Integrative modeling of multi-omics data for predicting tumor mutation burden in lung cancer patients

Peng Chen

Shenzhen University

Jun Wang

Shenzhen University

Mingyang Su

Shenzhen University

Guocheng Zhong

Shenzhen University General Hospital

Shasha Zhang

Shenzhen University

Deming Gou (✉ dmgou@szu.edu.cn)

Shenzhen University <https://orcid.org/0000-0002-3312-841X>

Primary research

Keywords: Immunotherapy, Multi-omics, Tumor mutation burden prediction, Lung adenocarcinoma, Machine learning

Posted Date: July 22nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-45668/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Immunotherapy has been widely used in the treatment of lung cancer, and one of the most effective biomarker for the prognosis of immunotherapy currently is tumor mutation burden (TMB). Although whole-exome sequencing (WES) could be utilized to assess TMB, several problems prevent its routine clinical application.

Methods

To develop a simplified TMB prediction model, patients with lung adenocarcinoma (LUAD) in The Cancer Genome Atlas (TCGA) were randomly split into training and validation cohorts, and categorized into TMB-high (TMB-H) and TMB-low (TMB-L) groups respectively.

Results

Based on the 610 differentially expressed genes, 50 differentially expressed miRNAs and 58 differentially methylated CpG sites between TMB-H and TMB-L patients, we constructed 4 predictive signatures and established TMB prediction model through machine learning methods that integrating the expression or methylation profiles of 7 genes, 7 miRNAs and 6 CpG sites. The multi-omics model exhibited excellent performance in predicting TMB with the area under curve (AUC) of 0.911 in the training cohort and 0.859 in the validation cohort. Besides, the significant correlation between the multi-omics model score and TMB was observed.

Conclusions

In summary, we developed a prognostic TMB prediction model by integrating multi-omics data in patients with LUAD, which might facilitate the further development of quantitative real time-polymerase chain reaction (qRT-PCR) based TMB detection assay.

1 Background

Lung cancer is one of the most common malignancies worldwide, and it is the first leading cause of tumor-related mortality with an increasing incidence in recent years [1]. It was reported that 2.1 million new cases of lung cancer were diagnosed around the world in 2018, which accounted for 11.6% of all new cancer patients [2, 3]. Despite the improvements in chemotherapy and targeted therapy, the 5-year overall survival (OS) for patients with lung cancer remains poor [1, 4]. Nevertheless, immunotherapy, especially the application of immune checkpoint inhibitors (ICIs), had made a great breakthrough in the

treatment of cancer and dramatically increased survival rate and quality of life for patients with lung cancer [5–9].

As the most successful representative of immunotherapy, programmed cell death-1/programmed cell death ligand-1 (PD-1/PD-L1) inhibitors had shown better performance over the conventional chemotherapy in terms of OS, response rate, and progression-free survival (PFS) for the treatment of lung cancer [10, 11]. Four ICLs that targeting PD-1 or PD-L1 have been approved by the U.S. Food and Drug Administration (FDA) for treatment in patients with non-small cell lung cancer (NSCLC) [12]. Furthermore, a large amount of clinical research had demonstrated that immunotherapy alone or in combination with chemotherapy could be used for first-line treatment of patients with metastatic lung cancer [13–16]. It was reported that patients with higher PD-L1 expression had better outcomes compared to patients with lower or no PD-L1 expression using the anti-PD-L1 antibody clone 22C3 [17]. Unfortunately, only 10%-20% of NSCLC patients have considerable curative effects and the majority of patients cannot benefit from immunotherapy [18–20], therefore biomarker is urgently needed to rationalize the utilization of immunotherapy for patients.

Tumor mutation burden (TMB) emerged recently as a reliable biomarker that significantly correlated with immunotherapy efficacy across a wide spectrum of tumor types. TMB is defined as the number of somatic mutations per megabase (Mb) of the genome examined. Previous studies found that higher TMB was associated with improved objective response, durable clinical benefit, and PFS in NSCLC patients treated with PD-1/PD-L1 inhibitors [21]. It had been reported that PFS among stage IV or recurrent NSCLC patients with high TMB was significantly longer with PD-1/PD-L1 plus cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) treatment than with chemotherapy [22]. Moreover, through analyzing 7,033 patients with different types of cancer, TMB was found to be a useful biomarker for predicting the response of ICLs across different types of cancer, and higher TMB (highest 20% in each histology) was associated with better OS [23].

Whole-exome sequencing (WES) is considered as the gold standard for the measurement of TMB, but it is time-consuming and carries high cost [24]. Thus, targeted next generation sequencing (NGS) that profiles the mutation landscape of selected genes has been adopted as an alternative approach for predicting the TMB. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) (468 genes) and FoundationOne companion diagnostic (CDx) (324 genes) are two extensively utilized targeted NGS methods, and both of them have been approved by the FDA for clinical application (<https://www.cancer.gov/news-events/cancer-currents-blog/2017/genomic-profiling-tests-cancer>). Despite the fact that targeted NGS is effective in predicting TMB, various problems arise for its routine clinical application, such as the limit of detection, germline mutation exclusion, and standard cutoff threshold determination [25]. Moreover, targeted NGS is still time-consuming and costly compared with other clinical molecular tests [25]. In an effort to establish a simplified, more cost-effective and faster approach to predict TMB in patients with lung adenocarcinoma (LUAD), this study was conducted to develop a multi-omics model that can precisely categorize patients into high TMB or low TMB for better immunotherapy prognosis.

In this study (Fig. 1), we firstly divided patients in The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) into TMB-high (TMB-H) and TMB-low (TMB-L) groups. Then, the differentially expressed genes, miRNAs and differentially methylated CpG sites were calculated between the two TMB groups in training cohort. Next, a multi-omics TMB prediction model (TPM) involving the expression profiles of selected genes, miRNAs and methylation profiles of CpG sites was established through the machine learning method with high sensitivity and specificity. Finally, the independent cohort (validation cohort) was used to verify the performance of TPM, therefore enabling the prediction of the TMB in patients with LUAD.

2 Materials And Methods

1.1 Multi-omics datasets acquisition from TCGA

Somatic mutation profiles of 567 samples, gene expression profiles of 594 samples, DNA methylation profiles of 507 samples and miRNA expression profiles of 495 samples were obtained from the TCGA database using either GDC tool (<https://portal.gdc.cancer.gov/>) or TCGAbiolinks R package (Additional file 6: Table S1) [26]. The somatic mutation profiles (mutation annotation format, MAF) were processed by Mutect software. Missense mutations, nonsense mutations, splice-site mutations, frameshift insertions, frameshift deletions, in-frame insertions or in-frame deletions identified in the samples were regarded as positive mutations. The gene expression profiles of 594 samples were annotated through g:Profiler website [27], and normalized using the scale method in limma package [28]. The low abundance profiles were eliminated. The DNA methylation profiles were annotated using IlluminaHumanMethylation450kanno.ilmnl12.hg19 R package. Quality control for the DNA methylation profiles was conducted through minfi R package to eliminate certain CpG sites [29], in which the single-nucleotide polymorphisms (SNPs) existed [30], multiple mapping to human reference genome was found [31], and the methylation information of any samples was not available. In addition, the CpG sites located in the sex chromosomes were excluded for analysis [32]. The miRNA expression profiles of 495 samples including 450 samples from LUAD tissue and 45 samples from matched normal lung tissue in TCGA database were downloaded from the University of California Santa Cruz (UCSC) Xena database (<https://xena.ucsc.edu/public>). Then, the miRNA expression profiles were transformed into reads per million (RPM) and the miRNAs expressed in more than 10% of patients with LUAD were extracted. The clinical information of 522 patients with LUAD from the TCGA database was obtained using TCGAbiolinks R package [26]. The clinical information covered submitter id, age, gender, tumor stage, state, weight, Body Mass Index (BMI), alcohol history, height, days to last follow up, years smoked and race, the information was used for subsequent analysis (Table 1).

1.2 TMB calculation and classification of patients

Somatic mutation profiles of the acquired samples were processed by Mutect software, and the identified somatic mutations, including base substitution, deletions, and insertions, were filtered according to the

following criteria: first, the minimum sequencing coverage for the mutations should be greater than or equal to 10; second, the variant allelic fraction should be greater than or equal to 5%. Then, the TMB was calculated as the total count of somatic mutations identified divided by 38 Mb, which is the length of exons in this study. According to the previously reported cutoff threshold of 10 in patients with LUAD [22, 33], they were divided into TMB-H group ($\text{TMB} \geq 10$) and TMB-L group ($\text{TMB} < 10$). Density plot of TMB-distribution for all patients with LUAD and boxplot of the correlation between tumor stage and TMB was drawn by ggplot2 R package .

1.3 Tumor-infiltrating immune analysis

Tumor-infiltrating immune analysis was performed through Tumor Immune Estimation Resource (TIMER) tool [34], which is a comprehensive resource for systematical analysis of immune infiltrates across diverse cancer types. The estimated abundances of six immune infiltrates (B cells, CD4(+) T cells, CD8(+) T cells, Neutrophils, Macrophages and Dendritic cells) were compared between TMB-H and TMB-L patients.

1.4 Multi-omics analysis between TMB-H and TMB-L patients

Differentially expressed genes between TMB-H and TMB-L patients in the training cohort were identified through limma R package with $\text{adj.p-value} < 0.01$ and $\log_2 \text{FoldChange} > 3$ [28], and then illustrated in volcano plot and heatmap by ggplot2 and pheatmap R package respectively . Differentially expressed miRNAs between TMB-H and TMB-L patients in the training cohort were identified through limma R package with $\text{adj.p-value} < 0.05$ and $\log_2 \text{FoldChange} > 0.35$ [28], and then illustrated in volcano plot and heatmap by ggplot2 and pheatmap R package respectively . In addition, the target genes of the differentially expressed miRNAs were searched and analyzed through miRWalk website tool (<http://mirwalk.umm.uni-heidelberg.de/>) [35]. Differentially methylated CpG sites between TMB-H and TMB-L patients in the training cohort were identified through limma R package with $\text{adj.p-value} < 0.05$ and $\log_2 \text{FoldChange} > 0.15$ [28], and then illustrated in volcano plot and heatmap by ggplot2 and pheatmap R package respectively and annotated by IlluminaHumanMethylation450kanno.ilmn12.hg19 R package .

1.5 Functional enrichment analysis

Gene Ontology (GO) database and Kyoto Encyclopedia of Genes and Genomes (KEGG) database are two extensively used approaches for functional enrichment studies of large-scale genes [36, 37]. We first converted the gene symbols into ENTREZ ID via org.Hs.eg.db R package, and then GO and KEGG analysis of differentially expressed genes were implemented using ggplot2, enrichplot and clusterProfiler R

packages [38]. Meanwhile, the GO and KEGG enrichment analysis were conducted for the target genes of differentially expressed miRNAs using the same method as described above.

1.6 Construction of the TPM

Firstly, we constructed 4 possible prediction biomarker signatures: gene signature (45 genes), miRNA signature (45 miRNAs), CpG site signature (45 CpG sites), and multi-omics signature (15 genes + 15 miRNAs + 15 CpG sites) using differentially expressed genes, miRNAs and differentially methylated CpG sites with a minimum p-value between TMB-H and TMB-L patients in the training cohort. Then, the expression profiles or methylation profiles of each gene, miRNA and CpG site from the 4 biomarker signatures in the training cohort were extracted. Next, the least absolute shrinkage and selection operator (LASSO) logistic regression model analysis was performed to select the optimal biomarker signature for predicting TMB through glmnet R package [39]. The predictive performance for each biomarker signature was evaluated by the lambda.min and matched area under curve (AUC). Finally, the differentially expressed or methylated genes, miRNAs and CpG sites identified with non-zero regression coefficients were used to construct the TPM. The TPM score was calculated using the regression coefficients from the LASSO analysis to weight the expression or methylation level of the chosen biomarkers. The validation cohort was used to evaluate the performance of the TPM through assessing the predicting sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and the AUC.

1.7 Principal component analysis (PCA)

Differentially expressed genes, miRNAs, and differentially methylated CpG sites identified through LASSO analysis were used to perform PCA. The expression or methylation profiles of the genes, miRNAs and CpG sites were extracted from each patient, and ggfortify R package was utilized to conduct the PCA .

1.8 ROC analysis

Receiver operating characteristic (ROC) curve is a useful tool to evaluate classifiers for biomedical and bioinformatics applications. In this study, ROC curve analysis was conducted using pROC R package to investigate the performance of TPM in predicting TMB [40].

1.9 Correlation analysis and regression analysis

The correlation between TPM score and TMB was analyzed by cor.test R function with the two-side pearson method. Samples were plotted in two-dimensional plots with the TPM score and TMB value. Regression analysis between TPM score and TMB was performed using lm R function.

3 Results

1.1 TMB-based division of patients with LUAD

The WES data of tumor tissue from a total of 567 patients were acquired from the TCGA-LUAD database, and the clinical characteristics of the patients were summarized (Table 1), the mean age of the patients was 65.800, among which 242 were males and 280 were females (Table 1). TMB was calculated as the number of somatic mutations identified per megabase (Mb) in the tumor tissue sample of each patient. It was found that most of patients with LUAD had a TMB ranging from 0 to 40 (Figure 2A). According to the cutoff threshold of TMB=10, 184 patients were classified as TMB-H and 383 patients were classified as TMB-L (Figure 2B). The TMB-H and TMB-L patients were found evenly distributed in different tumor stages as expected (Figure 2C). From the TCGA-LUAD project, tumor samples from 440 patients were found having coupled WES, RNA-seq, miRNA-seq, and DNA methylation data (Figure 3A, Additional file 6: Table S1), among which 148 patients belong to TMB-H group and 292 patients belong to TMB-L group (Figure 3B). Patients were then randomly split into the training cohort (70%, 103 TMB-H patients Vs 204 TMB-L patients) and the validation cohort (30%, 45 TMB-H patients Vs 88 TMB-L patients) without overlap for developing a multi-omics model to predict TMB.

1.2 The landscape of tumor-infiltrating immune cells in patients with LUAD

The proportions of different tumor-infiltrating immune cells between TMB-H patients and TMB-L patients were calculated through the website tool “TIMER”, the landscape of the tumor-infiltrating immune cells was summarized via the heatmap, in which the abundance of CD4(+) T cells ($p = 0.030$) showed more abundant density in TMB-L patients compared with TMB-H patients (Figure 4A, Additional file 1: Figure S1). Whereas the B cell, CD8(+) T cell, Dendritic cell, Macrophage cell and Neutrophil cell had similar density between TMB-H patients and TMB-L patients (Additional file 1: Figure S1, Figure 4A). Meanwhile, the correlations among different tumor-infiltrating immune cell types were moderate or weak (Figure 4B). These results suggested that the TMB status might be associated with the abundance of CD4(+) T cell.

1.3 Multi-omics analysis of transcriptome, miRNAome and methylome between TMB-H and TMB-L patients

Differentially expressed genes, miRNAs and differentially methylated CpG sites between TMB-H patients and TMB-L patients in the training cohort were identified. In summary, 480 genes and 36 miRNAs were upregulated in TMB-H patients, whereas 130 genes and 14 miRNAs were downregulated in TMB-H patients (Figure 5A, B, C, D). Moreover, 10 CpG sites were hypermethylated and 48 CpG sites were hypomethylated in TMB-H patients (Figure 5E, F). GO-enrichment analysis with clusterProfiler R package suggested that the differentially expressed genes were mainly involved in the biological processes including nuclear division, chromosome segregation, organelle fission and so on (Additional file 2: Figure S2A). Moreover, the differentially expressed genes were also enriched in the cellular component including

nuclear chromosome part as well as in the molecular function including damaged DNA binding, cell adhesion molecule binding, single-stranded DNA binding and ribonucleoprotein complex binding. KEGG pathways enrichment analysis suggested that the differentially expressed genes were mainly related to pyrimidine metabolism (Additional file 2: Figure S2B). These results demonstrated that differentially expressed genes might correlated with carcinogenesis-related processes [41]. In addition, GO and KEGG enrichment analysis were also performed for the target genes of differentially expressed miRNAs. The target genes were found to be mainly enriched in netrin-activated signaling pathway, DNA-binding transcription activator, single-stranded RNA binding, MAP kinase activity and so on (Additional file 3: Figure S3A). KEGG analysis suggested that the target genes were enriched in Glycerolipid metabolism, Platinum drug resistance, EGFR tyrosine kinase inhibitor resistance and Endocrine resistance (Additional file 3: Figure S3B). In addition, most of the differentially methylated CpG sites were found to locate in the gene body regions (Additional file 4: Figure S4), and 5 CpG sites were found to locate in the TSS1500 (sequence region from -200 to -1500 nt upstream of the transcription start site) and TSS200 (sequence region -200 nt upstream of the transcription start site) region of genes (Additional file 4: Figure S4, Additional file 7: Table S2).

1.4 Machine learning based construction of TPM

To develop TPM, we firstly generated 4 possible prediction biomarker signatures including the gene signature, miRNA signature, CpG site signature and multi-omics signature, which were composed of expression profiles of top 45 differentially expressed genes, top 45 differentially expressed miRNAs and top 45 differentially methylated CpG sites between TMB-H and TMB-L patients respectively (Additional file 8: Table S3). Using 10-fold cross-validation method and two-class logistic regression (type.measure = 'auc'), we then implemented the LASSO logistic analysis to select the optimal signature based on the expression or methylation profiles of the genes, miRNAs and CpG sites from the training cohort. The optimal biomarkers for the 4 prediction biomarker signature were obtained with non-zero regression coefficients (Figure 6, Table 2), and as a result, the multi-omics signature with maximum measure (0.868) was selected as the optimal biomarker signature for predicting the TMB (Figure 6D, Table 2). PCA using the shrunk multi-omics signature suggested that TMB-H patients and TMB-L patients could be separated obviously (Additional file 5: Figure S5). Based on the multi-omics signature, we finally constructed TPM by weighing the expression or methylation of the genes, miRNAs and CpG sites through the regression coefficients from the LASSO analysis (Table 3). The TPM was showed as the following math formula:

$$\text{TPM score} = -1.555696454 * \text{cg02031308} - 0.939485314 * \text{cg03286742} - 0.532855695 * \text{cg04046889} - 1.603385472 * \text{cg12095807} - 1.171295176 * \text{cg16794961} - 1.341848062 * \text{cg24553235} + 0.203290638 * \text{YBX2} + 0.000323171 * \text{HLTF} + 0.355814358 * \text{KLC3} + 0.017454209 * \text{WRNIP1} + 0.010739241 * \text{CKS1B} + 0.013056543 * \text{RNF26} + 0.039397451 * \text{ZYG11A} + 0.582628142 * \text{hsa-miR-571} + 3.954182602 * \text{hsa-miR-586} + 0.068239671 * \text{hsa-miR-151b} + 0.000724033 * \text{hsa-miR-378i} + 0.25824073 * \text{hsa-miR-6727-5p} - 0.731679875 * \text{hsa-miR-502-3p} - 0.007119299 * \text{hsa-miR-6798-3p}.$$

AUC of the constructed TPM in the training cohort was 0.911 showing its superior predictive accuracy (Figure 7A). Besides, the p-value of

two-side t-test was 3.40×10^{-48} between TPM score and TMB (Figure 7B), which suggested that TPM score was highly correlated with TMB in patients with LUAD.

1.5 Evaluation of the predicting accuracy of TPM in the validation cohort

To evaluate the predicting efficacy of the TPM constructed in the training cohort, the validation cohort was used in the subsequent analysis. We used the expression or methylation profiles of genes, miRNAs and CpG sites in the patients from the validation cohort as input parameters for calculating the TPM score. According to the threshold of -3.366 , 41 patients from TMB-H group were predicted as TMB-H, and 66 patients from the TMB-L group were predicted as TMB-L. In summary, the TPM has a sensitivity of 0.911, specificity of 0.750 and accuracy of 0.805 in predicting the TMB in the validation cohort, and the PPV was 0.651, NPV was 0.943 (Additional file 9: Table S4). ROC analysis revealed the AUC of the TPM in the validation cohort was 0.859 (Figure 8A), and the p-value of the two-side t-test was 1.19×10^{-14} between the TPM score and TMB (Figure 8B). These results suggested that the TPM performed relatively high TMB-predicting accuracy.

4 Discussion

Immunotherapy has been demonstrated particularly successful in NSCLC treatment and is being adopted as a first-line treatment option worldwide [13, 22, 42]. Nevertheless, only a small portion of the unselected patients can benefit from immunotherapy [25, 43]. Therefore, biomarkers for patient selection become important to achieve effective therapy. TMB has been recognized as the effective prognostic biomarker in NSCLC patients according to the results from numerous clinical trials [22, 23, 44, 45]. Although the targeted NGS has been proved to be an alternative approach of WES for the prediction of TMB, the accuracy and simplicity of targeted NGS remain challenging as various parameters should be taken into consideration [46]. In this study, we developed a mathematic multi-omics model that could precisely predict the TMB in patients with LUAD through LASSO logistic regression analysis, and the prediction accuracy of the model was validated in an independent cohort with high sensitivity and specificity (Fig. 8). Furthermore, as the input parameter in this model includes expression profiles of 7 genes, 7 miRNAs, and the methylation profiles of 6 CpG sites, which could be obtained through quantitative real time-polymerase chain reaction (qRT-PCR). This model paved the way for the further development of the simplified qRT-PCR based clinical assay for TMB prediction.

The tumor microenvironment refers to the network of cells and structures that surround a tumor cell, and it consists of immune cells, mesenchymal cells, endothelial cells, extracellular matrix (ECM) molecules, and inflammatory mediators [47]. High TMB indicates the presence of more neoantigens in the tumor microenvironment, which promote the inflammatory response and result in the alteration of transcriptomic and epigenetic signature [47]. It has been proved that gene expression signatures in the tumor microenvironment were associated with the prognosis in NSCLC [48–51]. In agreement with

previous studies, the differentially expressed genes between TMB-H and TMB-L patients identified in this study were found to enrich in the immune-related damaged DNA binding, nuclear division, nuclear chromosome segregation, organelle fission, single-stranded DNA binding, ribonucleoprotein complex binding and pyrimidine metabolism (Additional file 2: Figure S2) [52–56]. The 7 genes used in constructing the TPM might be involved in the carcinogenesis, for instance, Y box binding protein 2 (YBX2) was differentially expressed between different subtypes of breast cancer and was one of RNA processing factors which contribute to subtype-specific splicing [57]. Meanwhile, it was found that LINC00958 promoted cell proliferation and migration in oral squamous cell carcinoma through miR-627-5p/YBX2 axis [58]. Helicase-like transcription factor (HLTF) was a tumor suppressor, the expression of HLTF was negatively correlated with colorectal cancer and its overexpression regulated the TGF- β /SMAD pathway to suppress the migration and invasion of CRC [59]. It was reported that the wild type alleles of kinesin light chain 3 (KLC3) Lys751Gln was significantly correlated with greater smoking intensity, and genetic variations may influence the progression of lung cancer [60]. Then, WRN helicase interacting protein 1 (WRNIP1) was inhibited by miR-22 to increase the radiosensitivity of small-cell lung cancer cells [61]. Besides, the expression of CDC28 protein kinase regulatory subunit 1 (CKS1B) in lung cancer cells developed the chemoresistance through Hsp90 and MEK1/2 pathway [62].

miRNAs expression in the tumor microenvironment plays a crucial role in mediating and controlling several immune and cell interactions, and convolute in the regulation of immune checkpoints, PD1 and PD-L1 [63]. It was reported that a 25 miRNA-based signature classifier could predict the TMB level with high accuracy [64]. A cluster of highly expressed miRNA including hsa-miR-492, hsa-miR-498, hsa-miR-320 were found to be correlated with tumorigenesis of retinoblastoma [65]. Moreover, the invasion, proliferation and migration of cervical cancer cells were found to be promoted by hsa-miR-6727-5p, which might play an important role in cervical cancer progress [66]. In this study, we mapped the differentially expressed miRNAs between TMB-H and TMB-L patients to their target genes, and enrichment analysis of the target genes suggested that DNA-binding transcription activator, single-stranded RNA binding, MAP kinase activity, Glycerolipid metabolism, Platinum drug resistance and EGFR tyrosine kinase inhibitor resistance related to lung cancer metabolism were affected in the tumor microenvironment. The 7 miRNAs used in constructing the TPM in this study include hsa-miR-571, hsa-miR-586, hsa-miR-151b, hsa-miR-378i, hsa-miR-6727-5p, hsa-miR-502-3p, hsa-miR-6798-3p. It was reported that hsa-miR-378i, hsa-miR-92a-3p, hsa-miR-93-5p and so on were important for identifying both colon and rectal cancer [67]. In previous study, it was demonstrated that gallbladder carcinoma metastasis was developed by lncRNA-HGBC through the activation of miR-502-3p-SET-AKT cascade [68].

Changes in DNA methylation is one of the most important epigenetic alteration in tumor microenvironment. A multicenter study in 15 hospitals suggested epigenomic profile based on a microarray DNA methylation signature (EPIMMUNE) could serve as an effective biomarker in predicting the outcomes of NSCLC patients treated with PD-1 inhibitors [69], and the FOXP1 could be associated with validated predictive biomarkers for better-selecting patients to benefit with immunotherapy [69]. The CpG sites signature also had a relatively high predictive performance (measure = 0.861) of TMB, suggesting its great value in NSCLC prognosis. Cg02849937 located in the TSS1500 region of C7orf13,

and its expression level was inverse associated with promoter methylation using whole-genome integrative analysis [70]. In addition, cg27281030 located in the TSS1500 region of NLRP12. Previous studies demonstrated that NLRP12 could regulate inflammation and tumor, and it is believed that hepatocellular carcinoma was negatively regulated by NLRP12 through the suppression of inflammation and proliferation of hepatocytes [71]. Besides, cg23179456 located in the TSS1500 region of ADCY4, which has been proved to be a biomarker for breast cancer [72].

Through multi-omics analysis, we integrated gene or miRNA expression and DNA methylation data to reflect the subtle alterations of the tumor microenvironment to precisely predict the TMB for better prognosis of patients with LUAD in immunotherapy. Fragments per kilobase per million mapped reads (FPKM) of YBX2, HLTF, KLC3, WRNIP1, CKS1B, RNF26, ZYG11A, RPM of hsa-miR-571, hsa-miR-586, hsa-miR-151b, hsa-miR-378i, hsa-miR-6727-5p, hsa-miR-502-3p, hsa-miR-6798-3p as well as beta-value of cg02031308, cg03286742, cg04046889, cg12095807, cg16794961, cg24553235 were extracted from the RNA-seq, miRNA-seq and Illumina HumanMethylation450 BeadChip respectively for calculating the TPM score, and the threshold of TPM for predicting the TMB level had been predetermined as -2.947 with the specificity of 0.887 and the sensitivity of 0.825 in the training cohort. Although the FPKM, RPM and beta value involved in the TPM were based on high-throughput sequencing or chip analysis, it is feasible to convert them to cycle threshold (Ct) value in qRT-PCR, and thus simplify the prediction of TMB by using benchtop qRT-PCR instrument. The conversion of FPKM in different samples to Ct values could be probably through the comparison of the targeted gene expression to reference gene expressions, such as actin and eukaryotic elongation factor (eEF), which have relative consistent expression under different tumor microenvironment, and the beta value of CpG sites could also be converted into Ct value though the quantitative MethyLight technology [73]. To our best knowledge, this is the first time to construct the TPM for patients with LUAD from multi-omics view.

5 Conclusion

In summary, the present study developed a multi-omics risk model with high specificity and sensitivity in predicting the TMB for patients with LUAD, and laid the base for a more simplified and cost-effective TMB test assay establishment. Nevertheless, this study was solely bioinformatics research, and clinical sample validation for the TPM had not been implemented. The training cohort and the validation cohort used in this study were relatively small in size and required further expansion to increase the accuracy.

Abbreviations

TMB: tumor mutation burden; WES: whole-exome sequencing; LUAD: lung adenocarcinoma; TCGA: The Cancer Genome Atlas; TMB-H: TMB-high; TMB-L: TMB-low; AUC: area under curve; qRT-PCR: quantitative real time-polymerase chain reaction; OS: overall survival; ICIs: immune checkpoint inhibitors; PD-1/PD-L1: programmed cell death-1/programmed cell death ligand-1; PFS: progression-free survival; FDA: Food and Drug Administration; NSCLC: non-small cell lung cancer; CTLA-4: cytotoxic T-lymphocyte-associated protein 4; NGS: next generation sequencing; MSK-IMPACT: Memorial Sloan Kettering-Integrated Mutation

Profiling of Actionable Cancer Targets; TCGA-LUAD:The Cancer Genome Atlas Lung Adenocarcinoma; TPM:TMB prediction model; SNPs:single-nucleotide polymorphisms; UCSC:University of California Santa Cruz; RPM:reads per million; BMI:Body Mass Index; GO:Gene Ontology; KEGG:Kyoto Encyclopedia of Genes and Genomes; LASSO:least absolute shrinkage and selection operator; PPV:positive predictive value; NPV:negative predictive value; PCA:principal component analysis; ROC:receiver operating characteristic; TSS1500:sequence region from – 200 to -1500 nt upstream of the transcription start site; TSS200:sequence region – 200 nt upstream of the transcription start site; ECM:extracellular matrix; EPIMMUNE:epigenomic profile based on a microarray DNA methylation signature; FPKM:Fragments per kilobase per million mapped reads; Ct:cycle threshold.

Declarations

8 Acknowledgments

We would like to thank TCGA project organizers as well as all study participants.

9 Authors' contributions

Conceptualization: PC, JW and DG; Funding acquisition and Supervision: DG; Methodology: PC, JW, MS, GZ and SZ; Writing – original draft: PC and JW; Writing – review & editing: PC, JW and MS. All authors have read and agreed to the published version of the manuscript.

10 Funding

This work was supported by National Natural Science Foundation of China (91739109, 81970053, 81570046, 81870045, and 81700054); Guangdong Provincial Key Laboratory of Regional Immunity and Diseases (2019B030301009); Shenzhen Municipal Basic Research Program (JCYJ20190808123219295 and JCYJ20170818144127727); and Interdisciplinary Innovation Team Project of Shenzhen University (843-00000325).

11 Availability of data and materials

All data are included in the article.

12 Ethics approval and consent to participate

Not applicable.

13 Consent for publication

All authors consent to publication.

14 Competing interests

The authors declare that there are no conflicts of interest.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* 2019;69(1):7–34.
2. Chen W, Sun K, Zheng R, Zeng H, Zhang S, Xia C, Yang Z, Li H, Zou X, He J. Cancer incidence and mortality in China, 2014. *Chin J Cancer Res.* 2018;30(1):1–12.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2018;68(6):394–424.
4. Siegel R, Ward E, Brawley O, Jemal A. Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. *CA Cancer J Clin.* 2011;61(4):212–36.
5. Takamori S, Toyokawa G, Takada K, Shoji F, Okamoto T, Maehara Y. Combination Therapy of Radiotherapy and Anti-PD-1/PD-L1 Treatment in Non-Small-cell Lung Cancer: A Mini-review. *Clin Lung Cancer.* 2018;19(1):12–6.
6. Meyers DE, Bryan PM, Banerji S, Morris DG. Targeting the PD-1/PD-L1 axis for the treatment of non-small-cell lung cancer. *Curr Oncol.* 2018;25(4):e324–34.
7. Luo W, Wang Z, Tian P, Li W. Safety and tolerability of PD-1/PD-L1 inhibitors in the treatment of non-small cell lung cancer: a meta-analysis of randomized controlled trials. *J Cancer Res Clin Oncol.* 2018;144(10):1851–9.
8. Li JX, Huang JM, Jiang ZB, Li RZ, Sun A, Lai-Han Leung E, Yan PY. Current Clinical Progress of PD-1/PD-L1 Immunotherapy and Potential Combination Treatment in Non-Small Cell Lung Cancer. *Integr Cancer Ther.* 2019;18:1534735419890020.
9. Khunger M, Jain P, Rakshit S, Pasupuleti V, Hernandez AV, Stevenson J, Pennell NA, Velcheti V. Safety and Efficacy of PD-1/PD-L1 Inhibitors in Treatment-Naive and Chemotherapy-Refractory Patients With Non-Small-Cell Lung Cancer: A Systematic Review and Meta-Analysis. *Clin Lung Cancer.* 2018;19(3):e335–48.
10. Brahmer J, Reckamp KL, Baas P, Crino L, Eberhardt WE, Poddubskaya E, Antonia S, Pluzanski A, Vokes EE, Holgado E, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer. *N Engl J Med.* 2015;373(2):123–35.

11. Herbst RS, Baas P, Kim DW, Felip E, Perez-Gracia JL, Han JY, Molina J, Kim JH, Arvis CD, Ahn MJ, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet*. 2016;387(10027):1540–50.
12. Remon J, Besse B, Soria JC. Erratum to: Successes and failures: what did we learn from recent first-line treatment immunotherapy trials in non-small cell lung cancer? *BMC Med*. 2017;15(1):82.
13. Aguilar EJ, Ricciuti B, Gainor JF, Kehl KL, Kravets S, Dahlberg S, Nishino M, Sholl LM, Adeni A, Subegdjo S, et al. Outcomes to first-line pembrolizumab in patients with non-small-cell lung cancer and very high PD-L1 expression. *Ann Oncol*. 2019;30(10):1653–9.
14. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Csoszi T, Fulop A, Gottfried M, Peled N, Tafreshi A, Cuffe S, et al. Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer. *N Engl J Med*. 2016;375(19):1823–33.
15. Landre T, Des Guetz G, Chouahnia K, Taleb C, Vergnenegre A, Chouaid C. First-line PD-1/PD-L1 inhibitor plus chemotherapy vs chemotherapy alone for negative or < 1% PD-L1-expressing metastatic non-small-cell lung cancers. *J Cancer Res Clin Oncol*. 2020;146(2):441–8.
16. Martinez P, Peters S, Stammers T, Soria JC. Immunotherapy for the First-Line Treatment of Patients with Metastatic Non-Small Cell Lung Cancer. *Clin Cancer Res*. 2019;25(9):2691–8.
17. Brahmer JR, Govindan R, Anders RA, Antonia SJ, Sagorsky S, Davies MJ, Dubinett SM, Ferris A, Gandhi L, Garon EB, et al. The Society for Immunotherapy of Cancer consensus statement on immunotherapy for the treatment of non-small cell lung cancer (NSCLC). *J Immunother Cancer*. 2018;6(1):75.
18. Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, Gadgeel SM, Hida T, Kowalski DM, Dols MC, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet*. 2017;389(10066):255–65.
19. Fehrenbacher L, Spira A, Ballinger M, Kowanetz M, Vansteenkiste J, Mazieres J, Park K, Smith D, Artal-Cortes A, Lewanski C, et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet*. 2016;387(10030):1837–46.
20. Hui R, Garon EB, Goldman JW, Leighl NB, Hellmann MD, Patnaik A, Gandhi L, Eder JP, Ahn MJ, Horn L, et al. Pembrolizumab as first-line therapy for patients with PD-L1-positive advanced non-small cell lung cancer: a phase 1 trial. *Ann Oncol*. 2017;28(4):874–81.
21. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 2015;348(6230):124–8.
22. Hellmann MD, Ciuleanu TE, Pluzanski A, Lee JS, Otterson GA, Audigier-Valette C, Minenza E, Linardou H, Burgers S, Salman P, et al. Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *N Engl J Med*. 2018;378(22):2093–104.

23. Samstein RM, Lee CH, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, Barron DA, Zehir A, Jordan EJ, Omuro A, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet.* 2019;51(2):202–6.
24. Allgauer M, Budczies J, Christopoulos P, Endris V, Lier A, Rempel E, Volckmar AL, Kirchner M, von Winterfeld M, Leichsenring J, et al. Implementing tumor mutational burden (TMB) analysis in routine diagnostics-a primer for molecular pathologists and clinicians. *Transl Lung Cancer Res.* 2018;7(6):703–15.
25. Buttner R, Longshore JW, Lopez-Rios F, Merkelbach-Bruse S, Normanno N, Rouleau E, Penault-Llorca F. Implementing TMB measurement in clinical practice: considerations on assay requirements. *ESMO Open.* 2019;4(1):e000442.
26. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al. TCGAblinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 2016;44(8):e71.
27. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* 2016;44(W1):W83–9.
28. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
29. Fortin JP, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics.* 2017;33(4):558–60.
30. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, Moran S, Heyn H, Vizoso M, Gomez A, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J Clin Oncol.* 2013;31(32):4140–7.
31. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, Robinson WP, Kobor MS. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin.* 2013;6(1):4.
32. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* 2013;8(2):203–9.
33. Ready N, Hellmann MD, Awad MM, Otterson GA, Gutierrez M, Gainor JF, Borghaei H, Jolivet J, Horn L, Mates M, et al. First-Line Nivolumab Plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer (CheckMate 568): Outcomes by Programmed Death Ligand 1 and Tumor Mutational Burden as Biomarkers. *J Clin Oncol.* 2019;37(12):992–1000.
34. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 2017;77(21):e108–10.
35. Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: An online resource for prediction of microRNA binding sites. *PLoS One.* 2018;13(10):e0206239.
36. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat*

- Genet. 2000;25(1):25–9.
37. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999;27(1):29–34.
 38. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16(5):284–7.
 39. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33(1):1–22.
 40. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77.
 41. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000;100(1):57–70.
 42. Gettinger S, Rizvi NA, Chow LQ, Borghaei H, Brahmer J, Ready N, Gerber DE, Shepherd FA, Antonia S, Goldman JW, et al. Nivolumab Monotherapy for First-Line Treatment of Advanced Non-Small-Cell Lung Cancer. *J Clin Oncol.* 2016;34(25):2980–7.
 43. Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, Felip E, van den Heuvel MM, Ciuleanu TE, Badin F, et al. First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. *N Engl J Med.* 2017;376(25):2415–26.
 44. Gandara DR, Paul SM, Kowanetz M, Schleifman E, Zou W, Li Y, Rittmeyer A, Fehrenbacher L, Otto G, Malboeuf C, et al. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nat Med.* 2018;24(9):1441–8.
 45. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, et al. Mutational landscape and significance across 12 major cancer types. *Nature.* 2013;502(7471):333–9.
 46. Patel ASD, Parikh P, Kumar M. Biomarkers in immuno-oncology: A review article. *International Journal of Molecular Immuno Oncology.* 2019;4(2):41–9.
 47. Altorki NK, Markowitz GJ, Gao D, Port JL, Saxena A, Stiles B, McGraw T, Mittal V. The lung microenvironment: an important regulator of tumour growth and metastasis. *Nat Rev Cancer.* 2019;19(1):9–31.
 48. Chen G, Dong Z, Wu D, Chen Y. Profiles of immune infiltration in lung adenocarcinoma and their clinical significant: A gene-expression-based retrospective study. *Journal of cellular biochemistry* 2020.
 49. Yue C, Ma H, Zhou Y. Identification of prognostic gene signature associated with microenvironment of lung adenocarcinoma. *PeerJ.* 2019;7:e8128.
 50. Li J, Li X, Zhang C, Zhang C, Wang H. A signature of tumor immune microenvironment genes associated with the prognosis of nonsmall cell lung cancer. *Oncol Rep.* 2020;43(3):795–806.
 51. Sun J, Zhang Z, Bao S, Yan C, Hou P, Wu N, Su J, Xu L, Zhou M. Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J Immunother Cancer* 2020, 8(1).

52. Slowikowski BK, Galecki B, Dyszkiewicz W, Jagodzinski PP. Decreased expression of cytochrome p450 1B1 in non-small cell lung cancer. *Biomed Pharmacother.* 2017;95:339–45.
53. Wassenaar CA, Ye Y, Cai Q, Aldrich MC, Knight J, Spitz MR, Wu X, Blot WJ, Tyndale RF. CYP2A6 reduced activity gene variants confer reduction in lung cancer risk in African American smokers—findings from two independent populations. *Carcinogenesis.* 2015;36(1):99–103.
54. Veale D, Kerr N, Gibson GJ, Kelly PJ, Harris AL. The relationship of quantitative epidermal growth factor receptor expression in non-small cell lung cancer to long term survival. *Br J Cancer.* 1993;68(1):162–5.
55. Scagliotti GV, Selvaggi G, Novello S, Hirsch FR. The biology of epidermal growth factor receptor in lung cancer. *Clin Cancer Res.* 2004;10(12 Pt 2):4227s–4232s.
56. Ogawa J, Iwazaki M, Inoue H, Koide S, Shohtsu A. Immunohistochemical study of glutathione-related enzymes and proliferative antigens in lung cancer. Relation to cisplatin sensitivity. *Cancer.* 1993;71(7):2204–9.
57. Stricker TP, Brown CD, Bandlamudi C, McNerney M, Kittler R, Montoya V, Peterson A, Grossman R, White KP. Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression. *PLoS Genet.* 2017;13(3):e1006589.
58. Chen F, Liu M, Yu Y, Sun Y, Li J, Hu W, Wang X, Tong D. LINC00958 regulated miR-627-5p/YBX2 axis to facilitate cell proliferation and migration in oral squamous cell carcinoma. *Cancer Biol Ther.* 2019;20(9):1270–80.
59. Liu L, Liu H, Zhou Y, He J, Liu Q, Wang J, Zeng M, Yuan D, Tan F, Zhou Y, et al. HLTF suppresses the migration and invasion of colorectal cancer cells via TGFbeta/SMAD signaling in vitro. *Int J Oncol.* 2018;53(6):2780–8.
60. Verde Z, Reinoso L, Chicharro LM, Resano P, Sanchez-Hernandez I, Rodriguez Gonzalez-Moro JM, Bandres F, Gomez-Gallego F, Santiago C. Are SNP-Smoking Association Studies Needed in Controls? DNA Repair Gene Polymorphisms and Smoking Intensity. *PLoS One.* 2015;10(5):e0129374.
61. Jiang W, Han X, Wang J, Wang L, Xu Z, Wei Q, Zhang W, Wang H. miR-22 enhances the radiosensitivity of small-cell lung cancer by targeting the WRNIP1. *J Cell Biochem.* 2019;120(10):17650–61.
62. Wang H, Sun M, Guo J, Ma L, Jiang H, Gu L, Wen H, Liao S, Chen J, Zeng B, et al. 3-O-(Z)-coumaroyloleanolic acid overcomes Cks1b-induced chemoresistance in lung cancer by inhibiting Hsp90 and MEK pathways. *Biochem Pharmacol.* 2017;135:35–49.
63. Iqbal MA, Arora S, Prakasam G, Calin GA, Syed MA. MicroRNA in lung cancer: role, mechanisms, pathways and therapeutic relevance. *Mol Aspects Med.* 2019;70:3–20.
64. Lv Y, Huang Z, Lin Y, Fang Y, Chen Z, Pan L, Zhang Y, Xu Z. MiRNA expression patterns are associated with tumor mutational burden in lung adenocarcinoma. *Oncoimmunology.* 2019;8(10):e1629260.
65. Zhao JJ, Yang J, Lin J, Yao N, Zhu Y, Zheng J, Xu J, Cheng JQ, Lin JY, Ma X. Identification of miRNAs associated with tumorigenesis of retinoblastoma by miRNA microarray analysis. *Childs Nerv Syst.* 2009;25(1):13–20.

66. Liu H, Pei G, Song M, Dai S, Wang Y. Influence of hsa-miR-6727-5p on the proliferation, apoptosis, invasion and migration of Caski, Hela and SiHa cervical cancer cells. *J BUON*. 2017;22(4):973–8.
67. Pellatt DF, Stevens JR, Wolff RK, Mullany LE, Herrick JS, Samowitz W, Slattery ML. Expression Profiles of miRNA Subsets Distinguish Human Colorectal Carcinoma and Normal Colonic Mucosa. *Clin Transl Gastroenterol*. 2016;7:e152.
68. Hu YP, Jin YP, Wu XS, Yang Y, Li YS, Li HF, Xiang SS, Song XL, Jiang L, Zhang YJ, et al. LncRNA-HGBC stabilized by HuR promotes gallbladder cancer progression by regulating miR-502-3p/SET/AKT axis. *Mol Cancer*. 2019;18(1):167.
69. Duruisseau M, Martinez-Cardus A, Calleja-Cervantes ME, Moran S, Castro de Moura M, Davalos V, Pineyro D, Sanchez-Céspedes M, Girard N, Brevet M, et al. Epigenetic prediction of response to anti-PD-1 treatment in non-small-cell lung cancer: a multicentre, retrospective analysis. *Lancet Respir Med*. 2018;6(10):771–81.
70. Etcheverry A, Aubry M, de Tayrac M, Vauleon E, Boniface R, Guenot F, Saikali S, Hamlat A, Riffaud L, Menei P, et al. DNA methylation in glioblastoma: impact on gene expression and clinical outcome. *BMC Genom*. 2010;11:701.
71. Udden SN, Kwak YT, Godfrey V, Khan MAW, Khan S, Loof N, Peng L, Zhu H, Zaki H. NLRP12 suppresses hepatocellular carcinoma via downregulation of cJun N-terminal kinase activation in the hepatocyte. *Elife* 2019, 8.
72. Fan Y, Mu J, Huang M, Imani S, Wang Y, Lin S, Fan J, Wen Q. Epigenetic identification of ADCY4 as a biomarker for breast cancer: an integrated analysis of adenylate cyclases. *Epigenomics*. 2019;11(14):1561–79.
73. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW. MethyLight: a high-throughput assay to measure DNA methylation. *Nucleic Acids Res*. 2000;28(8):E32.

Tables

Table 1. Clinical information of 522 TCGA-LUAD patients.

Variables	Statistics
Gender	
Male (%)	242 (46.4%)
Female (%)	280 (53.6%)
Age	
80~89 (%)	30 (5.8%)
70~79 (%)	150 (28.7%)
60~69 (%)	146 (28.0%)
50~59 (%)	83 (16.0%)
40~49 (%)	25 (4.8%)
30~39 (%)	2 (0.4%)
Not reported (%)	86 (16.3%)
Race	
White (%)	393 (75.3%)
Black or African American (%)	53 (10.2%)
Asian (%)	8 (1.5%)
American Indian or Alaska native (%)	1 (0.2%)
Not reported (%)	67 (12.9%)
Status	
Alive (%)	334 (64.0%)
Dead (%)	188 (36.0%)
Tumor stage	
I (%)	279 (53.4%)
II (%)	124 (23.8%)
III (%)	85 (16.3%)
IV (%)	26 (5.0%)
Not reported (%)	8 (1.5%)

LUAD, lung adenocarcinoma.

Table 2. The performance of 4 optimal biomarker signatures obtained by LASSO regression analysis.

Biomarker signature	Optimal biomarkers	Lambda.min	Measure
Gene	GTF2IRD1, FTSJ1, CHMP4B, KLC3, DMAC2, GIT1, SOHLH2, SYNGR3, SAP130, LRRC1, FN3KRP, POU4F1, ZNF526, KRT80, UBE2C, FOXE1, MEX3D, CIDECP1, PRR19, DHX16, FANCG, AC010632.1, AC019171.1	0.018	0.884
miRNA	hsa-miR-22-5p, hsa-miR-486-5p, hsa-miR-492, hsa-miR-561-5p, hsa-miR-151b, hsa-miR-3677-5p, hsa-miR-3923, hsa-miR-4425, hsa-miR-4434, hsa-miR-4536-5p, hsa-miR-4679, hsa-miR-5702, hsa-miR-6727-5p, hsa-miR-6858-5p, hsa-miR-7107-5p, hsa-let-7g-3p, hsa-miR-136-3p, hsa-miR-155-3p, hsa-miR-371a-3p, hsa-miR-491-3p, hsa-miR-432-3p, hsa-miR-574-3p, hsa-miR-3074-3p, hsa-miR-3622b-3p, hsa-miR-3679-3p, hsa-miR-3150b-3p, hsa-miR-4639-3p, hsa-miR-4655-3p, hsa-miR-6798-3p, hsa-miR-6847-3p	0.017	0.734
CpG site	cg01862650, cg02031308, cg02916472, cg07184316, cg07729440, cg10120778, cg10488199, cg11002952, cg20151576, cg20297017, cg20671274, cg21827634, cg22773522, cg23049130, cg25841348	0.015	0.845
Multi-omics	cg01862650, cg07729440, cg20671274, cg21827634, cg22773522, GTF2IRD1, FTSJ1, TTI1, CHMP4B, KLC3, HNRNPUL1, UBE2S, BCL2L12, SYNGR3, KRT80, FOXE1, AC006213.3, hsa-miR-22-5p, hsa-miR-492, hsa-miR-4536-5p, hsa-miR-6727-5p, hsa-miR-7107-5p, hsa-miR-136-3p, hsa-miR-3679-3p, hsa-miR-6816-3p	0.010	0.938

LASSO, least absolute shrinkage and selection operator, TMB, tumor mutation burden.

Table 3. Coefficient of each biomarker of multi-omics signature in LASSO model analysis.

	Biomarkers	Coefficient
Multi-omics	cg01862650	-1.818730262
	cg07729440	-6.256940647
	cg20671274	-0.911730577
	cg21827634	-3.867697524
	cg22773522	-2.460934084
	GTF2IRD1	0.039415926
	FTSJ1	0.012286235
	TTI1	0.042479337
	CHMP4B	0.009500278
	KLC3	0.454704618
	HNRNPUL1	0.015834511
	UBE2S	0.014978452
	BCL2L12	0.078730525
	SYNGR3	0.192724739
	KRT80	0.017563239
	FOXE1	0.011660062
	AC006213.3	0.175314579
	hsa-miR-22-5p	1.230322969
	hsa-miR-492	-0.185026622
	hsa-miR-4536-5p	7.457452192
	hsa-miR-6727-5p	0.560095048
	hsa-miR-7107-5p	-0.131067725
	hsa-miR-136-3p	0.93577811
	hsa-miR-3679-3p	-0.007961529
	hsa-miR-6816-3p	1.359432308

LASSO, least absolute shrinkage and selection operator.

Figures

Figure 1

Flowchart of the analysis process in this study. TMB, tumor mutation burden; TMB-H, TMB-high; TMB-L, TMB-low; PCA, principal component analysis; LASSO, least absolute shrinkage and selection operator; ROC, receiver operating characteristic.

Figure 2

Division of patients with LUAD into TMB-H and TMB-L subgroups. (a) The distribution of TMB in patients with LUAD; (b) Number of TMB-H and TMB-L patients with LUAD; (c) The distribution of TMB across different tumor stages. TMB, tumor mutation burden; LUAD, lung adenocarcinoma; TMB-H, TMB-high; TMB-L, TMB-low; OS, overall survival.

Figure 3

Multi-omics data obtained from TCGA for patients with LUAD. (a) 440 patients with LUAD were found having coupled WES, DNA methylation, RNA-seq and miRNA-seq data; (b) 148 patients were classified as TMB-H and 292 patients were classified as TMB-L. WES, whole-exome sequencing; TMB-H, TMB-high; TMB-L, TMB-low.

Figure 4

The landscape of tumor-infiltrating immune cells in TMB-H patients and TMB-L patients. (a) Relative proportions of infiltrating immune cells in TMB-H patients and TMB-L patients; (b) Correlation matrix of all the proportions of 6 detected immune cell types. TMB-H, TMB-high; TMB-L, TMB-low.

Figure 5

Characterization of the top50 differentially expressed genes, miRNAs and differentially methylated CpG sites between TMB-H and TMB-L patients. Volcano plot showed the differentially expressed genes (a) and miRNAs (c) or differentially methylated CpG sites (e) between TMB-H and TMB-L patients. The red dots represented upregulated genes, miRNAs or hypermethylated CpG sites; the blue dots represented downregulated genes, miRNAs or hypomethylated CpG sites; the black dots represented genes, miRNAs or CpG sites with no significantly differential expression or methylation. Hierarchical clustering heatmap of differentially expressed genes (b) and miRNAs (d) or differentially methylated CpG sites (f) between TMB-H and TMB-L patients. Orange indicated the upregulated genes, miRNAs or hypermethylated CpG sites; Blue indicated the downregulated genes, miRNAs or hypomethylated CpG sites. TMB-H, TMB-high; TMB-L, TMB-low.

Figure 6

LASSO regression analysis for 4 possible prediction biomarker signatures. 10-fold cross-validation in LASSO regression analysis for gene signature (a), miRNA signature (b), CpG site signature (c), and multi-omics signature (d). LASSO, least absolute shrinkage and selection operator; AUC, area under curve.

Figure 7

The performance of TPM in the training cohort. (a) ROC analysis of the TPM score in the training cohort; (b) The TPM score is highly correlated with TMB. TPM, TMB prediction model; AUC, area under curve.

Figure 8

The performance of TPM in the validation cohort. (a) AUC of ROC analysis was 0.859 showing the great predictive accuracy of TPM; (b) The TPM score is highly correlated with TMB with p-value = 1.19×10^{-14} . TPM, TMB prediction model; AUC, area under curve.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile9TableS4.docx](#)
- [Additionalfile8TableS3.docx](#)
- [Additionalfile7TableS2.docx](#)
- [Additionalfile6TableS1.docx](#)
- [Additionalfile5FigureS5.jpg](#)
- [Additionalfile4FigureS4.jpg](#)
- [Additionalfile3FigureS3.jpg](#)
- [Additionalfile2FigureS2.jpg](#)
- [Additionalfile1FigureS1.jpg](#)