Richter GM, Kruppa J, Munz M, Wiehe R, Häslker R, Franke A, et al. A combined epigenome- and transcriptome-wide association study of the oral masticatory mucosa assigns CYP1B1 a central role for epithelial health in smokers. *Clin Epigenetics* 2019;11:105.
- [37] Richtmann S, Wilkens D, Warth A, Lasitschka F, Winter H, Christopoulos P, et al. FAM83A and FAM83B as prognostic biomarkers and potential new therapeutic targets in NSCLC. *Cancers* 2019;11:652.
- [38] Dhandapani L, Yue P, Ramalingam SS, Khuri FR, Sun SY. Retinoic acid enhances TRAIL-induced apoptosis in cancer cells by upregulating TRAIL receptor 1 expression. *Cancer Res* 2011;71:5245–54.

- [39] Milewski D, Balli D, Ustiyani V, Le T, Dienemann H, Warth A, et al. FOXM1 activates AGR2 and causes progression of lung adenomas into invasive mucinous adenocarcinomas. *PLoS Genet* 2017;13:e1007097.
- [40] Yang X, Gao L, Zhang S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. *Brief Bioinform* 2017;18:761–73.
- [41] Rauch T, Wang Z, Zhang X, Zhong X, Wu X, Lau SK, et al. Homeobox gene methylation in lung cancer studied by genome-wide analysis with a microarray-based methylated CpG island recovery assay. *Proc Natl Acad Sci U S A* 2007;104:5527–32.
- [42] Shiraishi M, Sekiguchi A, Oates AJ, Terry MJ, Miyamoto Y. *HOX* gene clusters are hotspots of *de novo* methylation in CpG islands of human lung adenocarcinomas. *Oncogene* 2002;21:3659–62.
- [43] Zhou X, Updegraff BL, Guo Y, Peyton M, Girard L, Larsen JE, et al. PROTOCADHERIN 7 acts through SET and PP2A to potentiate MAPK signaling by EGFR and KRAS during lung tumorigenesis. *Cancer Res* 2017;77:187–97.
- [44] Yamaguchi T, Hosono Y, Yanagisawa K, Takahashi T. NKX2-1/TTF-1: an enigmatic oncogene that functions as a double-edged sword for cancer cell survival and progression. *Cancer Cell* 2013;23:718–23.
- [45] Zhang Y, Xu X, Zhang M, Wang X, Bai X, Li H, et al. MicroRNA-663a is downregulated in non-small cell lung cancer and inhibits proliferation and invasion by targeting JunD. *BMC Cancer* 2016;16:315.
- [46] Zhang C, Chen B, Jiao A, Li F, Sun N, Zhang G, et al. miR-663a inhibits tumor growth and invasion by regulating TGF- $\beta$ 1 in hepatocellular carcinoma. *BMC Cancer* 2018;18:1179.
- [47] Cho JG, Park S, Lim CH, Kim HS, Song SY, Roh TY, et al. ZNF224, Krüppel like zinc finger protein, induces cell growth and apoptosis-resistance by down-regulation of p21 and p53 via miR-663a. *Oncotarget* 2016;7:31177–90.
- [48] Chang RM, Xiao S, Lei X, Yang H, Fang F, Yang LY. miRNA-487a promotes proliferation and metastasis in hepatocellular carcinoma. *Clin Cancer Res* 2017;23:2593–604.
- [49] O'Reilly D, Dienstbier M, Cowley SA, Vazquez P, Drozd M, Taylor S, et al. Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. *Genome Res* 2013;23:281–91.
- [50] Stadler PF, Chen JLL, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, et al. Evolution of vault RNAs. *Mol Biol Evol* 2009;26:1975–91.
- [51] Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat Commun* 2019;10:5640.
- [52] Powrózek T, Krawczyk P, Kucharczyk T, Milanowski J. Septin 9 promoter region methylation in free circulating DNA - potential role in noninvasive diagnosis of lung cancer: preliminary report. *Med Oncol* 2014;31:917.
- [53] Shen SY, Singhanian R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* 2018;563:579–83.
- [54] Hulbert A, Jusue-Torres I, Stark A, Chen C, Rodgers K, Lee B, et al. Early detection of lung cancer using DNA promoter hypermethylation in plasma and sputum. *Clin Cancer Res* 2017;23:1998–2005.
- [55] Sur I, Taipale J. The role of enhancers in cancer. *Nat Rev Cancer* 2016;16:483–93.
- [56] Lee KWK, Pausova Z. Cigarette smoking and DNA methylation. *Front Genet* 2013;4:132.
- [57] Freeman JR, Chu S, Hsu T, Huang YT. Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. *Oncotarget* 2016;7:69579–91.
- [58] Tessema M, Yingling CM, Liu Y, Tellez CS, Van Neste L, Baylin SS, et al. Genome-wide unmasking of epigenetically silenced genes in lung adenocarcinoma from smokers and never smokers. *Carcinogenesis* 2014;35:1248–57.
- [59] Huang T, Chen X, Hong Q, Deng Z, Ma H, Xin Y, et al. Meta-analyses of gene methylation and smoking behavior in non-small cell lung cancer patients. *Sci Rep* 2015;5:8897.
- [60] Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* 2016;18:71–86.
- [61] Li D, Zhang B, Xing X, Wang T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* 2015;72:29–40.
- [62] Li J, Xing X, Li D, Zhang B, Mutch DG, Hagemann IS, et al. Whole-genome DNA methylation profiling identifies epigenetic signatures of uterine carcinosarcoma. *Neoplasia* 2017;19:100–11.
- [63] Rauch TA, Zhong X, Wu X, Wang M, Kernstine KH, Wang Z, et al. High-resolution mapping of DNA hypermethylation and hypomethylation in lung cancer. *Proc Natl Acad Sci U S A* 2008;105:252–7.
- [64] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518:317–30.