Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2090286, 10 pages
http://dx.doi.org/10.1155/2016/2090286

Research Article

Uncovering Driver DNA Methylation Events in Nonsmoking Early Stage Lung Adenocarcinoma

Xindong Zhang,[1] Lin Gao,[1] Zhi-Ping Liu,[2] Songwei Jia,[1] and Luonan Chen[3,4,5]

[1] *School of Computer Science and Technology, Xidian University, Xi'an 710000, China*
[2] *Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Shandong 250061, China*
[3] *Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[4] *Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan*
[5] *School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China*

Correspondence should be addressed to Lin Gao; lgao@mail.xidian.edu.cn

Received 28 May 2016; Revised 28 June 2016; Accepted 5 July 2016

Academic Editor: Quan Zou

As smoking rates decrease, proportionally more cases with lung adenocarcinoma occur in never-smokers, while aberrant DNA methylation has been suggested to contribute to the tumorigenesis of lung adenocarcinoma. It is extremely difficult to distinguish which genes play key roles in tumorigenic processes via DNA methylation-mediated gene silencing from a large number of differentially methylated genes. By integrating gene expression and DNA methylation data, a pipeline combined with the differential network analysis is designed to uncover driver methylation genes and responsive modules, which demonstrate distinctive expressions and network topology in tumors with aberrant DNA methylation. Totally, 135 genes are recognized as candidate driver genes in early stage lung adenocarcinoma and top ranked 30 genes are recognized as driver methylation genes. Functional annotation and the differential network analysis indicate the roles of identified driver genes in tumorigenesis, while literature study reveals significant correlations of the top 30 genes with early stage lung adenocarcinoma in never-smokers. The analysis pipeline can also be employed in identification of driver epigenetic events for other cancers characterized by matched gene expression data and DNA methylation data.

1. Introduction

As a leading cause of death worldwide, lung cancer is mainly attributed to smoking in both men and women [1, 2], of which the most common histological subtype is adenocarcinoma. However, as smoking rates decrease, proportionally more cases occur in never-smokers [3]. Lung adenocarcinoma in never-smokers shows obvious distinctions in clinical and molecular mechanism to those cigarette smoking [4]. Both genetics and epigenetics in cancer genomes have been suggested to account for the development of lung adenocarcinoma.

As one of the vital epigenetic mechanisms, DNA methylation regulates gene expression without alterations in DNA sequence [5, 6] and plays key roles in X chromosome inactivation, genome stability, chromatin structure, embryonic development, differentiation, and maintenance of pluripotency in normal somatic cells [7, 8]. Genome-scale methylation-profiling techniques have confirmed the existence of widespread aberrations of DNA methylation patterns in human cancer genome [9–12]. Studies of DNA methylation have suggested that both global DNA hypomethylation and gene-specific hypermethylation may contribute to the initiation and progression of tumorigenesis, as well as gene body methylation [13–15]. It is challenging but of great significance to distinguish genes whose methylation changes are crucial in cancer occurrence, progression, or metastasis from genes whose methylation changes merely have effects on the process of tumorigenesis in cancer research and therapy [13]. Unlike somatic mutations in the genome, DNA methylation is inherently reversible and serves as potential drug targets in cancer intervention [16, 17].

Numerous studies have focused on discovering genes whose DNA methylation potentially plays key roles in tumorigenesis of lung adenocarcinoma, including integration of genome-scale DNA methylation and gene expression [18–21]. The main idea of these works is to search genes whose gene expression fluctuations are highly correlated to DNA methylation changes. However, there is a deficiency derived on the complexity of the gene expression regulation. Both genetic and epigenetic alterations can contribute to gene expression as well as other transcriptional factors in sophisticated manners in complex diseases [22, 23]. In tumors, a differential gene expression may be induced by an aberrant DNA methylation in the promoter of the gene but also may be a consequence regulated by its upstream genes in regulatory mechanisms. These appeal to a great attention in uncovering driver DNA methylations, which play major roles in methylation-associated gene silencing and drive malignant transformation [5, 13]. In this work, we refine the generalized description of driver methylation as two properties. (1) Driver DNA methylation should induce distinctive expressions in tumors with differential DNA methylation (T-DM) when compared to expressions in matched adjacent nontumor (normal) and tumors with nondifferential DNA methylation (T-NDM), and (2) driver methylation should induce a distinct regulation module in the network perspective. The first property guarantees the major role of DNA methylation in the regulation of gene expression, while the second property guarantees the functional effects of driver genes on tumorigenesis.

Focusing on genes differentially expressed among matched adjacent nontumors (normal), tumors with aberrant DNA methylation (T-DM), and tumors without aberrant DNA methylation (T-NDM), we integrate genome-wide DNA methylation data and gene expression data to uncover driver methylation events in never-smokers in early stage lung adenocarcinoma. Differential network analyses show significant changes of DNA methylation-responsive modules in network topology across normal, T-DM, and T-NDM, which imply potential mechanisms of identified driver genes underlying the tumorigenesis.

2. Materials and Methods

2.1. Data Sets. Both the DNA methylation data and gene expression data are downloaded from NCBI Gene Expression Omnibus (GEO) with accession number GSE32867 [18]. The series contains 59 samples with paired genome-scale DNA methylation profiling and gene expression. Stage I and stage II are merged as early stage and stages III-IV are labeled with late stage [18]. After removing noisy data [18], 22 samples are labeled with "never smoking" and "early stage" simultaneously. Paired DNA methylation data and gene expression data of these 22 samples are collected to further analysis. Probes in gene expression data are firstly mapped to Entrez gene ID and expression values sharing same Entrez gene IDs are averaged among samples.

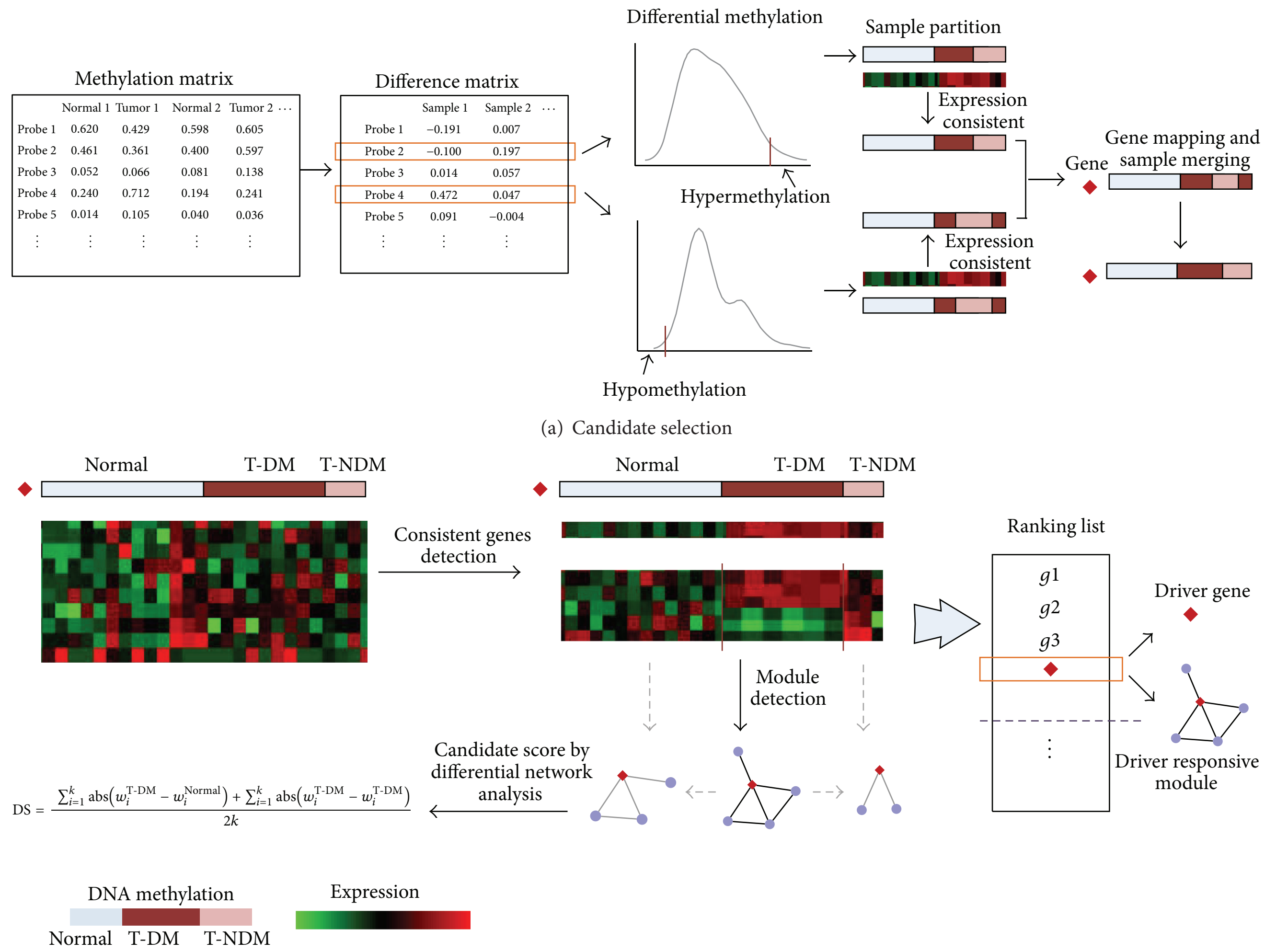
2.2. Schematic Overview of the Analysis Pipeline. The schematic overview of the analysis pipeline is shown in Figure 1, and detailed procedures are described in the following sections.

2.2.1. Candidate Driver Gene Selection. Figure 1(a) shows a brief schematic overview of this procedure. The difference matrix is firstly created to measure differences of beta values of DNA methylation between tumor and normal. The kernel probability distribution with normal smoothing function is used to estimate the probability density distribution for each probe in the difference matrix (Figure 1(a)). The hypothesis is that the differences of beta values for given probes come from distributions with the mean 0 and unknown variances. The cumulative density function (CDF) is used to estimate the probability of a beta value falling within given interval. Hypermethylation and hypomethylation are determined by the upper bound CDF > 0.95 and the lower bound CDF < 0.05, respectively. For each probe, tumors are partitioned into two groups, tumors with differential methylation group (T-DM) and tumors without differential methylation group (T-NDM).

Then, the two-sample t-test is used to evaluate differential expression under conditions [24], and p values are adjusted by the procedure introduced by Storey [25]. The mapping from DNA methylation to gene expression is performed by shared Entrez gene ID. Probes remain if the mapped genes are differentially expressed in T-DM when compared to normal and T-NDM (adjusted p value < 0.05), which implies that the differential methylation of given probes in T-DM is more likely to induce significant expression changes. Probes mapping to same genes are removed if hypermethylation and hypomethylation coexist in more than 5 samples. Then samples in T-DMs and T-NDMs merge, respectively, by shared Entrez gene ID and serve as T-DM and T-NDM of the gene.

We then search for genes whose expressions are highly discriminative and consistent in T-DM when compared to normal and T-NDM. Many types of statistics, such as Wilcoxon score, Pearson correlation coefficient (PCC), or mutual information (MI), could be used to score the relationship between gene expression and class labels, and a T-score method is used in this work [26]. For a given gene, let a be the gene expression levels across samples with class c and the discriminative score $s(a, c)$ is defined as the t-test statistic. To determine whether the discriminative level of the gene among groups is consistent, we permute the class c by 1000 times and obtain a background distribution of the discriminative scores $S'(a, c)$ derived on the gene expression levels a and permuted class c'. Genes with significant values (p value < 0.05) among groups (normal versus T-DM and T-DM versus T-NDM) are considered differentially methylated and served as candidates for further analysis.

2.2.2. Detection of DNA Methylation-Responsive Module. To construct the DNA methylation-responsive module for a


Methylation matrix
Normal 1 Tumor 1 Normal 2 Tumor 2 ⋯
Probe 1 0.620 0.429 0.598 0.605
Probe 2 0.461 0.361 0.400 0.597
Probe 3 0.052 0.066 0.081 0.138
Probe 4 0.240 0.712 0.194 0.241
Probe 5 0.014 0.105 0.040 0.036
Difference matrix
Sample 1 Sample 2 ⋯
Probe 1 −0.191 0.007
Probe 2 −0.100 0.197
Probe 3 0.014 0.057
Probe 4 0.472 0.047
Probe 5 0.091 −0.004
Differential methylation
Hypermethylation
Hypomethylation
Sample partition
Expression consistent
Expression consistent
Gene
Gene mapping and sample merging
Normal
T-DM
T-NDM
Consistent genes detection
Ranking list
g1
g2
g3
Driver gene
Driver responsive module
Module detection
Candidate score by differential network analysis
$DS = \frac{\sum_{i=1}^{k} \text{abs}(w_i^{\text{T-DM}} - w_i^{\text{Normal}}) + \sum_{i=1}^{k} \text{abs}(w_i^{\text{T-DM}} - w_i^{\text{T-DM}})}{2k}$
DNA methylation
Normal T-DM T-NDM
Expression

(a) Candidate selection

(b) Driver prioritization

FIGURE 1: Schematic overview of the pipeline proposed in this work. (a) Candidate gene selection. Methylation matrix of continuous beta values is converted into difference matrix and discretized by kernel distribution function, which partition samples into normal, T-DM, and T-NDM. Probes are mapped to genes after noise filtering and genes passing the consistent test are collected as candidate driver genes. (b) For each candidate gene, a subset of DM responsive genes is collected and DM responsive modules are constructed by the CLR method. Candidate driver genes are ranked by differential scores derived on the differential network analysis.

candidate gene g, we firstly recognize a set of genes whose expressions are highly discriminative among groups defined by DNA methylation profiles of g. These genes are potentially responsive to aberrant DNA methylation of g.

The Context Likelihood of Relatedness (CLR) method [27] is used to assess regulatory relationships among these genes. CLR estimates MI for each pair of variables and corrects the MI via a background-corrected procedure. In particular, for mutual information $I(X_i; X_j)$, CLR scores the relatedness between a pair of variables X_i and X_j by the joint likelihood measurement:

$$z_{ij} = \sqrt{z_i^2 + z_j^2}, \qquad (1)$$

where

$$z_i = \max\left(0, \frac{I(X_i; X_j) - \mu_i}{\sigma_i}\right), \qquad (2)$$

where μ_i and σ_i are the mean and standard deviation derived on the empirical distribution of MI between X_i and arbitrary variables X_k ($k = 1, 2, \ldots, n$) and $I(X_i; X_j)$ is the mutual information of X_i and X_j.

CLR employs B-spline smoothing and discretization method [28] to estimate the MI for a pair of variables. However, it is time-consuming in this work under diversiform conditions and permutations. Thus, we use the following estimation method to calculate MI for pair of variables X_i and X_j [29]; that is,

$$I(X_i; X_j) = -\frac{1}{2}\log(1 - \rho^2), \qquad (3)$$

where ρ is the PCC of X_i and X_j.

An experienced threshold δ is necessary when CLR is employed. A larger threshold results in a higher precision but a smaller size of responsive modules. The size of more than 70% modules is less than three when $\delta = 4.46$, while the size of 80% modules is larger than 3 when $\delta = 4.46$ and approximate ranking lists of top 30 genes are obtained when δ falls in the interval between 3.96 and 5.46. Thus, we set $\delta = 4.46$ in this work.

2.2.3. Scoring Candidate Driver Genes by Differential Network Analysis. Differential network analysis reveals dynamic changes of pathways and potential mechanisms in complex diseases including cancers [30]. For each candidate gene, we calculate CLR scores for edges in responsive modules under normal and T-NDM. Differential scores are calculated to estimate network differences among groups. The differential score (DS) is yielded by the following equation:

$$\mathrm{DS} = \frac{\sum_{i=1}^{k} \mathrm{abs}\left(w_i^{\text{T-DM}} - w_i^{\text{Normal}}\right) + \sum_{i=1}^{k} \mathrm{abs}\left(w_i^{\text{T-DM}} - w_i^{\text{T-TDM}}\right)}{2k}, \quad (4)$$

where w_i is the CLR score of the ith edge and k is the number of edges in driver methylation-responsive module. Then candidate genes are prioritized by DS scores in descending order.

3. Results

We focus on the detection of differentially methylated genes which play key roles in tumorigenesis ("driver methylation gene") and modules responsive to aberrant methylation of these genes. Rather than genes with consistent expressions to DNA methylation levels in whole tumors, we detect genes differentially expressed and consistent with DNA methylation in T-DM when compared to normal and T-NDM.

3.1. Identification of Candidate Driver Genes in Tumorigenesis. By integrating DNA methylation and corresponding gene expression data, the samples are partitioned into three groups (normal, T-DM, and T-NDM) for each gene (Figure 1(a)). Firstly, we remove genes that are not differentially expressed in T-DM when compared to normal and T-NDM. Then a permutation test is performed to determine the significance of the consistency of gene expression changes in T-DM when compared to T-NDM. To obtain a significant level of differences, we randomly permute T-DM and T-NDM and calculate differences. After 1000 times permutation, a background distribution of differences is constructed. After removing genes with the absolute mean beta value less than 0.1, 135 genes remain in the candidate list (see Supplementary File in Supplementary Material available online at http://dx.doi.org/10.1155/2016/2090286). We perform a functional enrichment analysis using DAVID [31, 32]. Of these 135 genes, 115 are annotated to GO terms including cancer-related functions such as response to stimulus, development process, cell differentiation, cell adhesion, cell growth and cell death, DNA repair, and apoptosis, which imply potential relationships between cancers and these 135 genes.

3.2. Detection Responsive Modules of Candidate Driver Genes. Biological network reveals cell's functional organization [33]. To characterize the functional implications of candidate driver genes in tumorigenesis, we detect modules responsive to differential methylation of candidate driver genes (Section 2). Totally, 130 of 135 modules have at least one edge when the threshold of CLR is set to 4.46, and the mean size of 130 modules is 15.

3.3. Prioritization of Candidate Driver Genes by Differential Network Analysis. We argue that a driver DNA methylation can induce not only a distinctive gene expression in T-DM, but also a distinctive module responsive to the alteration. We score each candidate driver gene by analysis of the differential level of the responsive module. Candidate driver genes are ranked by differential scores in descending order.

We testify the significance of the differential score to a background distribution derived from random permutations. For a given candidate driver gene, genes are randomly selected from its possible responsive genes with module size maintained, and a new module is constructed by CLR with $\delta = 4.46$ as well as a differential score. A sequence of DS' consisting of random differential scores is obtained after 1000 times random permutation. Of 135 candidate driver genes, 130 genes pass the test with p value < 0.01.

We also perform a differential network analysis of responsive modules under different CLR thresholds from 1.96 to 6.96 with step 0.5. Almost all modules obtain significant differential scores under CLR cutoffs (Supplementary File). Table 1 lists details of top 30 genes.

4. Discussion

We build two lists as background to testify the accuracy of the ranked list. The first consists of genes that show absolute mean fold change larger than 0.2 in T-DM and literature annotated in lung cancer. Totally, 29 genes are contained in the first list and denoted as Standard_Lit. The other one comes from Selamat et al. of 76 genes [18]. In fact, this list is not very suitable because genes in Selamat et al. are confused with differentially methylated genes under smoking and late stage. Thus, we select genes covered by list from Selamat et al. and our list. Totally 19 genes are in the list and denote as Standard_Sel. Genes in these two lists are listed in Supplementary File.

We test the accuracy of our list to Standard_Lit and Standard_Sel; Figure 2(a) shows the ROC curves with AUC = 0.686 and AUC = 0.628, respectively, which means that over half of genes in two standard lists are high-ranked in our list. Figure 2(b) shows the overlaps of the top 30 genes in our list to Standard-Lit and Standard-Sel. For Standard-Lit, 12 of 29 genes are overlapped (Fisher exact test p value = 0.0018), while for Standard-Lit, 10 of 29 genes are overlapped (Fisher exact test p value = $2.67E - 04$).

The ranked list is also validated by literature annotation. Of the top 30 genes, 27 genes are previously reported to be cancer-relevant, while 17 of them are lung cancer or non-small-cell lung cancer-related (Table 1).

Table 1: Top 30 genes ranked by differential score in lung adenocarcinoma.

Gene symbol[a]	Differential score	Number of samples in T-DM group[b]	p value
FAM107A [34]	16.301	20	$7.80E - 06$
SPARCL1 [35, 36]	14.920	20	$1.40E - 07$
TRPC6 [37]	14.649	11	$<1.0E - 10$
CRYAB [38]	14.508	12	$3.84E - 10$
WFDC3	14.483	−14	$<1.0E - 10$
EFEMP2 [39]	13.958	20	$<1.0E - 10$
MX2 [40, 41]	13.895	−18	$2.12E - 05$
PLA2G4C [42]	13.870	−8	$<1.0E - 10$
ST6GALNAC5 [43]	13.848	9	$<1.0E - 10$
PLAT [44]	13.690	8	$2.45E - 04$
TCF21 [45]	13.664	22	$<1.0E - 10$
SOX17 [46]	13.368	22	$<1.0E - 10$
SH3GL2 [47]	13.300	5	$<1.0E - 10$
MAMDC2 [18]	13.274	19	$4.54E - 07$
GCNT3 [48]	13.238	−14	$<1.0E - 10$
MSR1 [49]	13.144	−16	$<1.0E - 10$
PPP1R14D [50]	13.057	−12	$<1.0E - 10$
COL5A2 [51]	13.045	19	$6.67E - 04$
PTPRH [52]	12.967	−16	$8.98E - 13$
HKDC1 [53]	12.961	−20	$<1.0E - 10$
CDH13 [54]	12.932	−20	$3.34E - 04$
CFI [55]	12.932	5	$1.20E - 04$
ARL14	12.880	−12	$2.06E - 04$
MMP9 [56]	12.866	7	$<1.0E - 10$
CELSR3	12.856	16	$4.65E - 10$
CDO1 [57]	12.846	22	$<1.0E - 10$
AGR2 [58]	12.836	−22	$<1.0E - 10$
S100P [59, 60]	12.828	−10	$2.29E - 04$
DOCK2 [61]	12.777	20	$2.54E - 03$
TNFRSF1B [62]	12.736	13	$<1.0E - 10$

[a] Bold: gene literature annotated to lung cancer.

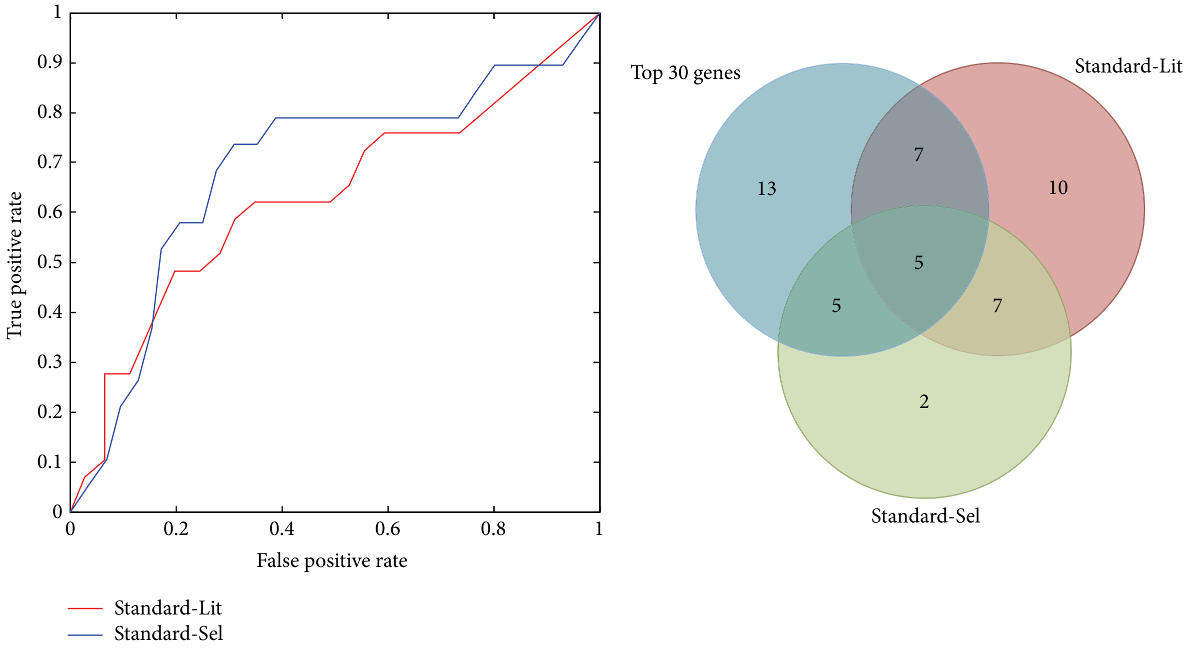
[b] −: Gene hypomethylated in samples.

Table 2: Functional annotation of driver-responsive network to KEGG signaling pathways (p value < 0.01).

Gene symbol	Enriched KEGG signaling pathway	p value
SPARCL1	CYTOSOLIC_DNA_SENSING	$3.22E - 03$
TRPC6	PPAR_SIGNALING	$9.50E - 03$
	P53_SIGNALING	$9.50E - 03$
	MTOR_SIGNALING	$7.16E - 03$
	NOTCH_SIGNALING	$6.47E - 03$
EFEMP2	NOTCH_SIGNALING	$9.70E - 03$
MX2	RIG_I_LIKE_RECEPTOR_SIGNALING	$9.33E - 04$
PLA2G4C	PPAR_SIGNALING	$9.50E - 03$
	P53_SIGNALING	$9.50E - 03$
	MTOR_SIGNALING	$7.16E - 03$
	NOTCH_SIGNALING	$6.47E - 03$
ST6GALNAC5	PPAR_SIGNALING	$9.50E - 03$
	P53_SIGNALING	$9.50E - 03$
	MTOR_SIGNALING	$7.16E - 03$
	NOTCH_SIGNALING	$6.47E - 03$
PLAT	TOLL_LIKE_RECEPTOR_SIGNALING	$3.09E - 03$
	NOD_LIKE_RECEPTOR_SIGNALING	$1.16E - 03$
	CYTOSOLIC_DNA_SENSING	$9.44E - 04$
	JAK_STAT_SIGNALING	$6.99E - 03$
TCF21	FC_EPSILON_RI_SIGNALING	$4.89E - 03$
GCNT3	NOTCH_SIGNALING	$9.70E - 03$
MSR1	NOTCH_SIGNALING	$9.70E - 03$
PTPRH	B_CELL_RECEPTOR_SIGNALING	$9.80E - 03$
HKDC1	PPAR_SIGNALING	$9.50E - 03$
	P53_SIGNALING	$9.50E - 03$
	MTOR_SIGNALING	$7.16E - 03$
	NOTCH_SIGNALING	$6.47E - 03$
CDH13	ERBB_SIGNALING	$1.55E - 03$
	T_CELL_RECEPTOR_SIGNALING	$2.38E - 03$
CFI	PPAR_SIGNALING	$2.90E - 03$
	MAPK_SIGNALING	$3.47E - 03$
ARL14	VEGF_SIGNALING	$3.93E - 03$
S100P	HEDGEHOG_SIGNALING	$6.47E - 04$
	TGF_BETA_SIGNALING	$1.51E - 03$
DOCK2	CHEMOKINE_SIGNALING	$3.49E - 05$
	TOLL_LIKE_RECEPTOR_SIGNALING	$4.85E - 03$
	NOD_LIKE_RECEPTOR_SIGNALING	$3.27E - 05$
	T_CELL_RECEPTOR_SIGNALING	$5.42E - 03$
	B_CELL_RECEPTOR_SIGNALING	$2.66E - 03$
TNFRSF1B	NOTCH_SIGNALING	$7.06E - 03$
	FC_EPSILON_RI_SIGNALING	$1.24E - 03$

We also annotate responsive modules of top 30 ranked genes to KEGG signaling pathways. Among them, responsive modules for 18 genes are enriched with KEGG signaling pathways with significance level p value < 0.01, which imply significant relations of these responsive modules to cancer processes (Table 2) and indicate potential mechanism changes induced by aberrant DNA methylation. The KEGG signaling pathways are collected from MsigDB [63, 64].

Of 30 top ranked genes, *FAM107A*, *MAMDC2*, *SOX17*, *TCF21*, *PTPRH*, and *CDO1* have been previously reported with aberrant DNA methylation in lung cancer [18, 34, 45, 46, 52, 57]. All these genes obtain higher occurrences ($n > 19$) in lung adenocarcinoma. *AGR2*, *CDH13*, *CRYAB*, *MX2*, *SH100P*, and *SH3GL2* are reported with aberrant gene expression [38, 40, 47, 54, 58, 59], while *AGR2*, *CDH13*, and *MX2* are of high occurrences in aberrant DNA methylation ($n \geq 18$). Differential expression of these genes has been reported playing crucial roles in key pathways in tumorigenesis or serving as potential prognostic targets. With higher occurrences, the



(a) (b)

Figure 2: Comparison of the ranked list to two standard sets denoted by Standard-Lit and Standard-Sel. (a) ROC curves of our ranked list compared to Standard-Lit and Standard-Sel with AUC equal to 0.686 and 0.628, respectively. (b) Venn diagram showing the overlap of top 30 ranked genes in our list to Standard-Lit and Standard-Sel.

correlation of differential gene expression and aberrant DNA methylation of *AGR2*, *CDH13*, and *MX2* have been reported relevant to lung adenocarcinoma [18].

Alpha B-crystallin (*CRYAB*) is one of the important members of the small heat-shock protein family with aberrant DNA methylation occurring in 12 of 22 samples. The upregulated expression of *CRYAB* is reported relevant to the poor survival of patients with non-small-cell lung cancer (NSCLC) [38]. Interestingly, we find a contrary expression pattern in early stage lung adenocarcinoma in nonsmoking patients (Figure 3). A decreased expression is observed in both T-DM (p value = $8.20E - 11$) and T-NDM (p value = $7.72E - 8$) when compared to normal, while a relatively weak difference is also observed between T-DM group and T-NDM group (mean fold change difference = 0.07, p value = 0.15), which implies multiple mechanisms in regulation of *CRYAB*, as well as DNA hypermethylation. The responsive module of *CRYAB* is highly changed in normal and T-NDM (DS = 14.508, p value = $3.84E - 10$). The similar case is *SH3GL2*, deletion of which downregulates tumor growth by modulating *EGFR* signaling [47].

Another interesting case is *S100P*, which has been reported as a key gene in tumor progression in both initial stage and advanced stage in lung adenocarcinoma [60]. The gene shows distinctive expressions among normal, T-DM, and T-NDM. There are nearly no changes existent in gene expression between normal and T-NDM, while in T-DM, upregulation is observed, which implies that the upregulation of *S100P* may be an important step in the early stage of lung adenocarcinomas.

Also some genes are relevant to cancers but lung cancer from literature study (*COL5A2* [51], *SPARCL1* [35], *EFEMP2* [39], *MSR1* [49], and *DOCK2* [61]). *APARCL1* and *DOCK2* have shown downregulation in types of cancer [36, 61], while both of them show downregulated gene expressions in T-DM with high occurrences of DNA hypermethylation. Similar to *CRYAB*, *EFEMP2* shows contrary expression patterns in our observation compared to which in gliomas [39]. *EFEMP2* has high occurrences of DNA hypermethylation and downregulated gene expression in totally 20 samples, while 2 samples in T-NDM show little differences when compared to matched normal. *COL5A2* also shows T-DM specific upregulation of gene expression and DNA hypermethylation with high occurrences.

We show the responsive module of *MSR1* in Figure 4(a) as a representation of responsive modules of cancer-related genes. All these genes exhibit significant changes in responsive modules in T-DM when compared to normal and T-NDM.

Besides cancer-related genes, three genes *ARL14*, *CELSR3*, and *WFDC3* are also observed in our list. These three genes show T-DM specific expression changes (Figure 3), and regulatory correlations in responsive modules show significant differences in T-DM when compared to normal and T-NDM (Figures 4(b)–4(d)) which also imply potential roles of the three genes in the tumorigenesis of lung adenocarcinoma.

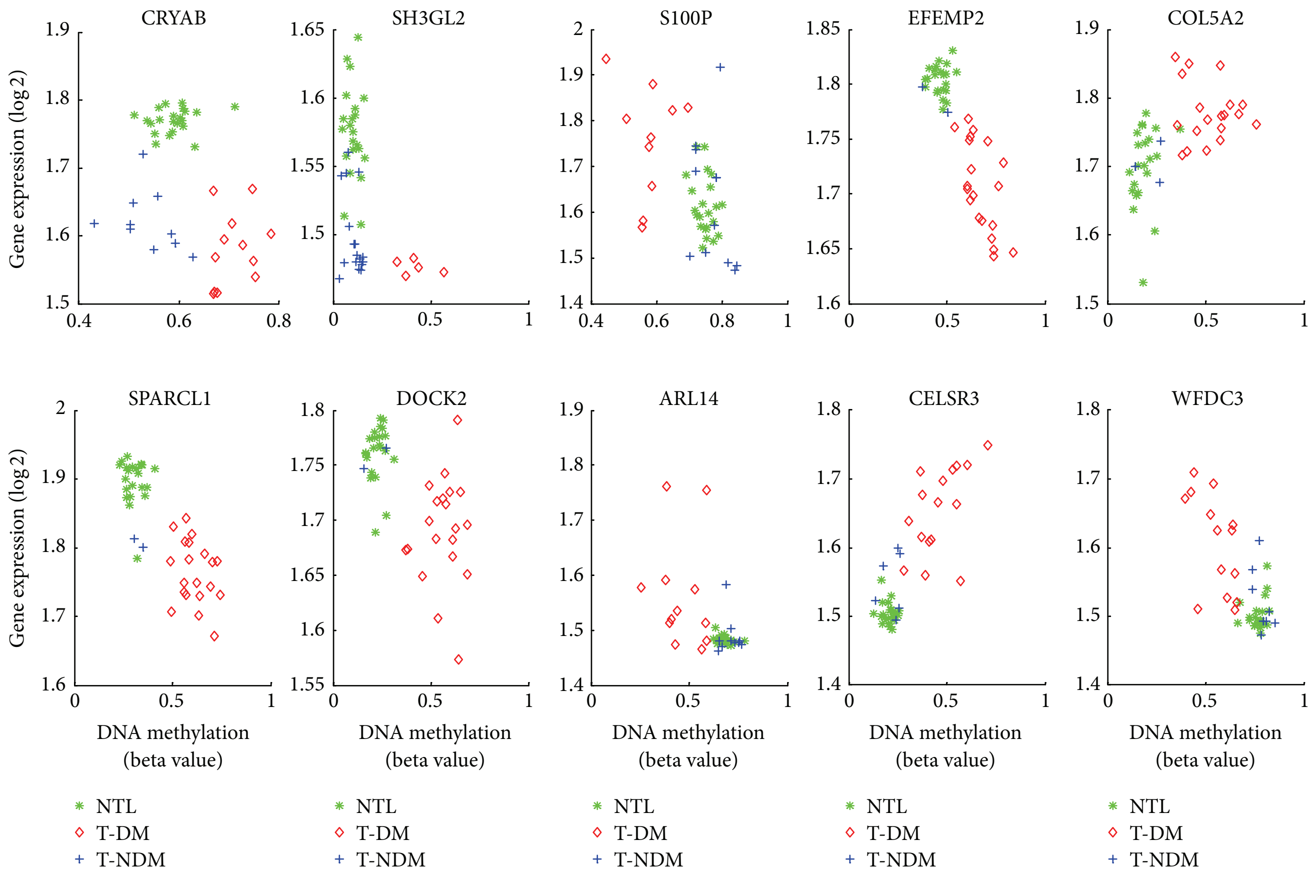


FIGURE 3: Genes show consistently significant changes in gene expression and DNA methylation in T-DM (red diamond) when compared to normal (green star) and T-NDM (blue plus). Results indicate different distributions of gene expression with altered DNA methylation in three groups of top ranked genes.

All top 30 genes show significant changes in responsive modules in T-DM, while detailed information of the top 30 genes and responsive modules are listed in Supplementary File.

5. Conclusions

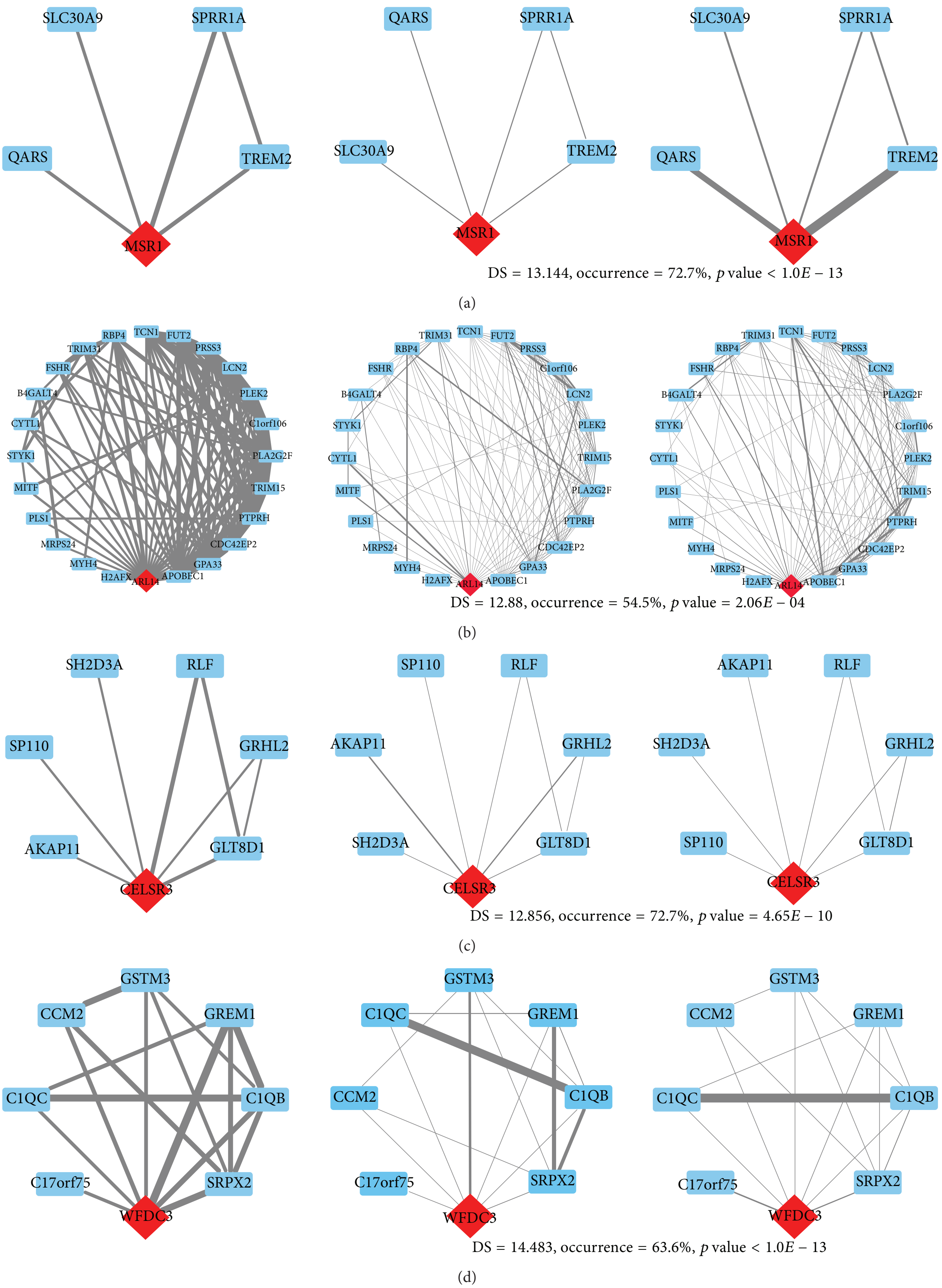
By integration of gene expression and DNA methylation data, we analyzed 22 matched lung adenocarcinoma/nontumor lung pairs for nonsmokers in early stage lung adenocarcinoma. By focusing on differences in gene expression patterns and responsive modules derived from T-DM compared to those in normal and T-NDM, we proposed a pipeline by employing a differential network analysis strategy. Totally, 135 candidate genes are analyzed, and top 30 genes are well studied in this work. All 135 genes are differentially expressed in T-DM when compared to matched normal and T-NDM, while 130 of them show significant changes in regulatory correlations of responsive modules. Literature mining of top 30 genes indicates a high proportion of lung cancer-relevant genes, which implies potential risks of these genes to disturb functions and pathways via differential methylation mechanisms, and further drives the tumorigenesis of lung adenocarcinoma in early stage. In conclusion, we provide a bioinformatics pipeline to identify driver genes with aberrant DNA methylation by fully considering differential expression and network changes in T-DM, normal, and T-NDM. The analysis pipeline can also be employed in identification of driver genes with aberrant DNA methylation of other cancers characterized by paired gene expression and DNA methylation.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grants nos. 61532014, 61572287, 61432010, and 61402349), the Fundamental Research Funds for the



DS = 13.144, occurrence = 72.7%, p value $< 1.0E-13$

(a)

DS = 12.88, occurrence = 54.5%, p value $= 2.06E-04$

(b)

DS = 12.856, occurrence = 72.7%, p value $= 4.65E-10$

(c)

DS = 14.483, occurrence = 63.6%, p value $< 1.0E-13$

(d)

Figure 4: Differential representation of responsive modules for *MSR1*, *ARL14*, *CELSR3*, and *WFDC3* in T-DM (left), normal (middle), and T-NDM (right). Significant changes of responsive modules for identified driver genes (red diamond) imply functional alterations of driver genes in tumorigenesis.

Central Universities (Grant no. BDZ021404), and Shandong Provincial Natural Science Foundation of China under Grant no. ZR2015FQ001.

References

[1] H. Zhang and B. Cai, "The impact of tobacco on lung health in China," *Respirology*, vol. 8, no. 1, pp. 17–21, 2003.

[2] J. Ferlay, P. Autier, M. Boniol, M. Heanue, M. Colombet, and P. Boyle, "Estimates of the cancer incidence and mortality in Europe in 2006," *Annals of Oncology*, vol. 18, no. 3, pp. 581–592, 2007.

[3] C.-K. Toh, F. Gao, W.-T. Lim et al., "Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity," *Journal of Clinical Oncology*, vol. 24, no. 15, pp. 2245–2251, 2006.

[4] Y. Hu and G. Chen, "Pathogenic mechanisms of lung adenocarcinoma in smokers and non-smokers determined by gene expression interrogation," *Oncology Letters*, vol. 10, no. 3, pp. 1350–1370, 2015.

[5] D. De Carvalho, S. Sharma, J. S. You et al., "DNA methylation screening identifies driver epigenetic events of cancer cell survival," *Cancer Cell*, vol. 21, no. 5, pp. 655–667, 2012.

[6] Y. Delpu, P. Cordelier, W. C. Cho, and J. Torrisani, "DNA methylation and cancer diagnosis," *International Journal of Molecular Sciences*, vol. 14, no. 7, pp. 15029–15058, 2013.

[7] D. D. De Carvalho, J. S. You, and P. A. Jones, "DNA methylation and cellular reprogramming," *Trends in Cell Biology*, vol. 20, no. 10, pp. 609–617, 2010.

[8] A. Meissner, "Epigenetic modifications in pluripotent and differentiated cells," *Nature Biotechnology*, vol. 28, no. 10, pp. 1079–1088, 2010.

[9] R. L. Momparler and V. Bovenzi, "DNA methylation and cancer," *Journal of Cellular Physiology*, vol. 183, no. 2, pp. 145–154, 2000.

[10] M. Kulis and M. Esteller, "DNA methylation and cancer," in *Advances in Genetics*, H. Zdenko and U. Toshikazu, Eds., pp. 27–56, Academic Press, 2010.

[11] N. The Cancer Genome Atlas Research, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, 2012.

[12] I. Balgkouranidou, T. Liloglou, and E. S. Lianidou, "Lung cancer epigenetics: emerging biomarkers," *Biomarkers in Medicine*, vol. 7, no. 1, pp. 49–58, 2013.

[13] S. Kalari and G. P. Pfeifer, "Identification of driver and passenger DNA methylation in cancer by epigenomic analysis," *Advances in Genetics*, vol. 70, pp. 277–308, 2010.

[14] D. Jjingo, A. B. Conley, S. V. Yi, V. V. Lunyak, and I. King Jordan, "On the presence and role of human gene-body DNA methylation," *Oncotarget*, vol. 3, no. 4, pp. 462–474, 2012.

[15] X. Yang, H. Han, D. D. DeCarvalho, F. D. Lay, P. A. Jones, and G. Liang, "Gene body methylation can alter gene expression and is a therapeutic target in cancer," *Cancer Cell*, vol. 26, no. 4, pp. 577–590, 2014.

[16] J. M. Teodoridis, G. Strathdee, and R. Brown, "Epigenetic silencing mediated by CpG island methylation: potential as a therapeutic target and as a biomarker," *Drug Resistance Updates*, vol. 7, no. 4-5, pp. 267–278, 2004.

[17] L. Sigalotti, E. Fratta, S. Coral et al., "Epigenetic drugs as pleiotropic agents in cancer treatment: biomolecular aspects and clinical applications," *Journal of Cellular Physiology*, vol. 212, no. 2, pp. 330–344, 2007.

[18] S. A. Selamat, B. S. Chung, L. Girard et al., "Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression," *Genome Research*, vol. 22, no. 7, pp. 1197–1211, 2012.

[19] M. Tessema, C. M. Yingling, Y. Liu et al., "Genome-wide unmasking of epigenetically silenced genes in lung adenocarcinoma from smokers and never smokers," *Carcinogenesis*, vol. 35, no. 6, pp. 1248–1257, 2014.

[20] A. Karlsson, M. Jönsson, M. Lauss et al., "Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome," *Clinical Cancer Research*, vol. 20, no. 23, pp. 6127–6140, 2014.

[21] T. Sato, E. Arai, T. Kohno et al., "Epigenetic clustering of lung adenocarcinomas based on DNA methylation profiles in adjacent lung tissue: its correlation with smoking history and chronic obstructive pulmonary disease," *International Journal of Cancer*, vol. 135, no. 2, pp. 319–334, 2014.

[22] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.

[23] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: a survey," *Briefings in Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2016.

[24] W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18, no. 1, pp. S96–S104, 2002.

[25] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 64, no. 3, pp. 479–498, 2002.

[26] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, Article ID e1000217, 2008.

[27] J. J. Faith, B. Hayete, J. T. Thaden et al., "Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, article e8, 2007.

[28] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska, "Estimating mutual information using B-spline functions—an improved similarity measure for analysing gene expression data," *BMC Bioinformatics*, vol. 5, no. 1, article 118, pp. 1–12, 2004.

[29] C. Olsen, P. E. Meyer, and G. Bontempi, "On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information," *Eurasip Journal on Bioinformatics and Systems Biology*, vol. 2009, Article ID 308959, 2009.

[30] T. Ideker and N. J. Krogan, "Differential network biology," *Molecular Systems Biology*, vol. 8, article 565, 2012.

[31] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[32] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[33] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[34] D. Pastuszak-Lewandoska, K. H. Czarnecka, M. Migdalska-Sęk et al., "Decreased FAM107A expression in patients with non-small cell lung cancer," *Advances in Experimental Medicine and Biology*, vol. 852, pp. 39–48, 2015.

[35] Y. Xiang, Q. Qiu, M. Jiang et al., "SPARCL1 suppresses metastasis in prostate cancer," *Molecular Oncology*, vol. 7, no. 6, pp. 1019–1030, 2013.

[36] P. Li, J. Qian, G. Yu et al., "Down-regulated SPARCL1 is associated with clinical significance in human gastric cancer," *Journal of Surgical Oncology*, vol. 105, no. 1, pp. 31–37, 2012.

[37] A. F. Pla and D. Gkika, "Emerging role of TRP channels in cell migration: from tumor vascularization to metastasis," *Frontiers in Physiology*, vol. 4, article 311, 2013.

[38] H. Qin, Y. Ni, J. Tong et al., "Elevated expression of CRYAB predicts unfavorable prognosis in non-small cell lung cancer," *Medical Oncology*, vol. 31, no. 8, article 142, 2014.

[39] L. Wang, Q. Chen, Z. Chen et al., "EFEMP2 is upregulated in gliomas and promotes glioma cell proliferation and invasion," *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 9, pp. 10385–10393, 2015.

[40] M. Watanabe, N. Komeshima, S. Nakajima, and T. Tsuruo, "MX2, a morpholino anthracycline, as a new antitumor agent against drug-sensitive and multidrug-resistant human and murine tumor cells," *Cancer Research*, vol. 48, no. 23, pp. 6653–6657, 1988.

[41] K. Kobayashi, M. Nishioka, T. Kohno et al., "Identification of genes whose expression is upregulated in lung adenocarcinoma cells in comparison with type II alveolar cells and bronchiolar epithelial cells in vivo," *Oncogene*, vol. 23, no. 17, pp. 3089–3096, 2004.

[42] C. Hartmann, L. Johnk, H. Sasaki, R. B. Jenkins, and D. N. Louis, "Novel PLA2G4C polymorphism as a molecular diagnostic assay for 19q loss in human gliomas," *Brain Pathology*, vol. 12, no. 2, pp. 178–182, 2002.

[43] P. D. Bos, X. H.-F. Zhang, C. Nadal et al., "Genes that mediate breast cancer metastasis to the brain," *Nature*, vol. 459, no. 7249, pp. 1005–1009, 2009.

[44] G. Buccheri and D. Ferrigno, "Lung tumour markers in oncology practice: a study of TPA and CA125," *British Journal of Cancer*, vol. 87, no. 10, pp. 1112–1118, 2002.

[45] L. T. Smith, M. Lin, R. M. Brena et al., "Epigenetic regulation of the tumor suppressor gene TCF21 on 6q23-q24 in lung and head and neck cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 4, pp. 982–987, 2006.

[46] D. Yin, Y. Jia, Y. Yu et al., "SOX17 methylation inhibits its antagonism of Wnt signaling pathway in lung cancer," *Discovery Medicine*, vol. 14, no. 74, pp. 33–40, 2012.

[47] S. Dasgupta, J. S. Jang, C. Shao et al., "SH3GL2 is frequently deleted in non-small cell lung cancer and downregulates tumor growth by modulating EGFR signaling," *Journal of Molecular Medicine (Berlin, Germany)*, vol. 91, no. 3, pp. 381–393, 2013.

[48] N. E. Reticker-Flynn and S. N. Bhatia, "Aberrant glycosylation promotes lung cancer metastasis through adhesion to galectins in the metastatic niche," *Cancer Discovery*, vol. 5, no. 2, pp. 168–181, 2015.

[49] Y. Chen, C. Sullivan, C. Peng et al., "A tumor suppressor function of the *Msr1* gene in leukemia stem cells of chronic myeloid leukemia," *Blood*, vol. 118, no. 2, pp. 390–400, 2011.

[50] K. Lokk, T. Vooder, R. Kolde et al., "Methylation markers of early-stage non-small cell lung cancer," *PLoS ONE*, vol. 7, no. 6, article e39813, 2012.

[51] H. Fischer, R. Stenling, C. Rubio, and A. Lindblom, "Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2," *Carcinogenesis*, vol. 22, no. 6, pp. 875–878, 2001.

[52] T. Sato, K. Soejima, E. R. I. Arai et al., "Prognostic implication of PTPRH hypomethylation in non-small cell lung cancer," *Oncology Reports*, vol. 34, no. 3, pp. 1137–1145, 2015.

[53] G.-H. Li and J.-F. Huang, "Inferring therapeutic targets from heterogeneous data: HKDC1 is a novel potential therapeutic target for cancer," *Bioinformatics*, vol. 30, no. 6, pp. 748–752, 2014.

[54] K. O. Toyooka, S. Toyooka, A. K. Virmani et al., "Loss of expression and aberrant methylation of the CDH13 (H-cadherin) gene in breast and lung carcinomas," *Cancer Research*, vol. 61, no. 11, pp. 4556–4560, 2001.

[55] M. Okroj, Y.-F. Hsu, D. Ajona, R. Pio, and A. M. Blom, "Non-small cell lung cancer cells produce a functional set of complement factor I and its soluble cofactors," *Molecular Immunology*, vol. 45, no. 1, pp. 169–179, 2008.

[56] D. Schveigert, S. Cicenas, S. Bruzas, N. Samalavicius, Z. Gudleviciene, and J. Didziapetriene, "The value of MMP-9 for breast and non-small cell lung cancer patients' survival," *Advances in Medical Sciences*, vol. 58, no. 1, pp. 73–82, 2013.

[57] J. Wrangle, E. O. Machida, L. Danilova et al., "Functional identification of cancer-specific methylation of CDO1, HOXA9, and TAC1 for the diagnosis of lung cancer," *Clinical Cancer Research*, vol. 20, no. 7, pp. 1856–1864, 2014.

[58] M. Alavi, V. Mah, E. L. Maresh et al., "High expression of AGR2 in lung cancer is predictive of poor survival," *BMC Cancer*, vol. 15, no. 1, article 655, 2015.

[59] B. Bartling, G. Rehbein, W. D. Schmitt, H.-S. Hofmann, R.-E. Silber, and A. Simm, "S100A2-S100P expression profile and diagnosis of non-small cell lung carcinoma: impairment by advanced tumour stages and neoadjuvant chemotherapy," *European Journal of Cancer*, vol. 43, no. 13, pp. 1935–1943, 2007.

[60] G. Rehbein, A. Simm, H.-S. Hofmann, R.-E. Silbar, and B. Bartling, "Molecular regulation of S100P in human lung adenocarcinomas," *International Journal of Molecular Medicine*, vol. 22, no. 1, pp. 69–77, 2008.

[61] H. Nishihara, M. Maeda, A. Oda et al., "DOCK2 associates with CrkL and regulates Rac1 in human leukemia cell lines," *Blood*, vol. 100, no. 12, pp. 3968–3974, 2002.

[62] X. Guan, Z. Liao, H. Ma et al., "TNFRSF1B +676 T>G polymorphism predicts survival of non-Small cell lung cancer patients treated with chemoradiotherapy," *BMC Cancer*, vol. 11, article 447, 2011.

[63] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. 1, pp. D457–D462, 2016.

[64] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.