Vol.2014-MPS-98 No.8
Vol.2014-BIO-38 No.8
2014/6/26

TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer

HIDEAKI UMEYAMA[†1] MITSUO IWADATE[†1]
Y-h. TAGUCHI[†2,†3]

Background
Non-small cell lung cancer (NSCLC) remains lethal despite the development of numerous drug therapy technologies. About 85% to 90% of lung cancers are NSCLC and the 5-year survival rate is at best still below 50%. Thus, it is important to find drug target genes for NSCLC to develop an effective therapy for NSCLC.
Results
Integrated analysis of publically available gene expression and promoter methylation patterns of two highly aggressive NSCLC cell lines generated by in vivo selection was performed. We selected eleven critical genes that may mediate metastasis using recently proposed principal component analysis based unsupervised feature extraction. The eleven selected genes were significantly related to cancer diagnosis. The tertiary protein structure of the selected genes were inferred by Full Automatic Modeling System, a profile based protein structure inference software, to determine protein functions and to specify genes that could be potential drug targets.
Conclusions
We identified eleven potentially critical genes that may mediate NSCLC metastasis using bioinformatic analysis of publically available data sets. These genes are potential target genes for therapy of NSCLC. Among the eleven genes, TINAGL1 and B3GALNT1 are possible candidates for drug compounds that inhibit their gene expression

1. Introduction

Currently, there is no effective therapy for non-small cell lung cancer (NSCLC), thus NSCLC remains lethal [1]. Five-year survival rate is at best still below 50%. In addition, NSCLC consists of several subtypes that require distinct therapies. Thus, from both a diagnosis and therapy point of view, the identification of genes critical to NSCLC is urgent. Few studies have identified NSCLC critical genes. Fawdar et al [2] recently found that mutations in FGFR4, MAO3K and PAK5 have critical roles in lung cancer progression. Li et al [3] also recently identified EML4-ALK fusion gene and EGFR and KRAS gene mutations were associated with NSCLC. Takeuchi et al [4] also reported that RET, ROS1 and ALK gene fusions were observed in lung cancer. However, it is likely that other critical gene candidates for NSCLC exist.

In this study, we attempted to identify new critical candidate genes important for NSCLC using recently proposed principal component (PCA) based unsupervised feature extraction (FE) mediated integrated analysis [5–8] of publically available promoter methylation and gene expression patterns of two NSCLC cell lines with and without enhanced metastasis ability. Most of the identified genes were previously reported as significant cancer-related genes. To understand the functionality of the selected genes, we predicted the tertiary structures of selected genes by Full Automatic Modeling System (FAMS) [9] and phyre2 [10] profile-based protein structure prediction software. This system also allowed the identification of drug target candidate genes.

†1 Department of Biological Science, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
†2 Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
†3Corresponding author

2. Results

2.1 The first principal components show no significant difference between samples

Fig. 1 shows two-dimensional embeddings of probes using PCA for gene expression and promoter methylation. To determine what each principal component (PC) represents, the contributions of samples to the first PC (PC1) are shown (Fig. 2). As previously observed [6, 8], the first PC did not identify distinct features among the samples, although they have major contributions (97% for gene expression and 87% for promoter methylation). Contributions of samples to PC1 are almost constantly independent of samples for gene expression and promoter methylation. Thus, we concluded that PC1 did not exhibit any significant differences among samples. It should be noted that this does not mean that PC1s are biologically meaningless, but rather that most gene expression and promoter methylation is sample independent; thus, the cell lines are very similar to each other independent of the ability for metastasis. This is not surprising, as they are similar NSCLC cell lines. Significantly different outcomes caused by sample dependence and/or metastasis presence would be unusual.

2.2 The second PCs demonstrate distinction between cell lines.

Because the first PCs did not distinguish between samples, we next considered second PCs (PC2s). As can be seen by two-dimensional embeddings of probes (Fig. 1), the second PCs have relatively smaller contributions. The second PC of gene expression has only a 1.7% contribution while for promoter methylation it is 9.6%. These values for contributions, especially for gene expression, can be ignored. In this case, since the samples are similar NSCLC cell lines, differences between samples are expected to be small as well. Thus, PCs with tiny contributions may represent biologically critical

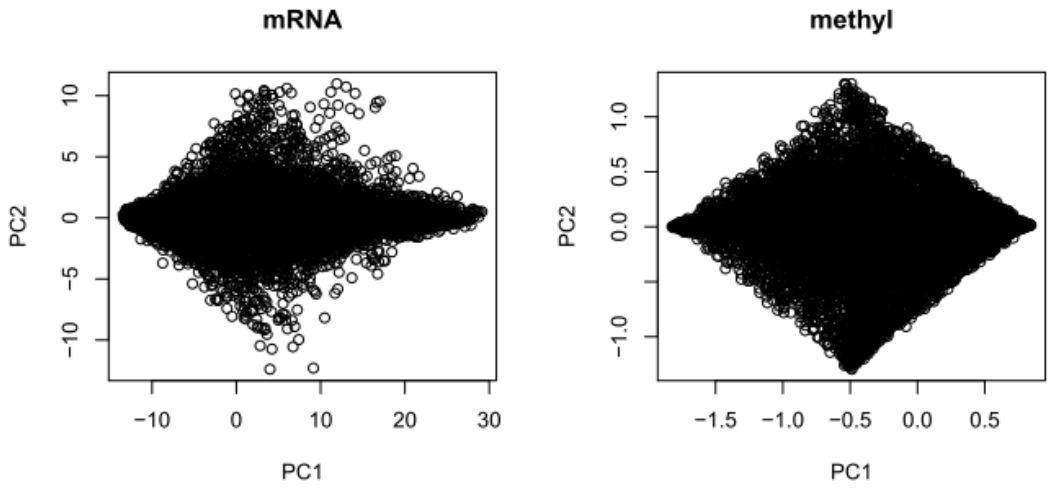
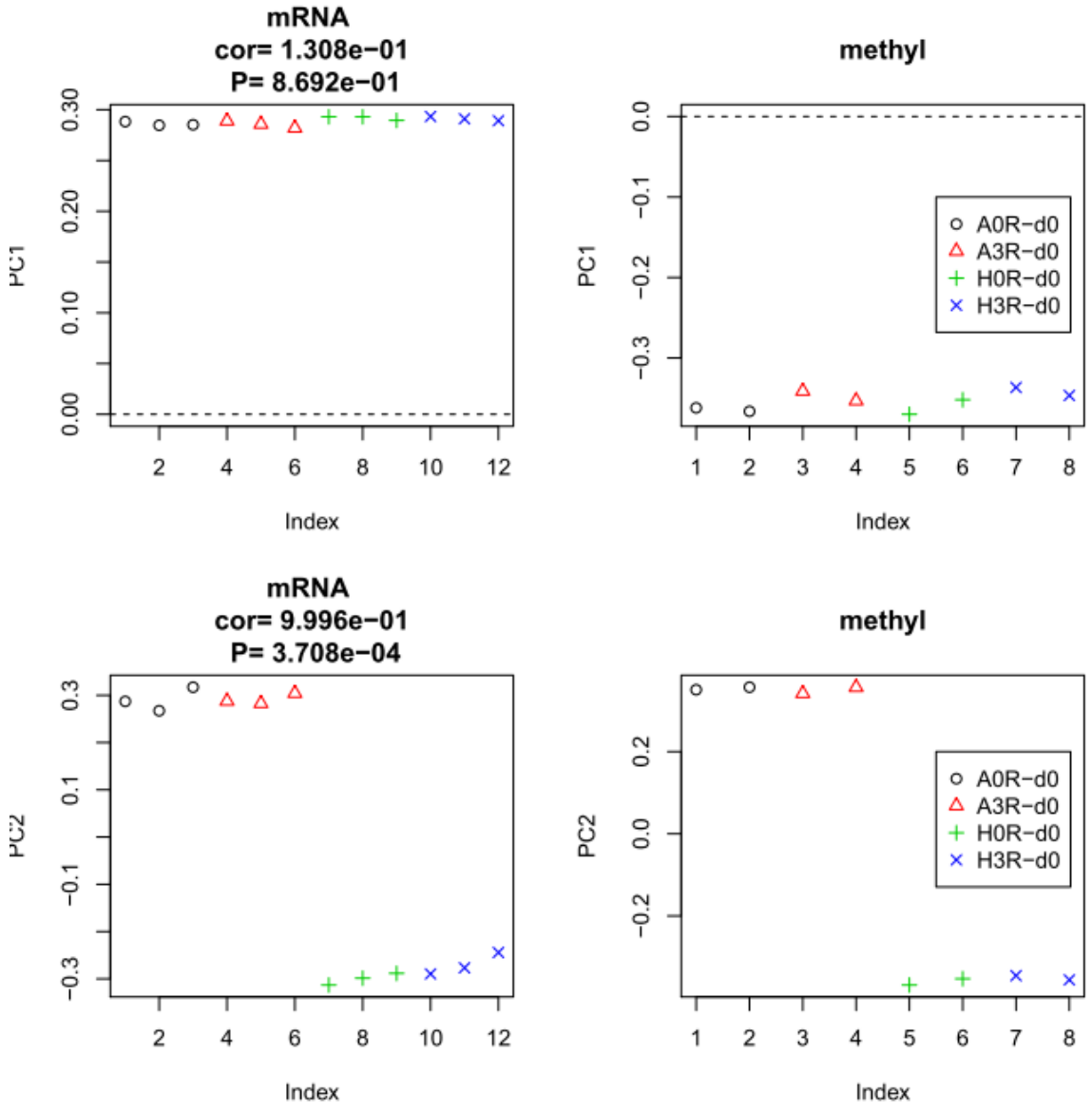


Figure 1 **Two-dimensional embeddings of probes using PCA**
Two-dimensional embeddings of probes (left: gene expression, right: promoter methylation) spanned by the first (horizontal axes) and the second (vertical) axes.



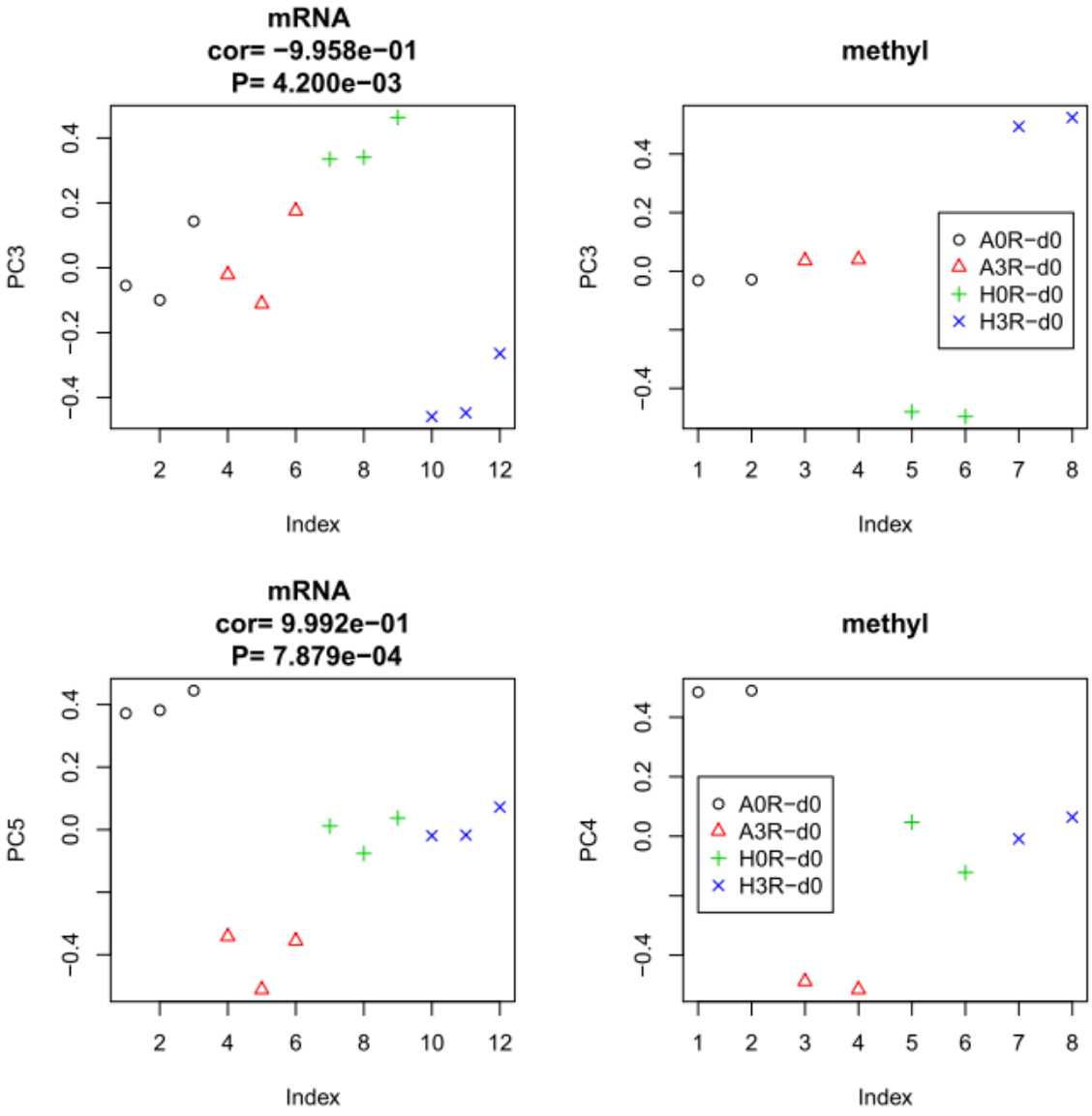


Figure 2 **Contributions of samples to PCs**
Contributions of samples (black open circles: A549 without metastasis, red triangles: A549 with metastasis, green crosses: HTB56 without metastasis, blue crosses: HTB56 with metastasis) to PCs. Left column: gene expression, right column: promoter methylation. "cor" indicates Pearson correlation coefficients of PCs between gene expression and promoter methylation averaged within each of four categories and "P" is attributed to "cor".

differences between samples, as shown in Fig. 2 where the contributions of samples to PC2s are demonstrated. It is obvious that PC2s did not distinguish between samples with and without metastasis ability, but could distinguish between A549 and HTB56 cell lines. Because we are interested in metastasis-causing genes in HSCLC, what PC2 expresses is out of scope of the present study. However, it is useful to identify genes associated with PC2 to determine which genes are different between the two cell lines, A549 and HTB56. PC2s showed good correlated between gene expression and promoter methylation. Thus, integrated analysis using PCA based unsupervised FE is applicable (Methods). P-values attributed to selected genes (Table 1) common between gene expression and promoter methylation are 4.1×10^{-9} and 5.1×10^{-12}, respectively (Methods). Thus, integrated analysis using PCA based unsupervised FE was successful. In contrast to expectations, the selected genes were frequently and significantly shown to be related to cancers by the Gendoo server [11] (Methods and Table 1). This suggests that HTB56 and A549 cell lines are potentially distinct to each other and should be considered separately. This is coincident with findings that when distinct genes are demonstrated to be present between samples with and without metastasis, they can also reflect differences between the HTB56 and A549 cell lines. Conversely, in contrast to the high correlation of PC2 for gene expression and promoter methylation, correlations between gene expression and promoter

methylation of individual genes were not significant. This might be because of too small a contribution of PC2s.

2.3 The third PCs distinguish differences between samples with and without metastasis for HTB56 but not for A549

Because no PCs reflected differences between samples with and without metastasis, we considered additional PCs. Fig. 2 shows the contributions of samples to the third PC (PC3). Although PC3s have even smaller contributions (0.2% for gene expression and 1.5% for promoter methylation) than PC1s or PC2s their correlation is high. Thus, genes associated with PC3 represent differences between samples with and without metastasis. Interestingly, PC3 exhibited differences between samples with and without metastasis only for the HTB56 cell line. However, since the two cell lines are distinct in terms of their oncogenic potential, it is not surprising that genes that exhibit differences between samples with and without metastasis for HTB56 did not exhibit differences between samples with and without metastasis for A549. Thus, we applied integrated analysis using PCA-based unsupervised FE. P-values attributed to selected genes (Table 1) common between gene expression and promoter methylation were 3.5×10^{-5} and 5.1×10^{-4} (Methods). Thus, integrated analysis using PCA based unsupervised FE was successful. The association of cancer disease and the selected genes are shown in Table 1. As expected, most of the selected genes were reported to be significantly associated with cancer disease. Correlations between gene expression and promoter methylation of individual genes were not significant.

2.4 The fourth PC of promoter methylation and the fifth PC of gene expression represent differences between samples with and without metastasis for A549 but not for HTB56

We further sought PCs that exhibited differences between samples with and without metastasis for A549. The fourth PC (PC4) of promoter methylation and the fifth PC (PC5) of gene expression demonstrated differences between samples with and without metastasis for the A549 cell line (Figs. 2 and 3). Because the correlation between these PC4 and PC5 were very high despite small contributions (0.6% for PC4 of promoter methylation and 0.09% for PC5 of gene expression), integrated analysis using PCA based unsupervised FE could still be used. P-values attributed to selected genes (Table 1) common between gene expression and promoter methylation were 9.8×10^{-8} (Methods). Thus, integrated analysis using PCA based unsupervised FE was successful. Cancer diseases associated with selected genes are listed in Table 1 and more than half were reported to be associated with cancer-related diseases. However, correlations between gene expression and promoter methylation of individual genes were not significant.

Table 1 Cancer disease association with genes selected in the present study based on Gendoo server.

Gene Symbol	Refseq mRNA	Cancer associations (P-value)
PC2 vs PC2		
SLC22A3	NM_021977	Gonadoblastoma (0.0002), Dysgerminoma (0.00075), Testicular Neoplasms (0.00456), Ovarian Neoplasms (0.0297), Cell Transformation, Neoplastic (0.0384)
DFNA5	NM_004403	Melanoma (0.006),
SPG20	NM_015087	Hepatoblastoma (0.0033), Liver Neoplasms (0.00496)
CYP1B1	NM_000104	Breast Neoplasms (1.13×10^{-45}), Endometrial Neoplasms (2.44×10^{-12}), Lung Neoplasms (1.56×10^{-9}), Prostatic Neoplasms (4.65e-9), Adenocarcinoma (6.03×10^{-6}), Ovarian Neoplasms (1.35×10^{-5}) Carcinoma, Squamous Cell (0.00018), Colorectal Neoplasms (0.000337), Head and Neck Neoplasms (0.00052), Adenoma, Liver Cell (0.0072), Urinary Bladder Neoplasms (0.012), Neoplasms (0.019), Carcinoma, Small Cell (0.028), Carcinoma, Non-Small-Cell Lung (0.0326)
ALX1	NM_006982	Carcinoma (0.000305), Chondrosarcoma (0.00129), Bone Neoplasms (0.0106), Uterine Cervical Neoplasms (0.011)
TFPI2	NM006528	Uterine Neoplasms (2.6×10^{-21}), Neoplasm Invasiveness (1.18×10^{-14}), Choriocarcinoma (2.33×10^{-13}), Fibrosarcoma (7.98×10^{-9}), Glioma (2.50×10^{-8}), Cystadenocarcinoma (1.68×10^{-5}), Lung Neoplasms (6.74×10^{-5}), Carcinoma, Non-Small-Cell Lung (0.00559)
HOXA9	NM_152739	Leukemia, Myeloid (2.0×10^{-48}), Leukemia, Myeloid, Acute

		(9.24×10^{-30}), Cell Transformation, Neoplastic (4.64×10^{-29}), Leukemia (9.46×10^{-19}), Leukemia, Myelogenous, Chronic, BCR-ABL Positive (2.64×10^{-14}), Precursor Cell Lymphoblastic Leukemia-Lymphoma (2.46×10^{-8}), Precursor B-Cell Lymphoblastic Leukemia-Lymphoma (1.65×10^{-6}), Myoma (0.00046), Leukemia, T-Cell (0.0012), Endodermal Sinus Tumor (0.0079), Seminoma (0.0157),
HOXA11	NM_005523	Uterine Neoplasms (8.23×10^{-7}), Choriocarcinoma (3.97×10^{-5}), Carcinoma, Endometrioid (0.0065), Adenocarcinoma, Clear Cell (0.00662), Wilms Tumor (0.0076),
PCSK1	NM000439	Bronchial Neoplasms (0.0022), Adenoma (0.0030), Adenoma, Islet Cell (0.0035), Bile Duct Neoplasms (0.011)
SPARC	NM_003118	Neoplasm Invasiveness (8.42×10^{-14}), Glioma (1.35×10^{-8}), Brain Neoplasms (1.01×10^{-7}), Melanoma (2.99×10^{-7}), Lung Neoplasms (1.43×10^{-5}), Carcinoma (0.00013), Carcinoma, Non-Small-Cell Lung (0.0009)
PC3 vs PC3		
HOXB2	NM_002145	Lung Neoplasms (0.000159), Leukemia, Myeloid (0.000326), Pulmonary Emphysema (0.00139), Carcinoma, Embryonal (0.0025), Adenocarcinoma (0.0054), Leukemia, Erythroblastic, Acute (0.0096), Leukemia, Promyelocytic, Acute (0.0124), Carcinoma, Small Cell (0.0148), Carcinoma, Non-Small-Cell Lung (0.0387)
CCDC8	NM_032040	
ZNF114	NM_153608	
DIO2	NM_000793	Choriocarcinoma (0.000616), Carcinoma, Papillary (0.00366), Hemangioma (0.0099), Adenoma (0.019), Neuroblastoma (0.025)
LAPTM5	NM_006762	Carcinoma, Hepatocellular (0.000396), Liver Neoplasms (0.000495), Multiple Myeloma (0.00947), Neoplasm Recurrence (0.010), Cell Transformation, Neoplastic (0.032)
RGS1	NM_002922	Burkitt Lymphoma (3.55×10^{-5}), Lymphoma, B-Cell (9.14×10^{-5}), Leukemia-Lymphoma, Adult T-Cell (0.0076), Lymphatic Metastasis (0.0329), Skin Neoplasms (0.0364), Stomach Neoplasms (0.0454), Melanoma (0.0455)
B3GALNT1	NM_003781	Neuroblastoma (0.0034)
PC5 vs PC4		
TINAGL1	NM_022164	Carcinoma, Hepatocellular (0.000119), Neoplasms (0.0295)
PMEPA1	NM_020182	Prostatic Neoplasms (2.30e-12), Carcinoma, Renal Cell (0.0233), Kidney Neoplasms (0.032)
CX3CL1	NM_002996	Neuroblastoma (0.0014)
ICAM1	NM_000201	Melanoma (0.00305), astrocytoma (0.00644), Granular Cell Tumor (0.0166), Colonic Neoplasms (0.0233), Lymphoma, AIDS-Related (0.023), Adenoma, Oxyphilic (0.0433)

2.5 Conclusions

This study performed the integrated analysis of promoter methylation and gene expression using PCA based unsupervised FE. It selected 11 genes that were differently expressed and which had different promoter methylation patterns between cell lines with and without metastasis ability. *P*-values attributed to the simultaneous selection between gene expression and

promoter methylation were significant and many cancer-related diseases were associated with the 11 genes selected. Two of selected eleven genes, *B3GALNT1* and *TINAGL1*, were identified as drug target candidates that might suppress metastasis in NSCLC. Further detailed and advanced studies are required to confirm these findings.

3. Methods

3.1 Promoter methylation and gene expression profiles

Promoter methylation profiles were downloaded from Gene Expression Omnibus (GEO) with GEO ID: GSE52144 that included two replicates of HTB56 cell lines with (H3R_d0) and without (H0R_d0) metastasis ability and A549 cell lines with (A3R_d0) and without (A0R_d0) metastasis ability. Gene expression profiles were downloaded from GEO with GEO ID: GSE52143 that included three replicates of the samples in GSE52144. For these two cell lines, data sets deposited in the "Series Matrix Files" were retrieved. Promoter methylation measured by sequencing was obtained from GEO with GEO ID: GSE52140. Within GSE52140_RAW.tar, eight files corresponding to those in GSE52144, (two replicates of H0R_d0. H3R_d0, A0R_d0 and A3R_d3) were used.

3.2 Integrated analysis of gene expression and promoter methylation using PCA based unsupervised FE

First, PCA was applied to gene expression and promoter methylation and each probe was embedded into a two dimensional space spanned with the first and the second PC scores. Then contributions of each probe to each PC were investigated and biologically meaningful PCs were selected. The 100 top outlier probes with larger (positively larger) or smaller (negatively larger) PC scores were extracted for each PC. The coincidence between selected probes for gene expression and promoter methylation were estimated as follows. If contributions of each probe to PCs were positively correlated between gene expression and promoter methylation, then intersections between gene expression outlier probes having larger (smaller) PC scores and promoter methylation outlier probes having smaller (larger) PC scores were sought, since gene expression and promoter methylation were expected to be negatively correlated with each other. Conversely, if contributions of each probe to PCs were negatively correlated between gene expression and promoter methylation, intersections between gene expression outlier probes having larger (smaller) PC scores and promoter methylation outlier probes having larger (smaller) PC scores were sought. P-values attributed to simultaneous selection of probes between gene expression and promoter methylation were computed by distribution that obeyed binomial distribution as follows:

$$1\text{-}P(x,100,100/y)$$

where x is the number of commonly selected probes between the top 100 outliers of gene expression and promoter methylation, y is total number of probes on the microarray, and P is the cumulative frequency of binomial distribution.

Reference

1) Goldstraw P, Ball D, Jett JR, Le Chevalier T, Lim E, Nicholson AG, Shepherd FA: Non-small-cell lung cancer. Lancet 2011, 378:1727-1740
2) Fawdar S, Trotter EW, Li Y, Stephenson NL, Hanke F, Marusiak AA, Edwards ZC, Ientile S, Waszkowycz B, Miller CJ, Brognard J: Targeted genetic dependency screen facilitates identification of actionable mutations in FGFR4, MAP3K9, and PAK5 in lung cancer. Proc Natl Acad Sci USA 2013, 110:12426-12431.
3) Li Y, Li Y, Yang T, Wei S, Wang J, Wang M, Wang Y, Zhou Q, Liu H, Chen J: Clinical significance of EML4-ALK fusion gene and association with EGFR and KRAS gene mutations in 208 Chinese patients with non-small cell lung cancer. PLoS One 2013, 8:e52093.
4) Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, Asaka R, Hamanaka W, Ninomiya H, Uehara H, Lim Choi Y, Satoh Y, Okumura S, Nakagawa K, Mano H, Ishikawa Y: RET, ROS1 and ALK fusions in lung cancer. Nat Med 2012, 18:378-381.
5) Murakami Y, Toyoda H, Tanahashi T, Tanaka J, Kumada T, Yoshioka Y, Kosaka N, Ochiya T, Taguchi YH: Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease. PLoS One 2012, 7:e48366.
6) Ishida S, Umeyama H, Iwadate M, Taguchi YH: Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery. Protein Pept Lett 2013, in press.
7) Taguchi YH, Murakami Y: Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers. PLoS One 2013, 8:e66714.
8) Kinoshita R, Iwadate M, Umeyama H, Taguchi YH: Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets. BMC Syst Biol 2014, 8:S4.
9) Umeyama H, Iwadate M: FAMS and FAMSBASE for protein structure. Curr Protoc Bioinformatics 2004, 4:5.2.1-5.2.16.
10) Kelley LA, Sternberg MJ: Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009, 4:363-371.
11) Nakazato T, Bono H, Matsuda H, Takagi T: Gendoo: functional profiling of gene and disease features using MeSH vocabulary. Nucleic Acids Res 2009, 37:W166-169.

Acknowledgments This study was supported by KAKENHI 23300357 and 26120528 and Chuo University Joint Research Grant.