Integrative CAGE and DNA Methylation Profiling Identify Epigenetically Regulated Genes in NSCLC

Masafumi Horie1,2,3, Bogumil Kaczkowski3, Mitsuhiro Ohshima4, Hirotaka Matsuzaki1, Satoshi Noguchi1, Yu Mikami1,5, Marina Lizio3, Masayoshi Itoh3,6, Hideya Kawaji3,6, Timo Lassmann3,7, Piero Carninci3, Yoshihide Hayashizaki6, Alistair R.R. Forrest3,8, Daiya Takai5, Yoko Yamaguchi9,10, Patrick Mickey11, Akira Saito1,2, and Takahide Nagase1

Abstract

Lung cancer is the leading cause of cancer-related deaths worldwide. The majority of cancer driver mutations have been identified; however, relevant epigenetic regulation involved in tumorigenesis has only been fragmentingly analyzed. Epigenetically regulated genes have a great theranostic potential, especially in tumors with no apparent driver mutations. Here, epigenetically regulated genes were identified in lung cancer by an integrative analysis of promoter-level expression profiles from Cap Analysis of Gene Expression (CAGE) of 16 non–small cell lung cancer (NSCLC) cell lines and 16 normal lung primary cell specimens with DNA methylation data of 69 NSCLC cell lines and 6 normal lung epithelial cells. A core set of 49 coding genes and 10 long noncoding RNAs (lncRNA), which are upregulated in NSCLC cell lines due to promoter hypomethylation, was uncovered. Twenty-two epigenetically regulated genes were validated (upregulated genes with hypomethylated promoters) in the adenocarcinoma and squamous cell cancer subtypes of lung cancer using The Cancer Genome Atlas data. Furthermore, it was demonstrated that multiple copies of the REP522 DNA repeat family are prominently upregulated due to hypomethylation in NSCLC cell lines, which leads to cancer-specific expression of IncRNAs, such as RPI-10G24.10, AL022344.4, and PCAT7. Finally, Myeloma Overexpressed (MYEOV) was identified as the most promising candidate. Functional studies demonstrated that MYEOV promotes cell proliferation, survival, and invasion. Moreover, high MYEOV expression levels were associated with poor prognosis.

Implications: This report identifies a robust list of 22 candidate driver genes that are epigenetically regulated in lung cancer; such genes may complement the known mutational drivers.

Introduction

Lung cancer is the leading cause of cancer-related deaths worldwide, with non–small cell lung cancer (NSCLC) representing the majority of cases (1). In recent years, the genomic landscape of NSCLC has been extensively investigated (2–4). However, for the majority of NSCLC patients, treatment options are still limited, and the overall prognosis remains poor. Thus, new robust biomarkers and therapeutic targets are required to guide and aim the treatment. Beyond the mutation profiles, epigenetic changes and deregulation of gene expression are decisive for cancer pathophysiology (5). Vogelstein and colleagues introduced a term “epigenetic driver” for genes that are altered by epigenetic mechanisms and confer a selective growth advantage by differential gene expression (6).

Although tens of mutational drivers are known, epigenetic drivers represent a largely unknown territory with a great potential for finding theranostic targets (6, 7). Optimal transcriptional biomarkers should be stably upregulated. As DNA methylation is involved in stable and long-term regulation, we hypothesize that differentially expressed genes that are caused by aberrant DNA methylation are optimal candidates for biomarkers. Differential expression driven by other factors, for
example, aberrant signaling or broken regulatory networks may be variable over time.

Here, we apply a novel approach to deeply integrate transcriptomics and epigenetics data to find robust biomarkers that are regulated at epigenetic level (epi-markers/drivers), including novel long noncoding RNAs (IncRNAs) and transcribed repetitive elements. We integrate the data from Functional AnnoTation Of Mammalian genome 5 (FANTOM5) and The Cancer Genome Atlas (TCGA) consortia, together with other available datasets, to find the optimal set of epigenetically regulated genes in lung cancer. Such genes can then be used as biomarkers, and we expect a subset of them to be epi-drivers.

As DNA methylation can have a different function depending on where on the gene it occurs (8), the optimal data integration has to be performed at promoter resolution. To do that, we utilized the unique Cap Analysis of Gene Expression (CAGE) dataset from the FANTOM5 consortium. CAGE is a 5’ sequence tag technology that globally determines transcription start sites (TSS) in the genome and their expression levels (9, 10). FANTOM5 dataset contains CAGE profiles of 16 NSCLC cell lines and 16 normal lung epithelial cells that allow differential expression analysis at the promoter/TSS level. This work is part of the FANTOM5 project. Data downloads, genomic tools, and copublished manuscripts are summarized online at http://fantom.gsc.riken.jp/5/.

Materials and Methods
FANTOM5 CAGE data
We obtained the CAGE expression data of 16 normal lung epithelial cells and 16 NSCLC cell lines from the FANTOM5 project (Supplementary Table S1). We used the expression table containing CAGE tag counts under 184,827 CAGE peaks, which represents genome-wide, promoter-level expression profiles (9). We selected 65,131 active promoters (tags per million > 1 in at least two samples) and annotated the CAGE peaks using the GENCODE v19, miTranscriptome, and PLAR annotations (11–13).

To identify up- and downregulated transcripts in NSCLC cell lines, we performed differential expression analysis using edgeR (14). The P values were adjusted for multiple testing by Benjamini–Hochberg method. The thresholds of absolute fold change > 2 and FDR < 0.05 were used. We also evaluated the expression status in binary fashion: ON (expressed, count > 0), OFF (not detected, count = 0), as described before (15). In short, we calculated the frequency of expression (OFF/ON analysis), and features expressed four times more frequently in NSCLC were determined as turned ON.

TCGA RNA-seq data
We obtained the gene expression profiles of 515 lung adenocarcinoma (LUAD) and 59 normal tissue controls, and 501 lung squamous cell carcinoma (LUSC) and 49 normal lung tissue controls from The Cancer Genome Atlas (TCGA) Data Portal (data status as of June 1, 2016). The level3 RNASeqV2 expression profiles of 20,502 genes were downloaded. The expression levels of recently annotated IncRNAs were downloaded from the Genomic Data Commons Data Portal (HTSeq pipeline, GDC.h38 GENCODE v22, https://gdc-portal.nci.nih.gov). The data were log2 transformed and used for the differential expression analysis by limma package (16). The thresholds of absolute fold change > 2 and FDR < 0.01 were used. We also performed OFF/ON analysis as described above.

DNA methylation array data
We obtained TCGA and GSE36216 (Supplementary Table S2) datasets generated by Illumina Infinium 450K methylation array (HumanMethylation450 BeadChip) platform (17). Probes that map to multiple genomic locations and nonspecific probes were removed from analyses according to the previous report (18). We processed the data using the minfi package to generate the β values for DNA methylation status, and to analyze the differences in methylation levels between cancer and normal groups (19). FDR < 0.05 was considered significant.

Integration of DNA methylation data and CAGE/RNA-seq
Each CAGE promoter was linked to DNA methylation sites within −1500 to +500 bp from the center of CAGE peak. Genomic positions of the probes were taken from the annotation file of the Infinium 450K methylation array. Gene symbols from the annotation file were used for integration of RNA sequencing (RNA-seq) and DNA methylation data. RNA-seq (E-MTAB-2706) and Infinium 450K methylation array (GSE36216) data for the same panel of NSCLC cell lines were used to validate the correlations between gene expression and methylation (20). Spearman ρ < −0.3 was used to show inverse correlations between expression and methylation levels in NSCLC cell lines or TCGA tumor samples.

Survival analyses in NSCLC tumors
For meta-analyses, we used 6 datasets for LUSC and 5 datasets for LUAD that used the Affymetrix GeneChip U133 Plus 2.0 microarray and consisted of more than 40 cancer tissue samples (GSE30219, GSE31210, GSE3141, GSE37745, GSE50081, E-MTAB-923, and E-MTAB-1727; Supplementary Table S2; refs. 21–27). Univariate and multivariate Cox models were fitted with the inclusion of recognized prognostic parameters; age (>70 vs. ≤70 years), gender, smoking status (ever-smoker vs. never-smoker), and stage at diagnosis (stage III–IV vs. I–II) as described previously (28).

Chromatin modifications in LUAD cells
Chromatin immunoprecipitation (ChIP)-seq, bisulfite (BS)-seq, and RNA-seq data of 26 LUSC cell lines and small airway epithelial cells (SAEC) were obtained from the DNA Data Bank of Japan (Supplementary Table S2; refs. 29, 30). Mapped sequence data were visualized using Integrative Genomics Viewer (31). Enrichment plots were generated using the ngs.plot package (32).

Cell cultures
Human lung cancer cell lines (A549, NCI-H441, NCI-H358, NCI-H1395, NCI-H1437, NCI-H1648, NCI-H1792, NCI-H1975, NCI-H2170, SK-MES-1, and L1210) were obtained from the ATCC. Lu99, EBC-1, and LC-2/ad were obtained from RIKEN BRC. Normal human bronchial epithelial (NHBE) cells and SAECs were purchased from TAKARA BIO INC. and LONZa, respectively. Human alveolar epithelial cell (ACEC) total RNA (HPAEpiCRNA, #3205) was from ScienCell.

Quantitative RT-PCR
Detailed procedures were described previously (33). GAPDH expression levels were used for the normalization. The PCR
Horie et al.

primers are as follows. GAPDH forward: GGTGAAAGTGGGGTCAACGG, reverse: GAGGATCTGGCTGGGCTGGAAGA. MYEOV forward: GACAGCCGATTACGCG, reverse: CATCCACGCGGCCACATAT.

siRNA experiment in A549 cells

siRNA against human MYEOV (Stealth RNAi, siMYEOV-1: HS1210166, siMYEOV-2: HS1210167) and the negative control (Stealth RNAi, siNC: #12935-300) were purchased from Life Technologies. A549 and H441 cells were transfected with 20 nM/siRNA using Lipofectamine RNAiMAX (Life Technologies). Cell proliferation, apoptosis, and invasion assays were performed as described previously (28, 33). Total RNA was extracted from A549 cells 24 hours after siRNA transfection using the RNAeasy Mini Kit. Microarray analysis was carried out using Affymetrix GeneChip Human Genome U133 Plus 2.0. Significance Analysis of Microarrays (SAM) was used for statistical analyses (34). The dataset was deposited in the GEO database (GSE65968).

The microarray CEL files were submitted to the ISMARA (The Integrated System for Motif Activity Response Analysis) online tool and selected MYC as the top transcription factor with highest activity significance (35). To find genes regulated by MYC, we obtained cMYC ChIP-seq peaks in A549 cells from ENCODE (36) and selected two nearest genes within 1 kb from the peak using GREAT 3.0.0 tool (37).

5-aza-dC and TSA treatment

SAECs were treated with 3 μg/mL of DNA demethylating agent, 5-aza-2'-deoxycytidine (5-aza-dC) on day 1, 3, and 1 μg/mL of histone deacetylase (HDAC) inhibitor, trichostatin A (TSA) on day 4. Total RNA was extracted on day 5, and microarray analysis was carried out using Agilent SurePrint G3 Human Gene Expression 8 × 60K v3. The dataset was deposited in the GEO database (GSE82214).

Miscellaneous statistical analyses

For cell culture experiments, differences were examined by ANOVA with Tukey post hoc test with JMP, version 11 (SAS Institute Inc.). Fisher exact test was used to assess the frequencies of promoters, genes, and repeat families. $P < 0.05$ was considered significant. All data are expressed as means ± SEM.

Results

To find the optimal set of epigenetically regulated genes in lung cancer, we performed an integrative analysis of datasets from the FANTOM5 and TCGA consortia. In the first step, we aimed to (i) identify differentially expressed (DE) promoters in NSCLC (FANTOM5 CAGE data); (ii) identify differentially methylated (DM) sites in NSCLC (69 NSCLC cell lines compared with 6 normal lung epithelial cells); (iii) link the differentially methylated sites to differentially expressed genes at promoter resolution, and (iv) validate the epigenetically regulated genes (DM-DE genes) in large lung cancer datasets from TCGA (LIAD and LIISC).

Upregulated promoters in NSCLC cell lines

Using the CAGE data of the FANTOM5 project, we compared the transcription of 16 NSCLC cell lines and 16 normal lung epithelial cells (Supplementary Table S1). We used the tag counts within 184,827 CAGE peaks as quantitative measures of promoter-level expression. Multidimensional scaling plot demonstrated the distinctive patterns of CAGE profiles of cancer and normal cells (Fig. 1A).

First, we performed DE analysis (edgeR, fold change > 2, FDR < 0.05) and found 10,433 upregulated promoters (8,512 coding genes). Within the upregulated promoters, we performed OFF/ON analysis (see Materials and Methods) and found 2,988 promoters (2,168 coding genes) that were turned ON in NSCLC cell lines (Table 1; Supplementary Table S3; ref. 15).

Hypomethylated sites in NSCLC cell lines

To identify DM sites in lung cancer, we analyzed the methylation profiles ($β$ values) in the dataset of 69 NSCLC cell lines and 6 normal lung epithelial cells (Infinium 450K methylation array, GSE36216; ref. 17). Multidimensional scaling showed the distinctive patterns of DNA methylation in cancer and normal cells (Fig. 1B). We performed differential methylation analysis and identified 66,826 hypermethylated and 8,634 hypomethylated sites in NSCLC cell lines compared with normal cells (absolute mean $β$ value difference > 0.2, FDR < 0.05).

Linking differential expression to the DNA methylation changes

To integrate the promoter-level differential expression with aberrant DNA methylation results, we directly linked the CAGE-defined promoters to DNA methylation probes located within $<1500$ to $>500$ bp from the promoters. In some cases, multiple probes were linked to a single promoter. In addition, we used the matched gene expression and methylation profiles in 58 NSCLC cell lines (RNA-seq, E-MTAB-2706; Infinium 450K methylation array, GSE36216) to select methylation probes whose methylation inversely correlated with gene expression (Spearman $ρ < −0.3$, $P < 0.05$).

We overlapped the upregulated promoters from FANTOM5 CAGE analysis with the hypomethylated sites and found 379 methylation probe–promoter pairs where the promoter was upregulated in FANTOM5 cancer cell lines, and the methylation site was hypomethylated in cancer versus normal cell comparison (Fig. 1C; Supplementary Table S4). Among them, we further selected 159 methylation probe–promoter pairs that showed inverse correlation (Spearman $ρ < −0.3$, $P < 0.05$) between DNA methylation of the probe and expression of the gene in 58 NSCLC cell lines.

At the gene level, this corresponded to 49 protein-coding genes and 10 noncoding RNAs that were upregulated in cancer and had at least one associated hypomethylated probe in NSCLC (Supplementary Table S4). Expression and methylation of representative genes in 58 NSCLC cell lines are shown in Fig. 1D.

Validation in TCGA dataset

To confirm that our candidates for epigenetically regulated biomarkers are relevant to clinical tumors, we performed a rigorous validation using RNA-seq and DNA methylation data of 426 LIAD and 359 LIISC samples from TCGA consortium. As above, we calculated the correlation between DNA methylation of the probes and expression of the associated genes (Spearman $ρ < −0.3$, $P < 0.05$). We also selected genes that were upregulated or turned ON in tumors compared with normal lung tissues (Supplementary Table S4). We found a robust signature of 22 epigenetically regulated biomarkers (epi-markers; 15 coding genes and
7 noncoding RNAs) that met the criteria in LUAD, LLUSC, or both (Supplementary Table S5; Table 2). Expression and methylation of representative genes in TCGA tumor (LUAD and LLUSC) or normal lung tissue samples are shown in Supplementary Fig. S1A and S1B.

Interestingly, 5 of the 7 lncRNA epi-markers were initiated from promoters overlapping repeat elements: 2 LTR elements and 3 REP522 elements (Table 2). Five of 15 protein-coding epi-markers were known cancer testis antigens (CTA; ref. 38). We also note that 13 of 22 epi-markers showed binary (OFF/ON) expression pattern: not expressed in normal (OFF) and expressed (ON) in cancer (Table 2), which is important in clinical applications for both biomarkers and therapeutic targets.

Figure 1.
Identification of upregulated/hypomethylated CAGE-defined promoters. A, Multidimensional scaling based on the CAGE tag counts of 500 most variant promoters. Red, 16 NSCLC cell lines; black, 16 normal lung epithelial cells. B, Multidimensional scaling of DNA methylation levels based on the β values of 10,000 most variant probes obtained by Infinium 450K methylation array (GSE36216). Red, 69 NSCLC cell lines; black, 6 normal lung epithelial cells. C, Study flowchart. Left, differential expression (DE) and OFF/ON analyses on CAGE profiles collected for the FANTOM5 project. Upregulated genes were selected using the thresholds of FDR < 0.05 and log2 fold change (FC) > 1. Right, differentially methylated regions (DMR) were defined as mean β value difference (Δβ) > 0.1 (hypomethylated in NSCLC) or < −0.2 (hypermethylated in NSCLC), and FDR < 0.05. Among hypomethylated and upregulated probe–promoter pairs, we further selected 59 genes that showed inverse correlation (Spearman ρ < −0.3, P < 0.05) between DNA methylation of the probe and expression of the gene in 58 NSCLC cell lines. D, Correlations between expression and methylation in 58 NSCLC cell lines. RNA-seq (E-MTAB-2706) and methylation array (GSE36216) datasets are integrated. The y-axis: log2 (1+RPKM). The x-axis: β values. The results of three coding genes (ARL14, CT83, and MYEOV), three lncRNAs (GS1-179L18.1, LINC00857, and RP11-371l1.2), and three lncRNAs whose promoters are within the REP522 repeat family (AL022344.4, PCAT7, and RPI-90G24.10) are shown.
Table 1. UP-/DOWN-regulated or turned OFF/ON promoters in NSCLC cell lines

<table>
<thead>
<tr>
<th>Category</th>
<th>UP</th>
<th>DOWN</th>
<th>OFF</th>
<th>Total (UP and DOWN)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein coding</td>
<td>6,887</td>
<td>1,397</td>
<td>4,131</td>
<td>11,018</td>
<td>64.1</td>
</tr>
<tr>
<td>Protein coding gene body</td>
<td>1,625</td>
<td>771</td>
<td>1,494</td>
<td>3,906</td>
<td>18.2</td>
</tr>
<tr>
<td>Antisense</td>
<td>298</td>
<td>110</td>
<td>72</td>
<td>506</td>
<td>2.2</td>
</tr>
<tr>
<td>Line</td>
<td>557</td>
<td>251</td>
<td>255</td>
<td>1,063</td>
<td>4.7</td>
</tr>
<tr>
<td>IncRNA</td>
<td>219</td>
<td>90</td>
<td>88</td>
<td>397</td>
<td>1.8</td>
</tr>
<tr>
<td>Pseudogene</td>
<td>285</td>
<td>99</td>
<td>67</td>
<td>451</td>
<td>2.0</td>
</tr>
<tr>
<td>Small ncRNA</td>
<td>36</td>
<td>10</td>
<td>29</td>
<td>75</td>
<td>0.4</td>
</tr>
<tr>
<td>Other</td>
<td>80</td>
<td>16</td>
<td>25</td>
<td>122</td>
<td>0.6</td>
</tr>
<tr>
<td>Not annotated</td>
<td>446</td>
<td>244</td>
<td>582</td>
<td>1,272</td>
<td>6.0</td>
</tr>
<tr>
<td>Total</td>
<td>10,433</td>
<td>2,988</td>
<td>6,743</td>
<td>17,176</td>
<td>100.0</td>
</tr>
</tbody>
</table>

NOTE: For the complete list of differentially expressed promoters see Supplementary Table S3.

Frequency of epi-marker upregulation

To investigate how often the epi-markers are upregulated in clinical NSCLC tumors, we calculated the fraction of tumors that express a given epi-marker above a normal (baseline) threshold, which we defined as the maximum expression value observed in normal lung control samples from TCGA (alternative thresholds are presented in Supplementary Table S6). The frequencies of upregulated epi-markers were 44% for IncRNAs and 40% for protein-coding genes. The frequencies were on average higher in LUSC than in LUAD (47% vs. 37%). Additionally, many epi-markers showed minimal or no expression in normal cells, while protein-coding genes demonstrated low to intermediate expression in normal samples (Table 2; Supplementary Figs. S2 and S3).

Epi-marker survival analysis

We then performed survival analysis to test whether the expression of epi-markers is associated with prognosis. Expression levels of 6 of 22 epi-markers (MYEOV, LINC00857, ARL14, CT83, IGF2BP3, and CEP55) were associated with poor prognosis in LUAD by multivariate Cox regression analysis. In LUSC, only one gene, MYEOV, was associated with poor prognosis (Table 2; Supplementary Table S7). Correlation analysis of 22 epi-markers showed that four epi-markers with prognostic impact were grouped into the same cluster of correlated genes in LUAD (MYEOV, LINC00857, ARL14, and CT83; Supplementary Fig. S4A and S4B).

Correlations with epigenetic factors

To investigate possible associations between epi-markers and known epigenetic factors, we calculated gene expression correlations between the 22 epi-markers and 683 genes from EpiFactor database (39). Intriguingly, in LUAD, the most positively correlated EpiFactors were strongly enriched for cell-cycle genes (12/14 genes, GO_CELL_CYCLE, FDR q-value = 5.6e−14) and were more specifically related to the G2–M checkpoint genes (8/14 genes, HALLMARK_G2M_CHECKPOINT, FDR q-value = 4.44e−13; Supplementary Fig. S4C).

The overlapping genes included TOP2A, which is upregulated across multiple cancer types (15, 40), and six kinases including AURKA and AURKB that are implicated in various cancer types (41). Among 22 epi-markers, IGF2BP3 and CEP55 (associated with poor survival in LUAD as described before) showed highest correlations with the cell-cycle genes (Supplementary Fig. S4C). We note however that those correlations at transcript level may not directly reflect causal relationships. No clear enrichments were observed in LUSC (Supplementary Fig. S4D).

DNA demethylation treatment

In search of genes whose expression is epigenetically silenced in normal lung epithelial cells and can be upregulated by DNA demethylation, we treated SAECs with a demethylating agent (5-aza-dc) and an HDAC inhibitor (TSA) and performed gene expression profiling (Supplementary Table S8). We observed 51 genes (45 coding genes and 6 noncoding RNAs) that were both upregulated by DNA demethylation and upregulated in NSCLC cell lines compared with normal epithelial cells (Supplementary Table S9). We found that none of them overlapped our candidate epi-markers (seven coding genes and two noncoding RNAs; Table 2).

Epigenetic activation of REP522 repetitive elements in NSCLC cell lines

We examined promoters that overlap DNA repetitive elements (RE) and have at least one annotated methylation probe and compared their expression and methylation levels between NSCLC cell lines and normal lung epithelial cells (Fig. 2A). Of particular note, the REP522 repeat family, an unclassified interspersed repeat, showed prominent hypomethylation (Fig. 2A, right) and upregulation (Fig. 2A, left) in NSCLC cell lines (Supplementary Table S10). There are 243 REP522 REs in the genome (hg19 RepeatMasker open-4.0.5) and 72 CAGE-defined promoters overlap 33 REP522 elements. Interestingly, the REP522 elements that carry promoters tend to be larger (~1.8 kb) and contain a bidirectional promoter (2 CAGE peaks, one sense, and one antisense) located in the center (Supplementary Table S11). Among 39 promoters within REP522 elements with methylation probes, 22 promoters were upregulated (14 promoters were turned ON), whereas none of them were downregulated. Analogously, 11 probes showed significant hypomethylation, whereas none of the 81 probes were hypermethylated in NSCLC cell lines (Fig. 2A; Supplementary Table S10).

To further illuminate the epigenetic regulation, we assessed histone3 modifications (data from ref. 30) within REP522 elements with or without the CAGE-defined promoters. In PC-14 LIAD cell lines, enrichment of H3K4me3, H3K27ac, and RNA polymerase II was noted in REP522 elements with promoters, but not in those without promoters, confirming transcriptional activation of promoter regions. No enrichments were observed in REP522...
<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Category</th>
<th># DM probes</th>
<th># UP promoters</th>
<th>Spearman (r) (&lt; 0.3)</th>
<th>TCGA differential expression</th>
<th>Max exp. in normal (RPKM)</th>
<th>Frequency (%) of exp. T &gt; max exp. N</th>
<th>Survival analysis RNA-Seq, TCGA LUAD Prognosis</th>
<th>Survival analysis RNA-Seq, TCGA LUSC Prognosis</th>
<th>Impact of DNA demethylating treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYEOV</td>
<td>Protein coding</td>
<td>3</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>UP</td>
</tr>
<tr>
<td>ADPRH1</td>
<td>Protein coding</td>
<td>4</td>
<td>4</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>UP</td>
</tr>
<tr>
<td>ARL14</td>
<td>Protein coding</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>UP</td>
<td>Poor</td>
<td>Poor</td>
<td>UP</td>
</tr>
<tr>
<td>KAZALD1</td>
<td>Protein coding</td>
<td>6</td>
<td>6</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>UP</td>
<td>Poor</td>
<td>Poor</td>
<td>UP</td>
</tr>
<tr>
<td>PLD1</td>
<td>Protein coding</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td></td>
<td>UP</td>
<td>Up</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>AM2</td>
<td>Protein coding</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>UP</td>
</tr>
<tr>
<td>IGF2BP3</td>
<td>Protein coding</td>
<td>3</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>UP</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>AIM2</td>
<td>Protein coding</td>
<td>9</td>
<td>9</td>
<td>Yes</td>
<td></td>
<td>Up</td>
<td>UP</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>MKRN3</td>
<td>Protein coding</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td></td>
<td>Up</td>
<td>Up</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>PON1</td>
<td>Protein coding</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td></td>
<td>Up</td>
<td>Off</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>MAGEA1</td>
<td>Cancer testis antigen</td>
<td>9</td>
<td>9</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>MAGEC2</td>
<td>Cancer testis antigen</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>PRAME</td>
<td>Cancer testis antigen</td>
<td>9</td>
<td>9</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>CEP55</td>
<td>Cancer testis antigen</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>UP</td>
<td>UP</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>CTR3</td>
<td>Cancer testis antigen</td>
<td>3</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>RP11-564D11.3</td>
<td>Pseudogene</td>
<td>1</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>LINC00857</td>
<td>linc</td>
<td>3</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>PCAT7</td>
<td>Antisense from REP522 promoter</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>RP1-90G24.10</td>
<td>Antisense from REP522 promoter</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>AL022544.4</td>
<td>linc from REP522 promoter</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>GS1-178t.18.1</td>
<td>linc from LTR/ERV1 promoter</td>
<td>3</td>
<td>3</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
<tr>
<td>RP11-371I2</td>
<td>linc from LTR/ERV1 promoter</td>
<td>2</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>ON</td>
<td>ON</td>
<td>Poor</td>
<td>Poor</td>
<td>Up</td>
</tr>
</tbody>
</table>

NOTE: Detailed information is available in Supplementary Tables S5–S9.
Abbreviations: DM, differentially methylated; T, tumor; N, normal.
elements of normal cells (SAECs), indicating inactivity of REP522 elements in normal lung epithelial cells (Fig. 2B).

**Upregulation of lncRNAs from REP522 elements in NSCLC**

The upregulated promoters that overlap REP522 elements were associated with lncRNAs, such as AL022344.4, PCAT7, and RPI-90C24.10. Interestingly, their expression levels were turned ON in NSCLC and highly cancer specific as confirmed in different RNA-seq datasets of NSCLC tissues (TCGA LUAD, TCGA LUSC, and GSE81089; Fig. 2C; Supplementary Fig. S1B and S1C).

We utilized BS-seq and RNA-seq data in 26 LUAD cell lines (30) and confirmed that DNA methylation and gene expression were negatively correlated for several REP522-associated lncRNAs, such as AL022344.4 (Fig. 2D). The expression of these lncRNAs was also associated with epigenetic signature typical of active promoters (enrichment of H3K4me3 and H3K27ac), as well as RNA polymerase II binding (Fig. 2E; Supplementary Fig. S5). Interestingly, several REP522-associated lncRNAs were upregulated in the same cell lines, such as PC-14 and VMRC-LCD cells, suggesting a possible genome-wide mechanism involved in transcriptional activation of REP522 repeat elements.

**Epigenetic regulation of MYEOV in NSCLC**

Among the upregulated/hypomethylated promoters in NSCLC cell lines, three promoters for MYEOV gene exhibited highly cancer-specific expression (Supplementary Table S5; Supplementary Fig. S6). In addition, MYEOV was the only gene that was...
prognostically relevant in both histologic subsets of lung cancer. As the function of MYEOV in NSCLC is largely unknown, we selected MYEOV for further studies.

First, we validated MYEOV expression in cultured NSCLC cell lines and lung epithelial cells by quantitative RT-PCR, which confirmed the cancer-specific expression of MYEOV (Fig. 3A). To investigate the DNA methylation status in NSCLC cell lines in more detail and resolution, we utilized BS-seq and (Fig. 3A). To investigate the DNA methylation status in NSCLC which con

Figure 3. Epigenetic regulation of MYEOV. A, Quantitative RT-PCR of MYEOV in normal lung epithelial cells (AEC, SAEC, NHBE) and NSCLC cell lines (H558, H441, H1359, H1437, H1648, H1729, H1975, A549, LC-2/ad, H270, SK-MES-1, LU1DLU-1, EBC-1, Lu99). B, Methylation and expression of MYEOV. Genomic coordinate, transcript, methylated CpG sites (BS-seq), and expression levels are shown as log(1+RPKM) in 26 LUAD cell lines. C, Histone3 modifications at the MYEOV locus in SAECs and A549 cells. Active marks (H3K4me1, H3K4me3, H3K9/14ac, and H3K27ac) are present at MYEOV locus in A549 cells. D, Quantitative RT-PCR of MYEOV in SAECs with or without 5-aza-dC (day 1, 3) and TSA (day 4) treatment. Total RNA was extracted on day 5.

Furthermore, we performed microarray analyses in A549 cells following MYEOV knockdown (Supplementary Table S12) and identified 38 commonly downregulated genes by two different siRNAs, which included tumorigenic genes such as TERT (Fig. 4D). To identify key transcription factors responsible for observed expression changes, we used ISMARA online computational tool (35). ISMARA models gene expression by integrating predicted transcription factor–binding sites, and it identifies transcription factors most likely to drive the expression changes observed in given experiment. ISMARA predicted MYC as the transcription factor that was most affected by MYEOV knockdown, and MYC activity decreased consistently in knockdown samples with both siRNAs (Fig. 4E). The biological interpretation is that genes with predicted MYC-binding sites in promoter region were downregulated after MYEOV knockdown.

To confirm the prediction, we used experimental binding sites of MYC in A549 cells (ChIP-seq data from ENCODE; ref. 36) and found that 17 of 38 downregulated genes (44.7%) have an MYC-binding site within 1 kb from the promoter and thus are potentially regulated by MYC. We marked those genes on Fig. 4D.

**MYEOV is associated with poor prognosis in NSCLC**

The univariate and multivariate Cox regression models in TCGA LIAD and LUSC datasets revealed that MYEOV is associated with poor prognosis as described above (Supplementary Table S7). We performed an extensive meta-analysis using six microarray datasets of LIAD that further substantiated the prognostic impact of MYEOV [Supplementary Fig. S7A, HR = 1.20; confidence interval (CI), 1.08–1.33; P = 0.0007]. The meta-analysis in five datasets of LUSC also confirmed MYEOV as an indicator of poor prognosis (Supplementary Fig. S7B, HR = 1.18; CI, 1.03–1.35; P = 0.0017).

We investigated the association between expression of epimarkers and mutation status of EGFR and KRAS in TCGA LIAD.
dataset (Supplementary Table S13). We found that the KRAS mutation–positive subgroup had higher MYEOV expression levels than wild-type group ($P = 0.03$), which indicates a possible link between MYEOV overexpression and KRAS mutation–associated oncogenic mechanisms.

**Discussion**

Our analyses identified 22 robust epigenetically regulated markers (epi-markers) in lung cancer. By directly integrating the CAGE and DNA methylation data, we were able to determine

---

**Figure 4.**

MYEOV knockdown in A549 and H441 LUAD cells. **A,** Cell numbers of siRNA-transfected A549 and H441 cells were counted on days 2, 4, and 6 after seeding. *, $P < 0.05$. **B,** Top, A549 and H441 cells transfected with siNC or siMYEOV were serum deprived for 48 hours to induce apoptosis. FITC-conjugated anti-Annexin V antibody and propidium iodide (PI) were detected by flow cytometry; bottom, the fraction of early apoptosis cells (Annexin V positive, PI negative). *, $P < 0.05$. **C,** Matrigel invasion assay in A549 cells. Data are representative of three independent experiments. *, $P < 0.05$. **D,** Heatmap showing the expression of significantly downregulated genes after knockdown of MYEOV with two different siRNAs (siMYEOV-1 and siMYEOV-2) in A549 cells. Genes with the cMYC-binding site within 1 kb from the promoter are marked with red. **E,** Inferred activity profile of MXI1_MYC_MYCN motif calculated by ISMARA (Integrated System for Motif Activity Response Analysis) tool (33). MXI1/MYC/MYCN was selected as the top motif by activity significance. The motif is shown in the top right corner of the figure.
precisely the methylation sites and promoters of the candidate epi-markers, whose promoter hypomethylation leads to upregulation of the associated transcript in cancer. Our unbiased, genome-wide analyses also allowed for the discovery of epigenetic activation of specific promoters located within repeat elements and identification of seven IncRNA epi-markers. Six of 15 protein-coding epi-markers (MYEOV, AIM2, ARL14, KAZALD1, MKRN3, and PRAME) are among 578 genes recently reported to show a negative correlation between expression and promoter methylation levels among NSCLC cell lines (42). Only one of our candidates (AIM2) was among 57 upregulated and hypomethylated genes in LIUADs reported by Selamat and colleagues (43). In addition, we observed no overlap between our candidates and genes reported by Mullapudi and colleagues, where 24 NSCLC tumors were compared with nontumor tissues, using a methylation-sensitive restriction enzyme–based HELP-microarray assay (44). The relatively small overlap indicates that despite previous reports, there is no consensus on epigenetically regulated genes in NSCLC. We propose that our integrative analyses based on both NSCLC cell line and large-scale tumor datasets provide a shorter and more robust list of candidates for further studies.

The frequencies of upregulation of candidate epi-markers are quite high in lung cancer tissues, and they average around 40% and are significantly higher that mutation frequencies of most common mutational drivers (KRAS 32%, EGFR 11.3%, and BRAF 7%) in the TCGA LIUAD cohort; ref. 2). The ontological link of our epi-markers with cell-cycle regulation and the significance of the associations noted with survival indicate a role in the tumorigenesis of lung cancer. However, more research is needed to elucidate the mechanisms of activation and functions of proposed epi-markers.

Five of 15 protein-coding epi-marker candidates are known CTAs, and three of them (MAGEA1, MAGEC2, and PRAME) were upregulated by demethylating treatment. This finding is consistent with previous reports showing that CTAs are epigenetically activated in multiple cancers (45) and that the expression of CTAs can be activated by a demethylating agent (46, 47). Upregulation of CTAs as immunogenic targets by a demethylating agent could be therapeutically attractive, in particular in combination with immunotherapy in NSCLC (48, 49). However, we note that three epi-markers activated by a demethylating agent (MYEOV, ARL14, and LINC00857) were also associated with poor prognosis, which can have implications when considering demethylating agents for therapeutic use in NSCLC, as oncogenic/carcinogenic genes can be activated by such treatment.

In our previous pan-cancer study, we showed for the first time the transcriptional activation of multiple copies of the REP522 interspersed repeat in cancer (15). In this study, we observed that the transcriptional activation is accompanied by parallel epigenetic changes: DNA hypomethylation and histone modifications characteristic to promoter activation. Interestingly, the activation of REP522 bidirectional promoters appears to be coordinated and occurs at multiple REP522 elements in the genome in the same cell line. With TCGA data, we were able to confirm the DNA hypomethylation and transcriptional activation (upregulation of associated transcript) in clinical tumors (both LUSC and LIUAD datasets). Our previous research indicates that REP522 activation is a pan-cancer phenomenon; thus, the epigenetic activation is likely to occur in other cancers. However, further research is needed to confirm this assumption.

We hypothesized that some of our candidate epi-markers could confer a selective growth advantage to cancer and thus act as epigenetic drivers (epi-drivers), and we selected one of our epi-markers, MYEOV, for further studies. MYEOV was initially reported as an overexpressed gene in multiple myeloma (50) and upregulated across multiple cancer types (15). In this study, we demonstrated for the first time that MYEOV is highly expressed in NSCLC cell lines and is induced by DNA demethylation in normal lung epithelial cells. Functional studies showed tumorigenic roles of MYEOV, and survival analyses revealed that MYEOV is an indicator of poor prognosis. Furthermore, MYEOV knockdown experiment followed by gene expression profiling suggested that MYEOV effects are linked to the MYC-regulatory pathway. However, further studies are needed to validate this finding and to elucidate the mechanism.

Although it is known that cancer cells take advantage of epigenetic changes by silencing tumor suppressor genes via DNA hypermethylation, our results indicate that specific (as opposed to global) hypomethylation can play a role in cancer by allowing for activation of cancer-associated genes. We believe that our selected epi-markers have a great theranostic potential in NSCLC and are a good starting point for further studies.

Disclosure of Potential Conflicts of Interest

P. Carninci has ownership interest (including patents) in RIKEN on transcriptome technologies and is a consultant/advisory board member for Dnaform, Inc. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

Conception and design: B. Kaczkowski, M. Ohshima, P. Carninci, Y. Hayashizaki, A.R.R. Forrest, P. Micke, A. Saito

Development of methodology: M. Horie, M. Itoh, P. Carninci

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M. Horie, M. Itoh, T. Lassmann, Y. Yamaguchi

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Horie, B. Kaczkowski, H. Matsuzaki, S. Noguchi, Y. Mikami, M. Liu, H. Kawaji, T. Lassmann, A. Saito

Writing, review, and/or revision of the manuscript: B. Kaczkowski, M. Ohshima, Y. Yamaguchi, P. Micke, A. Saito

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Horie, H. Kawaji, D. Takai, T. Nagase

Study supervision: M. Ohshima, P. Carninci, A.R.R. Forrest, A. Saito, T. Nagase

Other (reviewed the manuscript): P. Carninci

Acknowledgments

The authors are grateful to Kousuke Watanabe, Takahiro Kishikawa, Hirokazu Unushiya, Masahiro Yoshinari, and Dijana Djureinovic for technical support and useful discussions. We would like to thank all members of the FANTOM5 consortium for contributing to generation of samples and analysis of the data set and also thank GeNAS for data production.

Grant Support

This work was supported by JSPS KAKENHI grant #16K18437 (to M. Horie), #26461185 (to A. Saito), grant #16K11843 (to Y. Yamaguchi), grant #15K15768 (to M. Ohshima), grant #16H02653 (to T. Nagase), a grants-in-aid of the Public Trust Fund for Clinical Cancer Research (to M. Horie), and the Health and Labour Sciences Research Grant for Research on Allergic Disease and Immunology (to T. Nagase) from the Ministry of Health, Labour and Welfare, Japan. B. Kaczkowski was supported by Foreign Postdoctoral Researcher [JPR] program from RIKEN, Japan. A.R.R. Forrest was supported by a Senior Cancer Research Fellowship from the Cancer Research Trust and funds raised by the MACA Ride to Conquer Cancer. FANTOM5 was made
The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 11, 2017; revised June 12, 2017; accepted June 28, 2017; published OnlineFirst July 11, 2017.