DNA Methylation Diagnosis Markers Detection on Lung Adenocarcinoma with Deep Learning

Xinyang Yuan (yuanxinyang@stu.scu.edu.cn)
Sichuan University

Research Article

Keywords:

Posted Date: May 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1653074/v1

DNA Methylation Diagnosis Markers Detection on Lung Adenocarcinoma with Deep Learning

Xinyang Yuan
School of Aeronautics and Astronautics, Sichuan University, Chengdu, China
Correspondence should be addressed to Xinyang Yuan
; yuanxinyang@stu.scu.edu.cn

Abstract: Lung Adenocarcinoma (LUAD) is one of the most fatal malignant tumors which has led to millions of deaths yearly all around the world. Arouse from new topic of selecting DNA methylation markers in cancer therapy diagnosis, we explore relationship between gene expression and DNA methylation, we applied machine learning unsupervised algorithms in narrowing the range of candidate DNA methylation probes and create a deep-learning-based diagnosis model which has high accuracy in predicting cells into malignant and benign.

1. Introduction

Lung cancer is currently one of the most lethal oncological conditions worldwide, with a five-year survival rate of less than 20%, of which 13% are diagnosed as small cell carcinoma (SCLC) and 83% as non-small cell carcinoma (NSCLC), of which the major subtypes are lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) [1]. In a recent study, Herbst proposed a full gene action network for lung cancer targeting and summarized immunotherapies for lung cancer [2], and for one of these, lung adenocarcinoma (LUAD), Kim et, al [3] explored the relevance of transcriptional repression of CPS1 by LKB1 via AMPK and reversal of CPS1 expression in human NSCLC patients in response to LKB1. Hewelt et al. [4] used Python to explore transcription-related gene pathways in lung cancer.

Simo-Riudalbas et al.[5] screened for methylation sites associated with facial anomaly syndrome (ICF) and performed an analysis on the role of methylation levels in relation to gene expression and RNA regulation. Were analyzed, while Teschendorff et al. established the BMIQ model to optimize the normalization model of methylation data and initiated the study of methylation markers [6, 7]. Current studies on methylation profiles focus on cancer targeting studies and the relationship between differentially methylated regions (DMR), which are closely related to the prognostic outcome of tumors, and are widely used in the diagnosis of various cancer subtypes as well as the determination and prediction of prognostic outcome. xu et al.[8] and Qiu et al.[9] screened liver cancer (HCC) by LASSO regression with cox hazard Ishihara et al.[10] identified breast cancer cell fraction markers by methylation markers, Tang et al. combined DNA methylation with miRNA to predict primary signatures of tumors[11], Wang et al. used deep learning combined with high-throughput sequencing (Hi-C) based 3D chromosome topology for the epigenetic study of leukemia [12], and Xue et al. [13] identified markers for colon cancer (COAD) through the study of hypermethylated regions (UMR) and hypomethylation regions (LMR) to better understand the molecular mechanisms as well as the prognostic role of methylation markers. In contrast, the study of methylation profiles has played a major role in the

study of lung adenocarcinoma. He et al. [14] explored the relationship between methylation abnormalities and lung adenocarcinoma prognosis through the relationship between abnormal methylation sites and gene expression, Song et al. [15] closely correlated PITX1 overexpression through hypermethylation expression, and Yang et al. [16] combined methylation profiles, gene expression and mRNA relationships to find diagnostic loci for LUAD.

In our experiments, we explore methylation diagnostic marker screening for lung adenocarcinoma from a gene expression perspective. We screened the gene expression data for differentially expressed genes (DEGs), statistically screened the methylation data, and further screened by unsupervised learning. This was followed by screening for markers of LUAD methylation by comparing methylation profiles with normal samples using LASSO regression and modelling the diagnosis by deep learning (Figure 1). Finally, the screened candidate methylation sites were further analyzed in relation to CGIs and the relationship between methylation and gene expression.

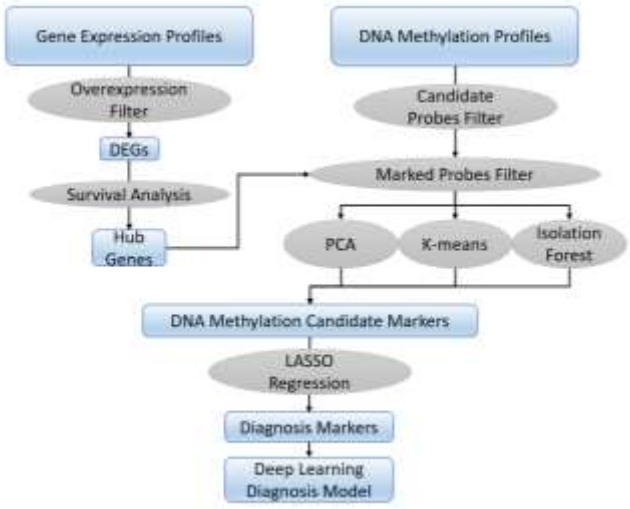

Figure 1 Study flowchart.

2. Results

Identification of DEGs

Three gene expression profiles (GSE43767, GSE68465, GSE75037) were selected for this study. Of these, GSE43767 contained 67 LUAD samples and 9 normal lung samples at 16 to 24 weeks, GSE68465 contained 443 LUAD samples and 19 normal lung samples, and GSE75037 contained 83 LUAD samples and 83 normal lung samples, respectively. Based on the criteria of $P<0.05$ and $|logFC|>=1$, a total of 5177 Candidate DEGs were identified in GSE43767, including 2731 up-regulated genes and 2440 down-regulated genes. 3222 Candidate DEGs were identified in GSE68465, including 1623 up-regulated genes and 1599 down-regulated genes. GSE75037 identified 3425 Candidate DEGs, including 1860 up-regulated genes and 1565 down-regulated genes. All Candidate DEGs were identified by comparing LUAD samples with normal lung

samples. Venn analysis (http://bioinformatics.psb.ugent.be/webtools/Venn/) was then performed to obtain intersections of the DEG profiles. Finally, 154 genes were significantly abnormally expressed in the intersection of the three groups, of which 68 were significantly up-regulated and 86 were down-regulated (Figure 2).

Functional enrichment analyses of DEGs

To gain further insight into the pathway disorders of lung adenocarcinogenesis, GO functional and KEGG pathway enrichment analyses were performed on DEGs containing different methylation sites using the online tool DAVID

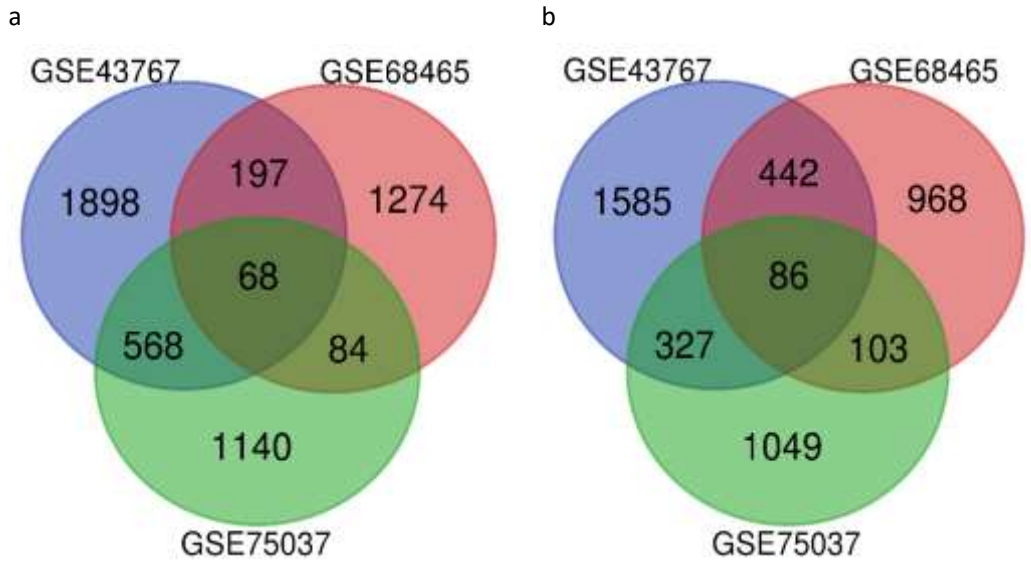

Figure 2 Detection of DEGs via Venn Plot (a) Up-regulated DEGs **(b)** Down-regulated DEGs

(https://david.ncifcrf.gov/). Results of GO analysis for "Biological Processes" (BP), "Cellular Composition" (CC) and "Molecular Function" (MF) were shown (Figure 3(a)), using The FDR P-value cut-off of 0.1 resulted in significant enrichment in 8, 7 and 5 GO terms, respectively. For BP,DEGs are enriched in RNA polymerase||negative regulation of promoter transcription, negative regulation of cell proliferation, nervous system development, positive regulation of cell proliferation, positive regulation of cell expression, negative regulation of transcription-DNA templating, negative regulation of apoptotic processes and oxygen transport. (oxygen transport, nervous system development, negative regulation of cell proliferation, negative regulation of transcription from RNA polymerase II promoter, positive regulation of gene expression, positive regulation of cell proliferation, negative regulation of apoptotic process, negative regulation of transcription, DNA-templated)

CC enriched in cytoplasm, cell membrane, plasma membrane components, Golgi apparatus, perinuclear region of cytoplasm, cell junctions, haemoglobin complex. (hemoglobin complex, cytosol, cell junction, integral component of plasma membrane, perinuclear region of cytoplasm, plasma membrane, Golgi apparatus)

MF analysis showed that DEGs were significantly enriched in protein binding, ferric ion binding, oxygen transport activity, oxygen binding, ferrous hemoglobin binding. (oxygen transporter activity, oxygen binding, iron ion binding, heme binding,

protein binding)

Meanwhile, KEGG pathway analysis (Figure 3(b)) showed that DEGs were mainly enriched in adherent spot kinases, signaling pathways regulating pluripotency of stem cells and proteoglycan-interacting pathways in cancer. (African trypanosomiasis, Focal adhesion, Malaria, Signaling pathways regulating pluripotency of stem cells, TGF-beta signaling pathway, Notably, there was a significant enrichment in KEGG for the "adherens spot kinase" pathway, a class of cytoplasmic

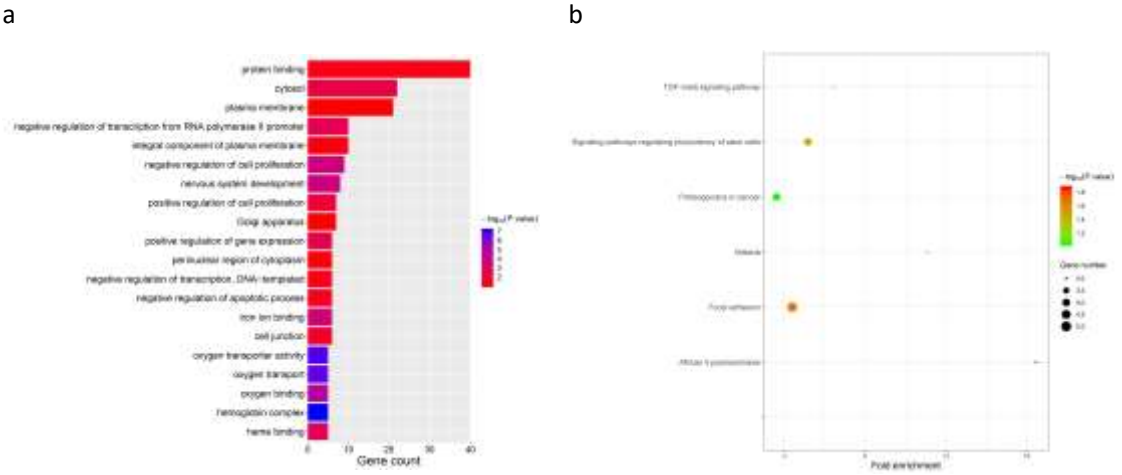

Figure 3 GO Analysis with DAVID Tools (a) Gene Ontology analysis and enrichment **(b)** KEGG pathway enrichment

non-receptor protein complex kinases associated with protein synthesis, reflecting the consistency with GO analysis.

PPI network construction and hub gene identification

The interactions between the proteins of the DEGs were predicted by STRING (http://string-db.org/cgi/input.pl?sessionId=74NZcK2I7h1B&input_page_show_search=on). The PPI network involved a total of 74 nodes (Figure 4). The node size represents how many nodes the node is associated with, i.e. the count value. The colours represent the up- and down-regulation of the proteins, where yellow represents up-regulation and blue represents down-regulation. edge indicates the degree of interaction between the linked differential proteins, and its thickness indicates the size of the score, with thicker indicating a stronger interaction between the two. We set a threshold value for each algorithm, selected the top genes for each algorithm, and then aggregated them. Combined with the Degree values, it can be visualised that Matrix metalloprotein 9 (MMP9) is the most significant gene with a degree of 16 (dynamic balance between degradation and remodelling of the extracellular matrix); the second is Receptor tyrosine-protein kinase erbB-2 (ERBB2) with a degree of 15. This is followed by Signal transducers and activators of transcription 1 (STAT1) with a degree of 11; Caveolin-1 (CAV1) with a degree of 9 (maintains the integrity of caveolae, transport of small cells, signalling); Vascular endothelial growth factor C (VEGFC), degree 8; endothelin receptor type B (EDNRB), degree 8; Serum amyloid A1 (SAA1), degree 8; Delta-

aminolevulinate synthase 2 (ALAS2), degree 7. Interferon-stimulated gene 15 (ISG15), degree 7; G1/S-specific cyclin-D2 (CCND2), degree 6; cyclin-dependent kinases 6 (CDK6), degree 6; A Disintegrin and metalloproteinase domain-containing protein 8 (ADAM8), degree 5; Tyrosine-protein kinase 6 (PTK6), degree 4; Carcinoembryonic antigen-related cell adhesion molecule 5 (CEACAM5), degree 4; DNA-binding protein inhibitor ID-1 (ID1), degree 3; where CAV1, VEGFC, EDNRB, CDK6, ALAS2, ID1 and CCND2 are upregulated genes, MMP9, ERBB2, STAT1, SAA1, ISG15, ADAM8, PTK6 and CEACAM5 were down-regulated genes.

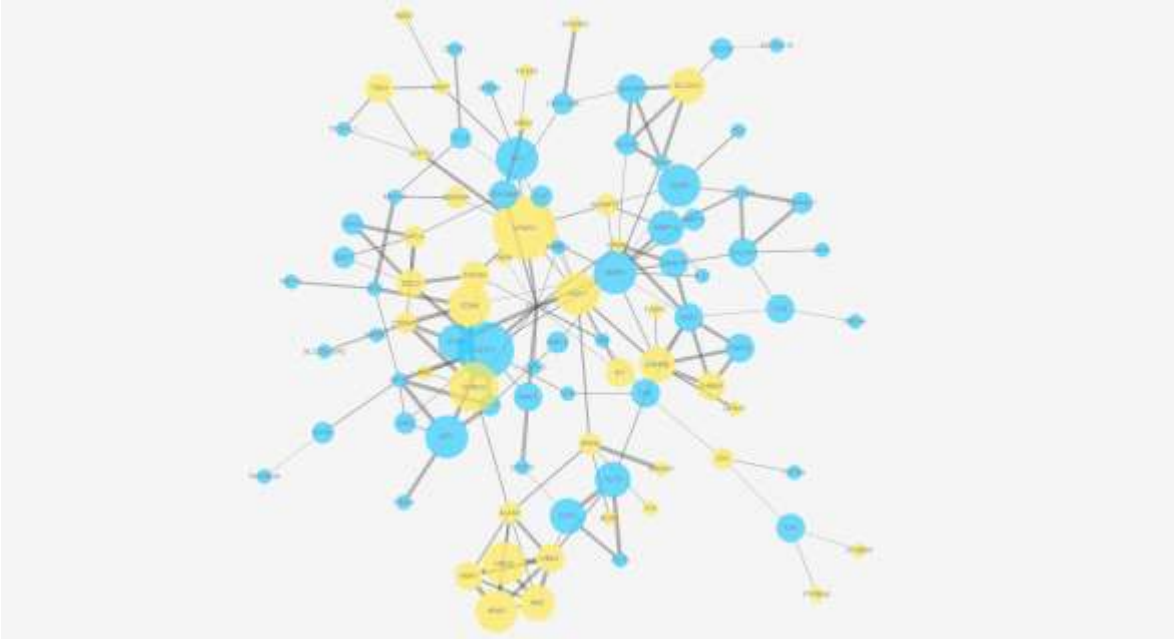

Figure 4 Protein-protein Interaction Network Yellow genes in the plot are up-regulated genes while blue genes are down-regulated genes, the size of each gene dots are the counts of genes.

Survival analysis of hub genes

To understand the prognostic value of 15 possible hub genes, we used the Kaplan-Meier Plotter (http://www.kmplot.com/analysis/index.php?p=background) analysis platform to do further survival analysis (Figure 5). The analysis ran on 720 adenocarcinoma patients. We found that high expression of six genes and low expression of three genes had a significant negative impact on the survival of adenocarcinoma patients. 6 of the genes were highly expressed: ADAM8, ERBB2, ISG15, SAA1, The six genes with high expression were ADAM8, ERBB2, ISG15, SAA1, ALAS1 and ID1; the three genes with low expression were CAV1, CCND2 and

EDNRB, with EDNEB having the most significant effect.

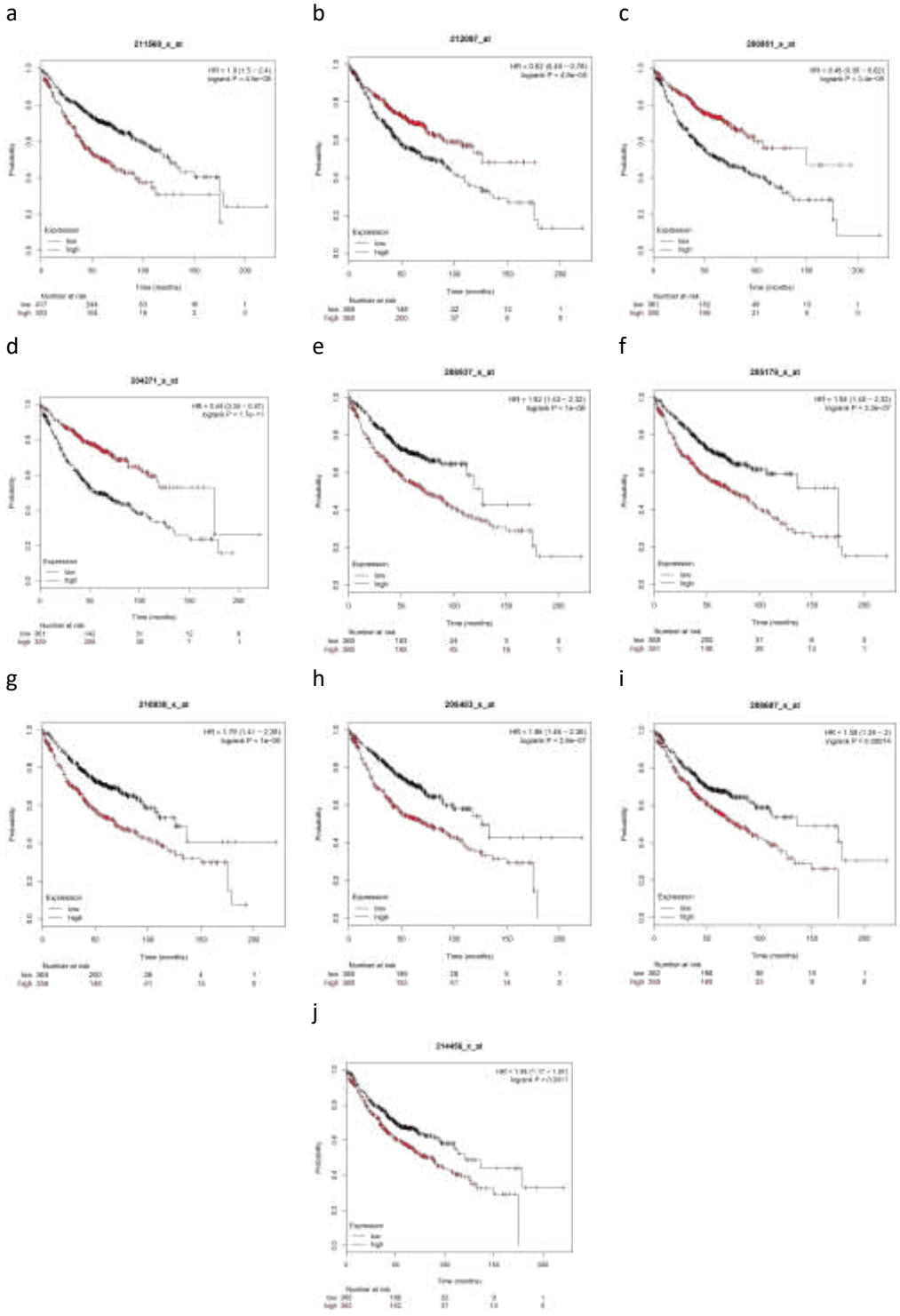

Figure 5 Kaplan-Meier Survival Analysis Plot for LUAD Hub Genes (a) ALAS2 **(b)** CAV1 **(c)** CCND2 **(d)** EDNRB **(e)** ID1 **(f)** ADAM8 **(g)** ERBB2 **(h)** ISG15 **(i)** SAA1(208607_s_at) **(j)** SAA1(214456_x_at)

Identification of candidate probes

To investigate the relationship between adenocarcinoma and loci, we selected loci with a high correlation with adenocarcinoma through multiple screens. Our data were genome-wide methylation data for GSE85845 (8 adenocarcinomas and 8 adjacent non-tumour tissues) and GSE75008 (40 adenocarcinomas and 40 normal lung tissues) from the NCBI database. The first screen statistically excluded interference and excluded probe data at sex chromosomes and SNPs. We set $|logFC| \geqslant 0.1$, $P.Value < 0.05$, removed Probe in Sex Chromatins and an SNPs, and obtained 185,592 probe data. The distribution of the data after preprocessing was observed (Figure 6), and the $|logFC| \geqslant 1.5$ was further set to select the loci with greater differentiation, and a total of 5633 entries were screened. We then used three unsupervised methods, PCA, K-Means and Isolated Forest, to perform unsupervised binary classification on each of the 5633 data items, and then obtained 3691 candidate probes by taking the intersection set.

The distribution of candidate probes was further analysed. Firstly, looking at the distribution of loci in CpG Island (Figure 7(e)), we found that 90.41% of the probes were in the Island region with 3337 entries; 2.44% of the loci were in N_Shore with 90 entries. For the distribution in Group (Figure 7(c)), we found that 27.80% of the probes were in Body with 1026, 17.07% in TSS1500 with 630, and 15.61% in 5'UTR with 576. There are 1687 Candidate probes located on the DMR, of which 94.13% are in the Island region with 1588 entries and 2.31% are in the N_Shore with 39 entries. For their distribution in groups, 29.64% were in Body with 500 entries, 18.73% were in TSS1500 with 316 entries and 14.05% were in 5'UTR with 237 entries.

Sclection of probes markers

We screened probe markers from candidate probes by LASSO regression. 17 probe markers were selected according to the set threshold. 17 of these probe markers were located in Island, accounting for 94.12% (Figure 7(f)); one was located in N_

Table 1 Distribution of Special Biological Loci from Markers Selection

Index	Detail	3691 Candidate Probes		1687 DMR Candidate Probes		17 Methylation Markers	
		Counts	Proportion	Counts	Proportion	Counts	Proportion
UCSC Gene Group	Body	1026	27.80%	500	29.64%	5	29.41%
	TSS1500	630	17.07%	316	18.73%	1	5.88%
	5'UTR	576	15.61 %	237	14.05%	3	17.65%
Relation to CpG Islands	Island	3337	90.41%	1588	94.13%	16	94.12%
	N-Shore	90	2.44 %	39	2.31%	0	0%

Shelf, accounting for 5.88%. For the distribution in Group (Figure 6(d)), we found that 29.41 % of the probes were located in Body with 5 loci, 5.88 % in TSS1500 with 1 locus, 17.65 % in 5'UTR with 3 loci and 11.76 % in TSS200 with 2 loci. 11.76 % of the loci were in 1stExon with 2 entries.

Heat map analysis of the 17 loci gave us an insight into the intrinsic linkage of these loci (Figure 7(a)). The data are the methylation beta values (0 to 1) for cancer tissue (left) and normal tissue (right) for 48 samples in GSE85845 and GSE75008. The

bluer the colour the greater the value, the redder the smaller. From the graph, it can be visualized that there is a significant difference between the methylation values of cancer and normal tissues in the clusters cg02036261, cg26419880 , cg23806894 , cg03662014 , cg13019491 , cg10903903 , cg22399133 , cg25774643.

Also to investigate the relationship between the 17 probe markers and the 9 hub genes, we made a heat map of gene expression values (Figure 7(b)). The data source is GSE75037 for 83 samples of cancer tissue (left) and normal tissue (right). The genes are refgene for hub gene and 17 probe markers. significant differences can be seen between the two clusters as well. We compared the relationship between LUAD site methylation and corresponding gene expression and could find that the higher the gene methylation compared to normal, the lower the corresponding gene expression; the lower the methylation, the higher the corresponding gene expression. In the analysis of 3691 candidate methylation sites, 1687 sites located in the differentially methylated region DMR and 17 marker sites screened by LASSO regression, we found that the sites with significant effects on LUAD were concentrated in the Body region, TSS1500 region and 5'UTR region. 'UTR region, and in the Island and N-Shore regions of the CGIs (Table 1).

To further integrate the findings with clinical diagnosis, we based the results on 17 loci from 96 samples of

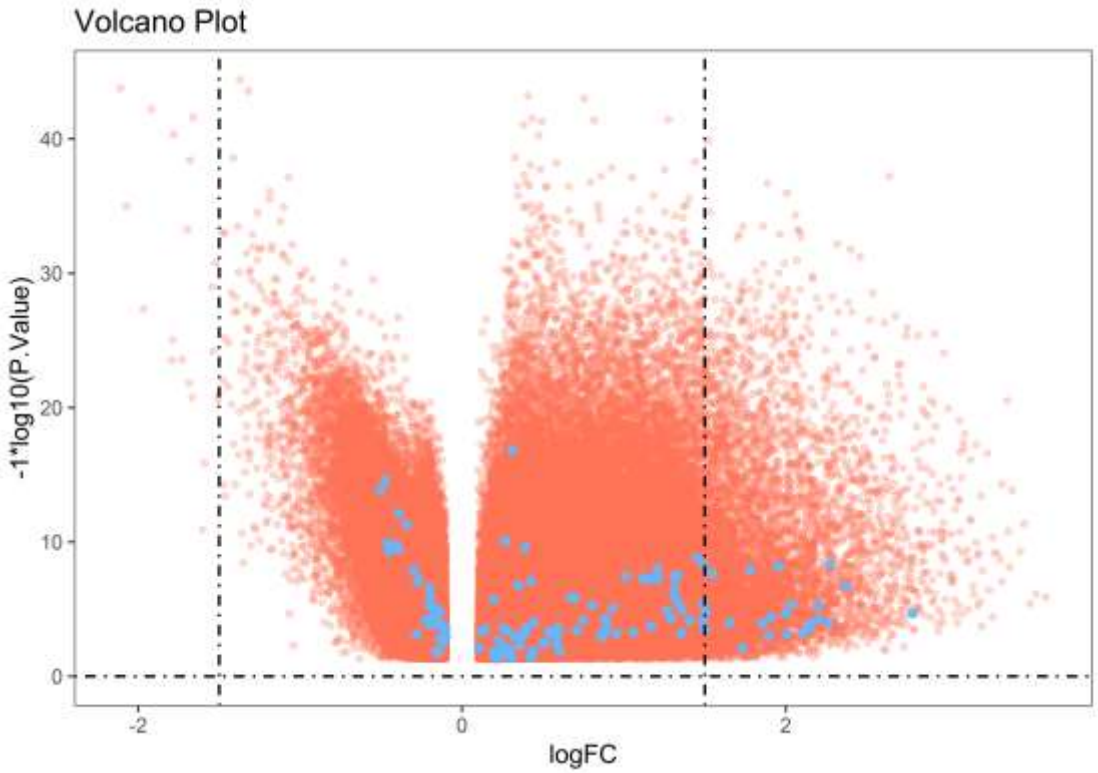

Figure 6 Volcano Plot Red dots in the plot are pre-processed DNA methylation data while blue dots are probes associated with 9 hub genes.

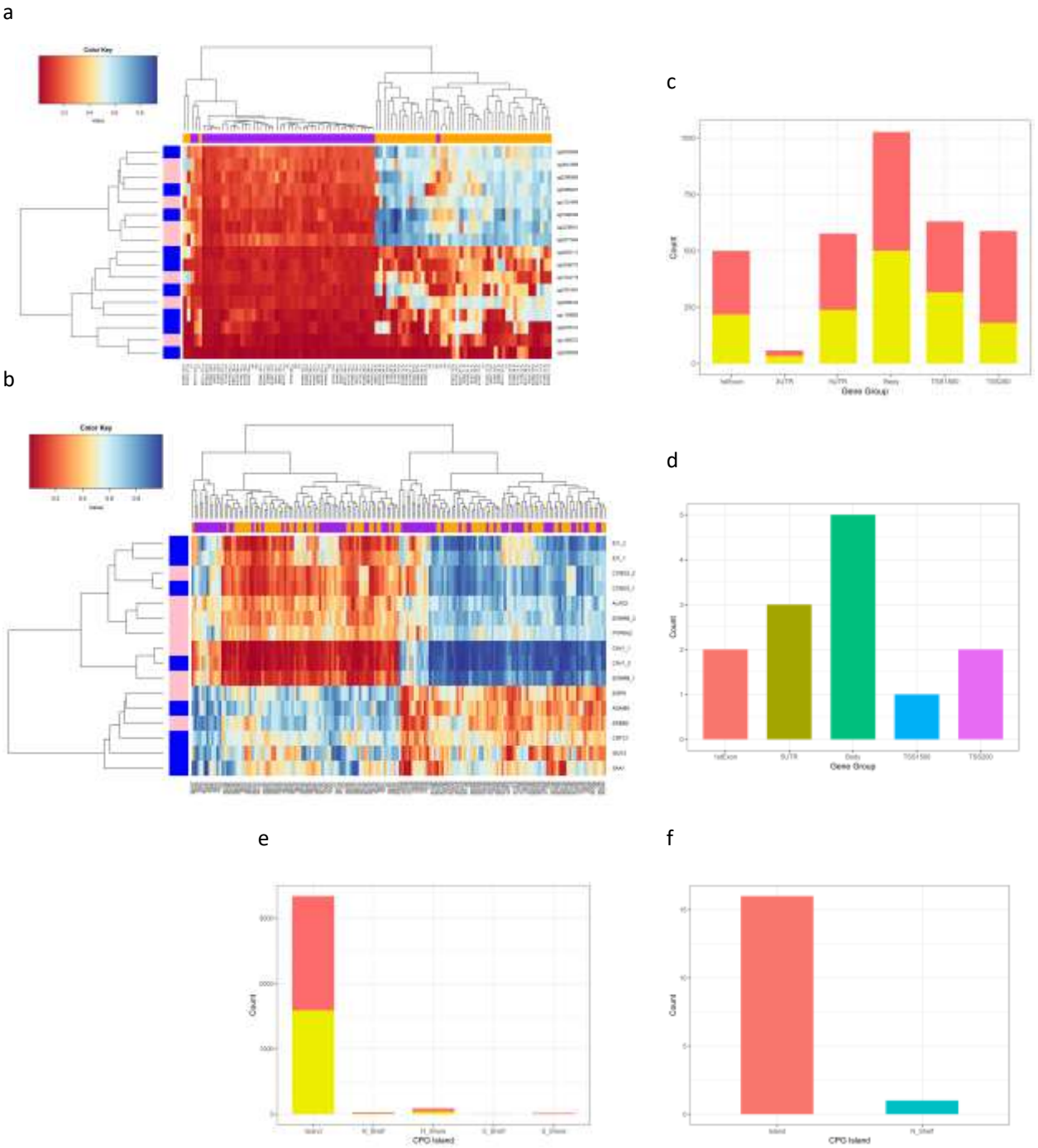

Figure 7 Genomic Feature of Candidate Probes and Markers (a) Heat map for DNA methylation profiles on 17 LUAD diagnosis markers from GSE85845 and GSE75008, the left half part are LUAD samples and the right half part are normal samples. **(b)** Heat map for Gene expression profiles on LUAD Hub Genes and UCSC Refgenes of 17 markers from GSE75037, the left half part are LUAD samples and the right half part are normal samples. **(c)** Gene groups of both 3691 candidate probes (red bars) and 1687 DMR probes (yellow bars) from candidate probes. **(d)** Gene groups of 17 methylation markers. **(e)** Relation to CGIs of both 3691 candidate probes (red bars) and 1687 DMR probes (yellow bars) from candidate probes. **(f)** Relation to CGIs of 17 methylation markers.

3. Discussion

DNA methylation, as one of the modifications, is of great importance to the study of gene expression and molecular epistasis, and thanks to advances in whole-genome methylation sequencing, we are able to further understand the mechanism of action of LUAD at a more microscopic level.

In our experiments, we first identified nine genes central to LUAD through three sets of gene expression profiles, in which upregulation of ALAS2, CAV1, CCND2, EDNRB, ID1 genes, and downregulation of ADAM8, ERBB2, ISG15 and SAA1 play key roles in LUAD. We identified 68 significantly up-regulated genes and 86 significantly down-regulated genes using the Gene Ontology GO tool, KEGG pathway analysis and the PPI protein interaction network, respectively. We obtained enrichment information for these DEGs by Gene Ontology GO analysis and KEGG pathway analysis, and obtained 15 hub genes by Cytoscape, and identified the final 9 core genes by Kaplan-Meier survival analysis. We tagged 17 loci located on these genes on the private DNA methylation profile data and identified 3691 candidate loci by annotation information as well as statistical methylation information through three unsupervised machine learning algorithms, principal component analysis, K-means and isolated forest, respectively, to find clusters of loci with comparable biological and statistical properties, followed by our use of GEO's LUAD DNA methylation profile data, and from constructing LASSO regressions on 96 samples (48 LUAD and 48 normal samples), 17 diagnostic markers for LUAD were screened out. We found significant differences in the methylation values of the screened marker loci from the normal control population samples by heat map. We counted 1687 loci in the differentially methylated region DMR (DMR) among 3691 candidate loci and 17 marker loci in terms of their gene part and relationship with CGIs, and found that LUAD methylation candidates and DMR loci were concentrated in the Island and N-Shore regions, and in the Body and 5' UTR region. We then constructed an end-to-end diagnostic model using a deep neural network for 96 sample marker loci and achieved an accuracy of 0.9868%.

We have additional advantages over previous work in that, firstly, our preliminary analysis is based on statistical gene expression data and methylation profile data, eliminating the need to analyse individual sample populations and narrowing down the pool of candidate methylation sites to focus on for further mechanistic studies through an unsupervised machine learning approach, and secondly, by using LASSO regression for marker screening, the training time for deep Secondly, with LASSO regression marker screening, the training time for deep learning is greatly reduced and the accuracy of the diagnostic prediction can be extremely high. However, there is still room for improvement in the prediction of lung cancer subtypes and tumour TNM (Tumor, Lymph Node, Metastasis) stages, as well as in clinical follow-up.

We hope that in future studies of LUAD, the construction of methylation sites can be combined more with LncRNA and Chip-Seq data to play a more significant role in the study of integrated network topology in biomics and immunotherapy. The application of deep learning has changed the previous biological experiments for clinical research, and we hope that in the future, LUAD can be used more often to build predictive and

simulation models for the whole mechanism of lung cancer, and even to build predictions for the whole cancer subtype and development period, which can provide better solutions for cancer early warning, targeted drug research and performance testing, and more cost-efficient development of clinical therapies to achieve better results.

4. Methods

Datasets and Preprocessing

In our research, we applied three LUAD gene expression profiles(GSE 43467, GSE 68465, GSE75037) to select hub genes, GSE 43467 were samples depicting lung tumor cell gene expression of four stages with 113 samples, GSE 68465 is LUAD dataset with 443 tumor samples and 19 normal samples based on Agilent GPL96 platform [HG-U133A](Affymetrix Human Genome U133A Array) and GSE43767 includes data from 83 LUAD 83 Non-malignant and was based on platform GPL6480 (Agilent-014850 Whole Human Genome Microarray 4x44K G4112F), GSE 75037 is microarray data with 83 adenocarcinomas and 83 non-malignant cell samples

Table 2 Component Selection and Codeing Job for DNA-methylation Profile

Column	Description	Encoding Job
logFC	Fold change with log2 base	(Numeric)Reserved
P.value	Probability value	(Numeric)Reserved
CHROMOSOME_36	Chromosome – genome build 36	(Multi-nomial)Exclude probes from X, Y or Multi Chromosome
STRAND	Design Strand	(Binomial)Forward and Reverse were marked respectively
UCSC_REFGENE_GROUP	Gene region feature category	(Binomial)Probes located on ”5’UTR”,”TSS200”,”1stExon” were marked as class 1, others as class 0
RELATION_TO_UCSC_CPG_ISLAND	Relationship to Canonical CpG Islands(CGIs)	(Multi-nomial)”CGIs” , ”N-Shore”, “S-Shore”, “N-Shelf”, “S-Shelf” or “Non-CGIs” were categorized respectively
PHANTOM	FANTOM-derived promoter	(Multi-nomial)”High-CGIs”, ”Low-CGIs” or “Non-CGIs” were marked respectively
DMR	Differentially methylated regions	(Binomial) Reprogramming- specific DMR (rDMR), cancer-specific DMR (CDMR) and DMR were marked as class 1, others as class 0
ENHANCER	Enhancer element	(Binomial)Enhancer element or not were marked respectively
DHS	DNAse Hypersensitive Site	(Binomial)DHS or not were marked respectively
UCSC_REFGENE_NAME	Gene Name(UCSC)	Reserved only for reference
UCSC_REFGENE_ACCESSION	Gene Acession(UCSC)	Reserved only for reference

from lung organ in 16 to 24 weeks after ovulation and was based on the Agilent GPL6884 platform (Agilent Whole Human Genome Microarray 4x44k Illumina Human WG-6 V3.0 expression beadchip). In order to select hub genes we first analyzed the three data with GEO2R tool and then set the threshold $|\log FC| \geq 1$ and $P.\text{value} < 0.05$. We retrieved the intersection of both up-regulated genes (68 genes) and down-regulated genes(86 genes) respectively (Figure 1).

Then we preprocessed our Illumina 450K DNA Methylation from GSE85845and

GSE75008. GSE85845 is genome wide DNA methylation profiling data of lung adenocarcinoma and non-tumor adjacent tissues. The Illumina Infinium 450k Human DNA methylation Beadchip was used to obtain DNA methylation profiles. Samples included eight lung cancer and adjacent non-tumor tissues excised from a cohort of 8 patients with lung adenocarcinoma and GSE75008 is genome wide DNA methylation profiling of lung cancer tissue samples and corresponding normal lung tissue samples of 40 patients. The Illumina Infinium 450k Human DNA methylation Beadchip was used to obtain DNA methylation profiles across approximately 480,000 CpGs. Samples included 40 normal lung tissue samples and 40 lung cancer samples. We analysed methylation data to calculate the overall statistic characteristics and then did encoding job before we fed the bioset to three unsupervised algorithms (Table 2), and then filtered probes located on sex chromatins and probes of single nucleotide polymorphism (SNPs) were excluded. To further select candidate probes, we set the threshold $|logFC| \geqslant 0.1$ and $P.value < 0.05$ (n=5633).

Hub Genes Detection with Gene Expression Profiles

After screening the gene expression profiles for processing thresholds, we used the Venn diagram analysis tool (http://bioinformatics.psb.ugent.be/webtools/Venn/) to extract the intersection of the three data sets for further processing (Supplementary Table 1)

To further analyse the biological properties of the screened core genes, we used the online tool Database for Annotation, Visualization and Integrated Discovery (DAVID, https://david.ncifcrf.gov/) to analyse the genes from We used the online tool Database for Annotation, Visualization and Integrated Discovery (DAVID,) to select genes from molecular functions (MF), cellular component (CC) and biological process (BP) with counts >10 for GO (Gene Ontology) analysis [citation needed] and counts >5 for KEGG pathway (Kyoto Encyclopedia of Genes and Genomes) analysis. Genes with a combined score >0.4 were then selected by the Search Tool for the Retrieval of Interacting Genes (STRING) database (https://string-db.org/cgi/input.pl) for Protein-Protein Interaction network (PPI network) [citation needed] was analyzed, visualized using Cytospace software, and selected for hub genes using the Cytohubba algorithm.

Binomial Selection of Candidate Probes with Unsupervised Algorithm

After the screening of hub genes, all loci on the hub genes were annotated (n=17) and three unsupervised analyses, namely principal component analysis (PCA), K-means and isolated forest (IF), were used to dichotomise the screened candidate probes to find loci with similar biometric properties. The first method we adopted was pca.

In the first approach, we adopted the pcaGopromoter package for principal component analysis. In principal component analysis [1], the covariance matrix 〖Cov〗_M (1), where n is the number of components, is first calculated for the statistical methylation matrix M that has been filtered and transposed after GEO2R analysis, and the orthonormal matrix P is found so that the matrix 〖Cov〗_D becomes a diagonal matrix (2), then P is the eigenvector matrix of 〖Cov〗_M and find the eigenvalue λ_PCA according to (3), where E is the unit matrix. The two principal components with the largest λ_PCA transformed by loci are selected to construct an orthogonal

coordinate system, and all loci are mapped to this plane, and then the 95% confidence interval is selected by the ellipse package to construct a correlation ellipse. We refer to the region inside the ellipse as the "high LUAD correlation region" (Class A) and the region outside the ellipse as the "low LUAD correlation outlier region" (Class B) (Figure 8(a)), and count the annotated loci in the All loci in the region with more loci were selected for further candidate site analysis.

$$Cov_M = \frac{1}{n} MM^T \tag{1}$$

$$Cov_D = \frac{1}{n} (PM)(PM)^T = PCov_M P^T \tag{2}$$

$$(Cov_M - \lambda_{PCA} E)M = 0 \tag{3}$$

In the second approach, we adopted the K-means algorithm from the h2o package [2] to further cluster the initially screened loci (Figure 8(a)). We chose K=2 to classify the "high LUAD-related" and "low LUAD-related outlier regions", and we set the class with more marker loci as Class A to indicate "high LUAD-related", and the class with fewer marker loci as Class A to indicate "high LUAD-related". We set the class with more marked loci as Class A, which means "high LUAD-related", and the class with fewer marked loci as Class B, which means "low LUAD-related outlier region", i.e., we randomly take two centroids S_i(i=1,2) in the loci data space, so that the loci belong to the nearest centroid iteratively to make Sum of the Squared Error (SSE) is minimized (4), and the iteration is considered complete when the location of the centroid no longer changes, where μ_i(i=1,2) is the mean of the probe spots in the centroid S_i(5)

$$\arg\min_s \sum_{i=1}^{2} \sum_{p \in S_i} |p_i - \mu_i|^2 \tag{4}$$

$$\mu_j = \frac{\sum_i \log ic\{S_i = j\} p_i}{\sum_i \log ic\{S_i = j\}} \quad (j = 1,2) \tag{5}$$

In the third approach we have adopted the Isolated Forest (IF) algorithm [1] to identify outliers for the already screened loci to further narrow down the range of candidate probes. In this decision tree based algorithm, outliers are identified as easily isolated outliers. The average path from the root node of the tree to each of the loci is calculated according to the decision tree(6), where H(i) is called the harmonic number valued as(7) and the isolated values are defined in exponential form(8), where Score$\in$(0,1) is identified as a normal fit sample when Score$\rightarrow$0 and as an outlier when Score$\rightarrow$1. We initially screened out the non-normal methylation sites (n=5633) and examined the distribution of the outlier Score (Figure 8(b)). Here we set Score=0.2 as the threshold and identified the set of samples below the threshold as Class A and the set of samples above the threshold as Class B.

$$\overline{Path(p_i)} = 2H(p_i - 1) - \frac{2(p_i - 1)}{p_i} \tag{6}$$

$$H(x) \approx In(x) + 0.577 \quad (7)$$

$$Score = 2^{-\frac{E(H(p_i))}{Path(p_i)}} \quad (8)$$

To circumvent the drawbacks of the three methods, we counted the distribution of labeled loci among the three unsupervised algorithms, selected the class with more labeled loci each (Table 3), and selected the intersection of the three unsupervised

Table 3 Candidate Probes Selected by Three Unsupervised Method

Algorithm	n_A	n_{HG-A}	n_B	n_{HG-B}	**Class Chosen**
Principal Component Analysis	5316	22	317	0	Class A
K-means (K=2)	5539	22	94	0	Class A
Isolation Forest (threshold=0.3)	3706	19	1927	3	Class A

A – Class A;B – Class B;HG – Hub Gene Probes.

methods (n=3691) by plotting the Wenn diagram (Figure 8(c)).

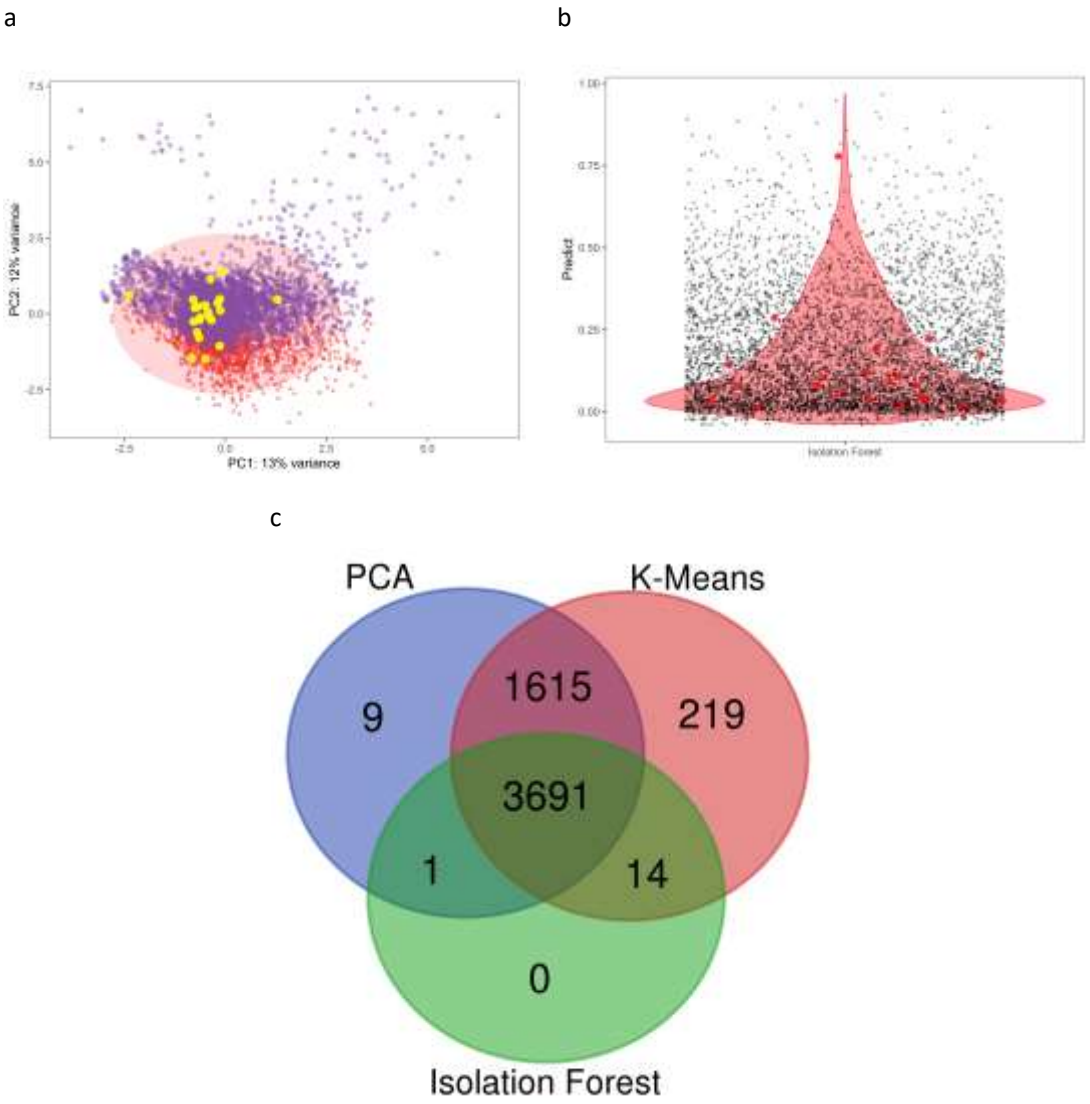

Figure 8 Distribution and Selection of Selected Candidate Probes. All the three methods were based on 5633 differentially methylated probes on the statistic LUAD DNA methylation profile **(a)**Selection results of both principal component analysis and k-means. The ellipse drawn is the 95% confident region of LUAD highly correlation region. Purple dots and red dots belong to class A and B in red respectively and yellow dots are 17 marked probes located on DEGs. **(b)**Violin plot for Isolation Forest scores on 5633 probes. Red dots are marked probes located on DEGs **(c)**Venn plot for finding interactive part of probes selected by three unsupervised algorithm.

Diagnosis Markers Selection and Validation

The selection of diagnostic markers was performed based on 3691 candidate loci. We used the linear model of the Least absolute shrinkage and selection operator (LASSO) [1], calling the glmnet package of R for screening (n=17) (Supplementary Table 1). (9) Here is m denotes the total number of samples, i

denotes the sample subscript, j denotes the methylation marker subscript, and λ is the penalty term. The effect of methylation sites on diagnostic parameters (i.e., coefficients) changes continuously during the λ adjustment process (Figure 9(a)), and we adjust the penalty term so that the minimum error is obtained in the case where λ = 0.02954903 (Figure 9(b)), the loci with β^_Probe $\neq$ 0 were derived as diagnostic identifiers.

$$\beta_{\mathrm{Pr}obe} = \arg\min_{\beta}\left[\frac{1}{m}\sum_{i}(Y_i + \beta X_i^T)^2 + \lambda\sum_{j}|\beta_j|\right] \quad (9)$$

After screening all the diagnostic markers we used Deep Learning (DL) [1] to construct a diagnostic model of LUAD on the screened methylation markers to test the efficacy of the screened marker loci. We used the Tensorflow-based Keras package to construct the fully connected layers in R (Figure 10), and we connected two adjacent fully connected layers by the activation function Rectified Linear Unit (ReLU) (11) and used Mean Square Error (MSE) as the loss function (12).

$$relu(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \quad (10)$$

$$MSE(y) = \frac{1}{N}\sum_i (y_i - \hat{y}_i)^2 \qquad (11)$$

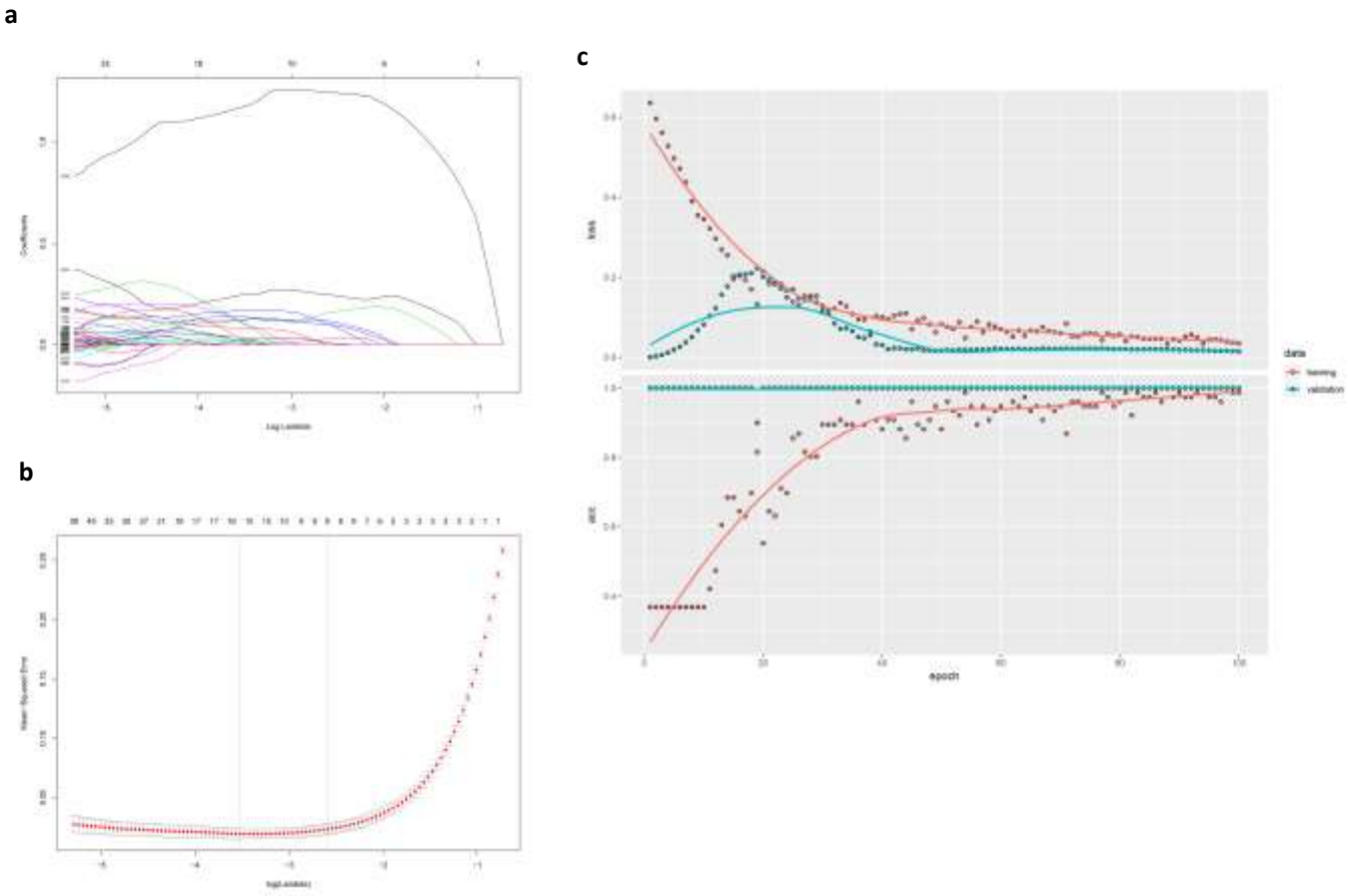

Figure 9 Using LASSO in further selecting DNA-Methylation Probes in Lung Adenocarcinoma and performance in Diagnosis. (a) Coefficient of candidate probes contributed to diagnosis results. The number of each line marks the order No. of the probes. **(b)**Prediction loss of lambda fine-tuning when on LASSO cross validation step. **(c)**The loss curve and the learning curve when applying deep learning on LUAD diagnosis based on DNA methylation diagnosis markers.

Here x represents the input value of the corresponding layer Hyperparameters, y_i represents the actual normal/LUAD diagnostic result, and (y_i)ˆ is the predicted result of the Deep Learning-based model diagnosis.

We took nodes (17, 17, 17, 16, 16, 8, 4, 2, 1) in order to first model the methylation beta values (12) of 17 DNA Methylagion Diagnosis Markers from 96 samples, where I_M denotes Methylated Allele Intensity, I_U denotes Unmethylated Allele Intensity, and then gradually extract the sequence information step by step in integer powers of 2, and output the probability of binary classification through a fully connected layer with a terminal node number of 1. Due to the small sample size, we inserted two Dropout layers to randomly discard 20% of the data each to prevent overfitting, and used RMSprop Optimizer with Epochs = 100 for training, and we obtained an accuracy of 0.9868% (Figure 9(c)).

$$\beta = \frac{I_M}{I_M + I_U + 100} \qquad (12)$$

Conclusion

Our bioinformatics analysis identified 154 DEGs in Based on gene expression datasets obtained from the GEO database, we identified 154 DEGs, of which 15 central genes are likely to be central to lung adenocarcinoma. Seven of these were up-regulated and eight were down-regulated. Of these, CAV1, VEGFC, EDNRB, CDK6, ALAS2, ID1 and CCND2 were up-regulated genes and MMP9, ERBB2, STAT1, SAA1, ISG15, ADAM8, PTK6 and CEACAM5 were down-regulated genes. Furthermore, we found that high expression of six of these genes and low expression of three genes had a significant detrimental effect on the survival of adenocarcinoma patients.

In studying the relationship between adenocarcinoma and loci, our further studies applied machine learning unsupervised algorithms to narrow down candidate DNA methylation probes and build a deep learning-based diagnostic model with 98.68% accuracy in predicting cells as malignant and benign.

References

[1] K. D. Miller *et al.*, "Cancer treatment and survivorship statistics, 2016," *CA: A Cancer Journal for Clinicians,* vol. 66, no. 4, pp. 271-289, 2016/07/01 2016.

[2] R. S. Herbst, D. Morgensztern, and C. Boshoff, "The biology and management of non-small cell lung cancer," *Nature,* Review Article vol. 553, p. 446, 01/24/online 2018.

[3] J. Kim *et al.*, "CPS1 maintains pyrimidine pools and DNA synthesis in KRAS/LKB1-mutant lung cancer cells," *Nature,* vol. 546, p. 168, 05/24/online 2017.

[4] B. Hewelt, H. Li, M. K. Jolly, P. Kulkarni, I. Mambetsariev, and R. Salgia, "The DNA Walk and its Demonstration of Deterministic Chaos— Relevance to Genomic Alterations in Lung Cancer," *Bioinformatics,* pp. bty1021-bty1021, 2019.

[5] L. Simo-Riudalbas *et al.*, "Genome-Wide DNA Methylation Analysis Identifies Novel Hypomethylated Non-Pericentromeric Genes with Potential Clinical Implications in ICF Syndrome," *PLOS ONE,* vol. 10, no. 7, p. e0132517, 2015.

[6] A. E. Teschendorff *et al.*, "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data," *Bioinformatics (Oxford, England),* vol. 29, no. 2, pp. 189-196, 2013.

[7] A. E. Teschendorff and M. Widschwendter, "Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions," *Bioinformatics,* vol. 28, no. 11, pp. 1487-1494, 2012.

[8] R.-h. Xu *et al.*, "Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma," *Nature Materials,* Article vol. 16, p. 1155, 10/09/online 2017.

[9] J. Qiu *et al.*, *CpG Methylation Signature Predicts Recurrence in Early-Stage Hepatocellular Carcinoma: Results From a Multicenter Study*. 2017, p. JCO2016682153.

[10] H. Ishihara, S. Yamashita, S. Fujii, K. Tanabe, H. Mukai, and T. Ushijima, "DNA methylation marker to estimate the breast cancer cell fraction in DNA samples," *Medical Oncology,* vol. 35, no. 11, p. 147, 2018/09/14 2018.

[11] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, "Tumor origin detection with tissue-specific miRNA and DNA methylation markers," *Bioinformatics,* vol. 34, no. 3, pp. 398-406, 2017.

[12] Y. Wang *et al.*, "Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks," *Scientific reports,* vol. 6, pp. 19598-19598, 2016.

[13] W. Xue, X. Wu, F. Wang, P. Han, and B. Cui, "Genome-wide methylation analysis identifies novel

prognostic methylation markers in colon adenocarcinoma," *Biomedicine & Pharmacotherapy,* vol. 108, pp. 288-296, 2018/12/01/ 2018.

[14] W. He, D. Ju, Z. Jie, A. Zhang, X. Xing, and Q. Yang, "Aberrant CpG-methylation affects genes expression predicting survival in lung adenocarcinoma," *Cancer Medicine,* vol. 7, no. 11, pp. 5716-5726, 2018/11/01 2018.

[15] X. Song *et al.*, "High PITX1 expression in lung adenocarcinoma patients is associated with DNA methylation and poor prognosis," *Pathology - Research and Practice,* vol. 214, no. 12, pp. 2046-2053, 2018/12/01/ 2018.

[16] Z. Yang, B. Liu, T. Lin, Y. Zhang, L. Zhang, and M. Wang, "Multiomics analysis on DNA methylation and the expression of both messenger RNA and microRNA in lung adenocarcinoma," *Journal of Cellular Physiology,* vol. 234, no. 5, pp. 7579-7586, 2019/05/01 2019.

[17] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics,* vol. 2, no. 4, pp. 433-459, 2010/07/01 2010.

[18] H. Xiong, J. Wu, and J. Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 2, pp. 318-331, 2009.

[19] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413-422.

[20] N. E. Breslow, C. Nazionale, C. Convegni, and S. Agostino, *Generalized Linear Models: Checking Assumptions and Strengthening Conclusions*. 1996.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, p. 436, 05/27/online 2015.

Supplementary Table 1 Selected DEGs	
Type	**Gene Names**
Up-Regulated	PRKCB BCHE HBZ ID1 SPTBN1 TSPAN7 RASIP1 SDC2 HBG2 TCP1 SNCAIP TYRP1 SNCA NRN1 F10 TBX3 GPER1 SRPX THRA GPC3 MAF CHRM3 ALAS2 FOXF2 ID3 EML1 VSNL1 ALDH1A2 HDGFRP3 VEGFC EDNRB PTPRN2 CXorf57 CAV1 BEX1 GSTM3 CDKN1C CTNNAL1 GPM6B RELN ADD2 HBA2 PSIP1 FGFR4 COLEC12 MOCS1 ADARB1 P3H2 CHST7 GFRA1 NSG1 SYNGR1 SLC2A3 CDK6 SGCG SEMA6A LEFTY2 TMSB15A TMSB15B ADAMTS1 STXBP6 MSX1 ITM2A CCND2 PPP1R3C TUBB1 HBG1 HBA1
Down-Regulated	STYK1 SLC2A5 SLC12A8 CXCL13 PGGHG IRF7 TMPRSS4 XAF1 CLDN3 ERBB2 CEACAM5 FUT3 TMC5 KRT16 COMP TNFRSF21 ADAM8 JUP SULF1 GGTLC1 LSR FUT2 KIF26B SLC22A18AS MMP13 MYO6 PRRG4 LRRC31 SFN GREM1 MUC4 ISG15 VDR MST1R MYH14 CD79A ST14 COL17A1 LMO3 PLEKHS1 AIM2 SIX1 FGD6 BIRC3 ANKRD36B CEACAM6 SLC1A7 SLC38A10 TRIM2 VWA1 ARHGEF16 MXRA5 PLPP2 STAT1 MMP12 MMP11 ADAM28 SRPX2 PLA2G2D LAD1 GSDMB SAA1 ICA1 SPSB1 GCNT3 CEACAM1 ELMO3 SPINK1 CILP ADAMDEC1 TFCP2L1 FUT6 RAMP1 RRBP1 LAMP5 FAP P2RY6 IL37 ABCA4 CFB CDCP1 COL10A1 PTK6 DPY19L1 ANXA11 MMP9

Supplementary Table 2 List of LUAD DNA Methylation Markers Selected by LASSO

Probe	Chromosome	Coordite	Strand	Gene Name	Accession	Gene Group	Relation to CGIs	DMR	Enhancer	Regulatory Feature Group	DHS
cg01219549	9	139803470	F	EHMT1	NM_024757	Body	Island				
cg01485157	16	4371585	R	VASN; CORO7	NM_138440; NM_024535	Body;				Unclassified_Cell_type_specific	
cg01875106	11	131792552	R	OPCML;	NM_002545; NM_001012393	3'UTR;				Unclassified_Cell_type_specific	
cg02391713	6	105507678	R				Island	RDMR			
cg02671826	6	157232955	R	ARID1B;	NM_017519; NM_175863; NM_020732	Body;			TRUE		
cg02710090	1	205291751	F	PFKFB2; YOD1	NM_001018053; NM_006212; NM_018566	TSS1500;	N_Shore	CDMR			
cg02856338	8	101891284	R						TRUE	Unclassified_Cell_type_specific	
cg03292206	2	111141385	F	BUB1	NM_004336	Body			TRUE		
cg04284535	17	38979671	R	ETV4;	NM_001986; NM_001079675	TSS1500	Island				
cg04549287	13	51850320	F	THSD1;	NM_199263; NM_018676	Body					
cg04837071	9	139438470	F	EXD3; NOXA1	NM_017820; NM_006647	TSS1500; Body	Island				

cg04877780	17	78303355	R	TBCD	NM_005993	5'UTR; 1stExon	Island			Promoter_Associated	
cg05261299	12	50723566	R								
cg05347948	10	73731686	F				S_Shelf				
cg06602478	3	196640585	R	ACAP2	NM_012287	Body	N_Shelf				
cg06677239	6	169831771	F	WDR27	NM_182552	5'UTR			TRUE		
cg06712559	1	958258	F	AGRN	NM_198576	Body	Island		TRUE	Unclassified	TRUE
cg07055259	2	235965594	F								
cg08263416	14	71790588	R	RGS6	NM_004296	Body			TRUE		
cg08539620	17	37371556	F	CNP	NM_033133	TSS1500	N_Shore			Promoter_Associated	
cg08610968	10	133647317	F				S_Shore			Unclassified_Cell_type_specific	TRUE
cg09464735	5	172108460	F				Island		TRUE		
cg09635874	13	97750519	F	FARP1	NM_005766	Body			TRUE		
cg09744759	7	5435781	F				Island			Unclassified_Cell_type_specific	
cg09968630	11	22171130	R	ANO5	NM_213599; NM_001142649	TSS200;	Island	DMR			TRUE
cg11519760	6	170439419	R	DLL1	NM_005618	Body	Island				
cg12109968	20	49817083	R	ATP9A	NM_006045	Body	Island				
cg12152566	5	176914516	F	FAM193B;	NR_024019; NM_019057	TSS1500;	S_Shore			Promoter_Associated	

cg12251075	17	46599716	R	NME2; NME1-NME2	NM_001018137; NM_001018138; NM_001018139; NM_001018136; NM_002512	Body	S_Shore			Promoter_Associated	
cg13023623	14	56346010	R	OTX2	NM_021728	5'UTR	Island				
cg13323752	12	7916834	F	SLC2A14	NM_153449	TSS200	Island				
cg13505279	15	53399173	R	PIGB	NM_004855	Body	S_Shore				
cg15689513	17	1581013	R	WDR81;	NM_001163673; NM_152348;NM_001163811;NM_001163809	Body	N_Shelf		TRUE		TRUE
cg16051195	6	166188173	R				Island				
cg16700364	2	156887135	R				S_Shore	RDMR		Unclassified_Cell_type_specific	
cg17610169	11	10707549	R								
cg19459207	5	49773530	F	EMB	NM_198449	TSS1500	S_Shore			Promoter_Associated_Cell_type_specific	
cg19533977	17	55074464	F	CLTC	NM_004859	Body			TRUE	Unclassified_Cell_type_specific	
cg20557603	12	67160220	F						TRUE		
cg21880624	21	39116548	R	ETS2	NM_005239	Body	Island				
cg22622477	17	44040441	R	HOXB7	NM_004502	Body	Island				
cg22747480	10	1156867	R	WDR37	NM_014023	Body				Unclassified_Cell_type_specific	

cg23371746	1	119334448	R	TBX15	NM_152380	TSS1500	S_Shore			
cg24331475	18	19850913	F	TTC39C	NR_024232; NM_153211; NM_001135993	Body; 5'UTR	S_Shore	RDMR		Promoter_Associated
cg25594100	7	4753469	F	FOXK1	NM_001037165	Body	Island			
cg26069973	7	79918149	F						TRUE	
cg26419880	1	6438211	F	ESPN	NM_031475	Body	Island			Unclassified_Cell_type_specific
cg26906217	11	17208532	R				S_Shore			Unclassified_Cell_type_specific