Prediction of Overall Survival for Early-stage Non-small Cell Lung Cancer via a Novel 7-gene Prognostic Signature and Allele Frequency Deviation (AFD)

Aisha AL-Dherasi

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China. Department of Biochemistry, Faculty of Science, Ibb University, Ibb, Yemen

Yuwei Liao

Yangjiang Key Laboratory of respiratory diseases, Yangjiang People's Hospital, Guangdong province, P.R. China

Qi-Tian Huang

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Yichen Wang

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Rulin Hua

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Sultan AL Mosiab

Department of Computer Science and Technology, Sahyadri Science Collage, Kuvempu University, India

Jingkai Zhang

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Ying Yu

State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, 2005 Songhu Road, Shanghai 200438, P.R. China

Yu Zhang

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Xuehong Zhang

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Chao Huang

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Sufiyan Sufiyan

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Dongcen Ge

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Wanting Bai

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Ruimei Liu

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Yanyan Shao

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Yulong Li

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Haithm Mousa

Department of Clinical Biochemistry, College of Laboratory Diagnostic Medicine, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Leming Shi

State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, 2005 Songhu Road, Shanghai 200438, P.R. China

Dekang Lv

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China

Zhiguang Li

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China. Division of Genetic and Molecular Toxicology, National Center for Toxicological Research, Food and Drug Administration

Quentin Liu (✉ liuq9@mail.sysu.edu.cn)

Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China. <https://orcid.org/0000-0002-0999-9805>

Keywords: AFD, Gene signature, Lung cancer, Overall survival, Prognostic marker

Posted Date: September 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-80073/v1>

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

[Read Full License](#)

Prediction of Overall Survival for Early-stage Non-small Cell Lung Cancer via a Novel 7-gene Prognostic Signature and Allele Frequency Deviation (AFD)

Running title: Early-stage NSCLC Survival Prognosis Model

Aisha AL-Dherasi^{1,4,#}, Yuwei Liao^{7,#}, Qi-Tian Huang^{1,#}, Yichen Wang^{1,#}, Rulin Hua¹, Sultan AL-Mosiab⁵, Jingkai Zhang¹, Ying Yu³, Yu Zhang¹, Xuehong Zhang¹, Chao Huang¹, Sufiyan Sufiyan¹, Dongcen Ge¹, Wanting Bai¹, Ruimei Liu¹, Yanyan Shao¹, Yulong Li¹, Haithm Mousa⁶, Leming Shi³, Dekang Lv¹, Zhiguang Li^{1,2}, Quentin Liu^{1,*}

Author Affiliations:

- 1 Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China.
- 2 Division of Genetic and Molecular Toxicology, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079.
- 3 State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, 2005 Songhu Road, Shanghai 200438, P.R. China.
- 4 Department of Biochemistry, Faculty of Science, Ibb University, Ibb, Yemen
- 5 Department of Computer Science and Technology, Sahyadri Science Collage, Kuvempu University, India.
- 6 Department of Clinical Biochemistry, College of Laboratory Diagnostic Medicine, Dalian Medical University, Dalian 116044, Liaoning, P.R. China.
- 7 Yangjiang Key Laboratory of respiratory diseases, Yangjiang People's Hospital, Guangdong province, P.R. China

#These authors contributed equally to this work.

***Correspondence author**

Quentin Liu Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, Liaoning, P.R. China.

Email: liuq9@mail.sysu.edu.cn

Zhiguang Li, Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, P.R.China.

Division of Genetic and Molecular Toxicology, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079

Email: zhiguangli88@gmail.com

Dekang Lv, Center of Genome and Personalized Medicine, Institute of Cancer Stem Cell, Dalian Medical University, Dalian 116044, P.R.China

Email: dekanglv@dmu.edu.cn

Abstract

Background: Due to the late and poor prognosis of non-small lung cancer(NSCLC), the mortality of patients is high, underlines the need to identify a credible prognostic marker for NSCLC patients. The aim of our study is to examine the association of allele frequency deviation (AFD) with the patient's survival, as well as identification and validation of a new prognostic signature to predict NSCLC overall survival(OS).

Methods: First, we developed a new algorithm to calculate AFD from whole-exome sequencing(WES) data, then we compared the predictability of the patient's survival between AFD, tumor mutation burden (TMB) and change of variants allele frequency (dVAF). Second, we overlapped the differentially expressed genes (DEGs) from our data with the genes associated with the survival of The Cancer Genome Atlas (TCGA) database to confirm all genes significantly related to the survival of lung cancer. We identified 149 genes, 31 of which are new genes and have not been reported for lung cancer, that was used to develop a new prognostic model. Lung cancer adenocarcinoma (LUAD) data from the TCGA database was used to validate the gene-signature model. The prognostic model relating to the genes was established and validated in training and LUAD validation groups.

Results: There was a significant association found between the high AFD value and poor survival among non-small cell lung cancer (NSCLC) patients. A novel seven genes (UCN2, RIMS2, CAVIN2, GRIA1, PKHD1L1, PGM5, CLIC6) were obtained through multivariate Cox regression analysis and significantly associated with NSCLC patients survival. Cox regression analysis confirmed that AFD and 7-gene signature are an independent prognostic marker in NSCLC patients. The AUC for 5-year survival in AFD and the AUC for 3-year survival in both training and validation groups were greater than 0.7.

Conclusion: As a result, AFD and 7-gene signatures were identified as new independent predictive factors used for predicting the survival among NSCLC patients.

Keywords:

AFD, Gene signature, Lung cancer, Overall survival, Prognostic marker.

Background

Despite of any advancement in lung cancer treatment, it remains one of the most common types and the leading cause of cancer-associated mortality among men and women worldwide[1]. Non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) are identified as two major types of lung cancer. The two main types of non-small cell lung cancer are lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD)[2], so these histological subtypes may determine the choice of treatment [2,3]. In recent years, the absolute and relative frequencies of lung cancer's incidence and mortality have risen dramatically around the world [4,5]. Overall, the five-year survival rate for lung cancer is 19% [6]. A total of 228,820 new lung cancer cases and 135,720 lung cancer deaths were expected to occur in 2020 [7].

However, more than half of NSCLC patients are diagnosed with either advanced or metastatic stage(third or fourth stage) disease, significant and longer survival rates can be obtained for those who are diagnosed at an early stage, but in advanced stages, curative treatment options are prolonged and limited, resulting in poor prognosis and low survival rates [8]. Time is certainly the crucial factor for all cancer patients, in addition to the fact that NSCLC is a heterogeneous group of diseases, all these reasons have led to the creation of a clear unmet medical need for the new marker that can aid the clinicians through facilitating the accurate and early diagnosis of lung cancer, enhancing predictive clinical outcomes, and guide the customized treatment [9,10].

Recently, a variety of studies have been conducted to identify predictive biomarkers to guide long-term NSCLC patients prognosis. Such biomarkers are segregated into: 1) single biomarker like SLC2A1, PKM, EPCAM, ALCAM, CADM1, HIF1A, and PTK7 [11-17], as well as molecular biomarker such as Tumor mutation burden(TMB), blood Tumor Mutation Burden(bTMB)[18], or some of the other recent markers currently being studied to predict prognostic condition or metastasis; and 2) gene markers that are found through the study of high-throughput gene expression profiles, they are built through using several prognostic genes. Many studies have demonstrated TMB as a biomarker for NSCLC patients [19], for example, Rizvi et al.[20] demonstrated that high TMB levels were correlated with improved ORR, prolonged PFS in retrospective analysis of NSCLC patients. Chae YK et al.[21] reported that treating with PD-1/PD-L1 inhibitors leads to a longer OS among patients

with high TMB in comparison to those with low TMB. Owada-Ozaki Y et al.[22] confirmed that high TMB is a poor prognosis factor in NSCLC patients. In addition, dVAF has also been identified as a predictor of clinical outcomes at NSCLC and UC[23]. Moreover, several studies began to identify gene biomarkers related to NSCLC prognosis [24-26]. Another study on AFD involved cervical cancer patients revealed that AFD is positively correlated with therapy response and helps in expecting the progression-free survival²⁷.

Based on these results, the investigation of a new prognosis biomarker for OS in early-stage NSCLC patients will be critical in the rapid assessment of diagnosis and therapeutic efficacy. Therefore, our study is considered the first to study the direct association of AFD with prognostication, as there is a lack of studies that paid due attention to the comparison of TMB and dVAF with AFD in order to predict the progression of NSCLC patients. We have concluded that AFD is an independent prognostic factor to be used to predict the survival of NSCLC patients (primary aim), we reached this by developing a new algorithm to determine the deviation of tumor from normal allele frequency through WES of tumor and normal samples. Moreover, we analyzed the RNA.seq data to identify the NSCLC-related gene biomarkers for better identification of a reliable gene signature relevant to NSCLC prognosis. This approach has helped us to recognize a novel 7-gene signature which can be used for successfully prediction of prognostic risk in NSCLC patients (secondary aim).

METHODS

Data source

Raw data including Whole Exome Sequencing (WES), RNA sequencing (RNA.seq), and relevant clinical information (including survival information) of 102 NSCLC Patients have been obtained from Fudan University. For the validation group, the data related to gene expression and clinical information of LUAD patients was downloaded from the TCGA dataset, consisted of a total 594(535 tumor sample and 59 normal samples) adenocarcinoma cases. The main characteristics of the analysis included the following: OS, age, tumor size, sex, pT stage, pathologic_stage and history of smoking; details of patient clinical information is described in **Table 1**.

Around 48% of the sample are males while 52% are females, the participating age ranged from 37 to 84 years, with a median age of 61.5 years. A total of 102 white blood cell (WBC) samples, 102 tumor samples for the same patients were included in RNA.seq analysis, 54 samples out of 102 were enrolled in WES analysis. The raw data of this study were obtained from Fudan University. The process of carrying out the data analysis for the study has been carried out according to the ethical standards (Fudan University Shanghai Cancer Center Institutional Review Board No. 090977-1). Collecting the samples from patients was conducted by the tissue bank of Fudan University Shanghai Cancer Center after having the consent of patients or their relatives [70].

Alignment and quality control

In-house pipelines were used to process the sequencing of 54 WES data. Tumor and normal samples quality data were evaluated using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), including sequence length distribution, GC content, aspect of per-base quality, sequence duplicate levels, Kemer content and over-represented sequences [27]. Sequencing readings were aligned with the human reference genome (hg38) using the Burrows-Wheeler Aligner (BWA) software package with the default parameters [28], and we removed the reads that were mapped in multiple genome positions. Then we assessed the quality of the map using SAMtools flagstat [29]; and All the genome sites for somatic variants were called by using VarScan [30] software with parameters of base quality higher than 30 and supporting reads ≥ 200 (**Supplementary Figure1**).

Calling of SNV from the whole-exome sequence

After mapping all the readings to the human reference genome (hg38) using (BWA)[28], Picard 1.67 was used to mark the duplicate readings that re-aligned around the known indels. We performed the base quality recalibration by using GATK version 3.7 [31]. Somatic mutations were called using Mutect2 when each of the following criteria was met: first, the frequency of allele variants was 1% in tumor samples and 1% in standard samples; second, both standard and tumor samples had sequencing coverage of 200; third, the alternative readings in tumor samples were 10; fourth, the corrected p-value was less than 0.05. SNVs have been annotated by using

ANNOVAR in multiple databases[32] and further filtered with population frequency in ExAC,1000 Genomes,dbSNP138,clinvar_20170130 and avsnp147.

Allele Frequency Deviation (AFD)

Allele frequency was calculated for 54 samples and WBC was used as a control to avoid the factors that could affect the AF at site of tumor sample. As displayed in **supplementary figure 2**, first, we made a scatter plot for all detected genomic sites of a patient, with Y axis representing AF of a tumor sample and X axis representing the AF of a paired normal sample. Second, we made a diagonal line, on which the points have the same AF between tumor and normal samples. The distance from each point to this diagonal line is calculated and defined as d_i of the i -th point. Third, we transposed the X,Y coordinates by -45° , so that d_i is equal to the absolute value of the Y axis of i point and can be calculated by this formula :

$$d_i = |y_i'| = |x_i * \sin\left(-\frac{\pi}{4}\right) + y_i * \cos\left(-\frac{\pi}{4}\right)|$$

The y_i' is the transposed Y-axis value of the i point, the x_i , y_i is the original X-axis and Y-axis value. Fourth, we made vertical lines across the X-axis so that they are divided into bins of equal distances. In the same way, we divided the Y-axis into equidistant bins. Thus, the coordinate system plane is divided into m grids of equal sizes (Supplementary Figure2. d-f). And then, we calculated $\bar{d}_k$ by averaging the distance values ($y'_{i,i \in \text{bin}m}$) of all points in a grid. And the density ρ_k of a grid is defined as the proportion of the number of points in this grid to the number of all points.

ρ_k equal points number in bin k /all point, point density

$$\rho_i = \rho_{k,k \in \text{bin}k}$$

By using density, we can avoid outlier and noise that would be given less weight due to their low density, in this case the AFD can reflect patient clinical situation.

Finally, the AFD of a patient was calculated by :

$$AFD = \frac{\sum_{i=1}^n d_i \rho_i}{n}$$

Where d_i represent distance values of all points in a grid, ρ_i represent density of grid, n represent the total number of grid.

Tumor Mutation Burden (TMB) calculation

The number of nonsynonymous single-nucleotide variants, insertion or deletion variants is known as the burden of tumor mutation [33]. It is known that WES is the gold and standard method to determine TMB [34], it is used to calculate TMB from tumor samples and their matched WBC normal samples. We used quantile method based on measurements of TMB to divided patients into quantiles [35].

Variant allele frequencies changes (dVAF) calculation

We calculated variant allele frequencies(VAF) of somatic mutations for all genes in WES for tumor samples and its corresponding WBC in our cohort consisting of 54 patients. Non-synonymous and synonymous variants were included in this calculation, synonymous mutations were considered purposefully to reduce sampling noise [36]. Only variants observed in normal samples were used for the tumor mean VAF calculation. The change in mean VAF(dVAF) was calculated as:

$$dVAF=(\text{meanVAF}_{\text{tumor}})-(\text{meanVAF}_{\text{normal}})$$

The association between the AFD and dVAF was tested. The statistical analysis of dVAF was performed to assess its ability to predict the progression of overall survival.

RNA.seq analysis

For Gene expression analysis of lung cancer RNA-seq data and TCGA dataset, the reads were mapped against the human genome(hg38) using STAR2 software [37]. The mapped reads with quality more than 10 were selected using Samtools. The read counts per gene were defined using featurecount [38] as the reference transcriptome. Differential expression analysis was performed using edgeR[39], comparing tumor samples to their matched normal samples. The selected genes are the ones statistically significant differentially expressed between tumor and normal samples and their FDR <0.05 and $\text{abs}(\log\text{FC}) > 1$ (**Supplementary Figure 3**).

Construction of a prognostic gene signature

We first determined the candidate prognostic genes which are significantly associated with OS, and select the common ones for constructing the gene signature model. An univariate Cox proportional hazard regression analysis as well as lasso regression were implemented with each gene to identify the relationship of genes with OS in the data[40]. P-value less than 0.05 was used as a cutoff to define and select the candidate genes related to patients survival. Finally, a multi Cox proportional hazard regression analysis was carried out to select the final list of genes- related to OS and then create the model of prognostic signature. The hazard ratio from the analysis of multivariate cox regression was used to assess the protective genes(HR < 1) and risky genes(HR > 1). The risk scoring for each patient was then estimated using the following equation to calculate the expression values pertaining to the selected genes weighted by regression coefficients in multivariate cox regression analysis.

$$\text{RiskScore} = \sum_{i=1}^n \text{Exp}_i * C_i^{\text{HR}}$$

Where n is the number of selected prognostic genes, Exp_i is the expression value of the prognostic gene i and C_i^{HR} represents the estimated regression coefficient for the corresponding gene i in the multivariate cox regression analysis. Subsequently, the median prognostic score was used to differentiate between the high-risk and low-risk groups. The patients with lower risk than median value were assigned to a low-risk group while the others were assigned to a high-risk group. The prognostic performance of the prognostic score model was measured by using the ROC curve by comparing the area under the respective receiver operating characteristic curve. Finally, the prognostic score model was examined to check its association with the survival of the NSCLC patient.

Statistical analysis

The Spearman correlation test was analyzed among the variables. The statistical difference in OS was determined by one-sided log-rank test. Survival curves have been estimated using the Kaplan-Meier method, and p-value was calculated. We conducted univariate and multivariate Cox proportional hazard regression models in order to adjust AFD, TMB, dVAF and 7-gene signature model with various variables in order to evaluate the impact of each variable on patients

survival. Cox proportional-hazards model was used to estimate the Hazard Ratios and 95% confidence intervals. The comparison of power between the factors was carried out using the ROC curve analysis. The p-value < 0.05 was defined as a nominal level of statistical significance. All statistical analyses were performed through using the version 3.5.1 of R language environment.

RESULTS

Patients characteristic

The median age of NSCLC patients during diagnosis was 61.5 years which is ranged from 37 to 84 years. Adenocarcinoma was the histological subtype for all patients in the current study. 48% (n=49) of our sample group were males and 52% females(n=53). Output status for all patients was 0 or 1, Seventy one(70%) were never smoked patients, and 31(30%) were former/current smokers, fifty nine patients(58%) had stage IA, 27 patients(26.4%) had stage IB, and sixteen patients(15.6%) had stage IIIA. No information was provided on treatment(**Table1**).

Relationship between factors (AFD, TMB, and dVAF)

(**Supplementary Figure 4.a**) shows the correlations between AFD and TMB in NSCLC patients. Based on the Spearman correlation coefficient, we observed that the p-value of the test is more than the significance level 0.05, therefore, the AFD and TMB are not significantly associated at a correlation coefficient 0.16 and p-value of 0.26 . On the other hand, almost a very weak significant correlation was observed between AFD and dVAF ($r = 0.27$, $p = 0.05$) in all patients in the identified survival study group(**Supplementary Figure 4.b**). Moreover, there was a weak significant correlation between TMB and dVAF ($r = 0.3$, $p = 0.03$) (**Supplementary Figure 4.c**). Overall, these findings indicate that the AFD values are significantly different and independent from the other two factors.

Allele frequency deviation shows an active power to predict patient outcomes

To evaluate the sensitivity and specificity of AFD, TMB and dVAF for OS predicting in NSCLC patients, we implemented a time-dependent ROC curve.

The TMB and AFD significantly achieved almost same AUC values 0.721 and 0.713 respectively(**Figure 1. a-right & b-right panel**). In contrast, the AUC for dVAF was not high(0.56), which means that the dVAF remains a challenge, this creates the need for more verifications of this factor through future investigation. In particular, AUC values of more than 0.7 for AFD in these data demonstrate a high OS predictive performance, and AUC values of less than 0.6 for dVAF show poor OS predictive performance(**Figure 1.c-right panel**). These results demonstrated that AFD has the power to predict the overall survival which is reflected by the AUC performance.

Overall survival

According to the fact that TMB, dVAF and AFD are continuous variables, and the cutting points for these variables are still not uniformly established, in this study, we divided the patients into quantile(quantile method) according to the measurements of each AFD, TMB, and dVAF. The mean value of AFD was 13.74(range From 0.15 to 33.18); it was 19.81(the range 2.5 to 32.97) for TMB, while the dVAF value was ranging from 0.014 to 1.24 with the mean of 0.091(**Table 1**), in regards of this method, the AFD cut-off point at 75 percent quantile was 17.93, 22.028 mutation/Mb and 0.0599 for TMB and dVAF respectively, which divided the patients into high and low-value groups, the mean follow-up duration was 44.9 months(3.6-72.7 months) for OS and 36.7 months(1-72.7 months) for RFS(**Table 1**).

Kaplan-Meier estimated the overall survival at 31 months as 89.7 percent(95 % confidence interval[CI], 80.6 to 99.8) in the low-value AFD group and 64.3 percent(95% CI, 43.5 to 95) in the high-value AFD group(**Table 2**). It is observed that there is a gradual decrease in survival from 78.6% at 12 months in the high-value group to 52.2 % at 35 months. Overall survival was found significantly longer in the low-value AFD group than in the high-value group, with 10 % lower risk of death and 42.8% higher risk of death for both low and high-value groups respectively(hazard ratio for death, 1.10; 95 % CI, 1.01 to 1.2, $p=0.03$). Patients for the high-value and low-value groups included in the survival analysis according to their cutoff points were 14 and 40 respectively(**Figure 1 a-left panel**). The one-sided stratified log-rank P-value was 0.0064, indicating a significant difference between the two groups as well as an increase in genetic variants in high-value AFD groups with any increase in the AFD

value in patients; therefore, the risk of death in this group of patients increased. In contrast, in the group of AFD low-value, the decrease of AFD values reflects the low genetic variants, so the death risk observed was lower than in the other group. Besides, the median of overall survival could not be estimated in both groups of AFD. On the other hand, the Kaplan-Meier curve for TMB showed that high-level Patients had significantly shorter overall survival than the low-level patients, with 35.7% higher risk of death(hazard ratio, 1.08; 95%CI, 0.96 to 1.2, $p=0.17$). Thus, overall survival was 62.5% at 31 months with(95% confidence interval [CI], 41 to 95.3)inhigh-level TMB and 89.9% (95%CI, 80.9 to 99.8) in low-level TMB group. The number of patients in the high-level group is 40 patients and 14 patients in the low-level group(**Table2**). Notably, the one-sided stratified long-range P-value was 0.03, which indicates the difference between the two groups in regards to overall survival. Similarly, in the AFD, the median overall survival of the high-level group and low-level group could not be estimated(**Figure 1. b-left panel**).

We also performed the Kaplan-Meier estimation for the dVAF. Our finding showed that there is no statistical significance between the high-level and low-level groups of dVAF. The p-value was 0.17 due to the one-side stratified long-rank estimation(**Figure 1. c-right panel**). The total number of patients in every group was the same while the number of patients in the case of AFD groups was 14 and 40 patients for the high and low-level groups respectively. Moreover, the Kaplan-Meier estimation revealed that patients in low and high-level of dVAF groups have the overall survival at 13 months 92.5%(95% CI, 84.7 to 100) and 78.6% (95% confidence interval, 59.8 to 100) respectively, while the overall survival at 35 months was 84.4%(95% CI, 73.6 to 69.7) for the low-level of dVAF group and in the high-level group, there is no event at 35 months, therefore, the overall survival could not be estimated(**Table 2**).

The Allele frequency deviation is an independent prognostic factor.

Herein, we conducted univariate and multivariate cox regression models in the NSCLC data. The AFD, TMB, and dVAF with other clinicopathological factors, including gender, smoking, age, pT, and tumor-size were used as covariates. **Table 3** illustrates the association between OS and these three factors. It was found through univariate regression analysis that age, sex, pT stage, and smoking were not significantly associated with the OS of NSCLC patients. Notably, the TMB in this analysis

presented remarkably poor prognostic ability to predict the overall survival($p= 0.17$). While the AFD and tumor size were identified to have a significant correlation with the OS($p= 0.03$ and 0.04 respectively) of NSCLC patients, the dVAF still showed a poor prognostic ability to predict the OS as in the KM analysis($p= 0.7$). Moreover, to assess whether the AFD was independent from TMB and dVAF and other clinical variables, a multivariate Cox regression analysis was carried out. The results showed that AFD had a significant independent predictive ability for OS of NSCLC patients($HR= 1.16$, 95% CI 1.027- 1.32, $P= 0.01$), While each of the two other factors TMB and dVAF were not related to the survival of patients($HR= 1.03$, 95% CI 0.85-1.25, $P= 0.75$) and ($HR= 0.35$, 95% CI 0.0065 -19.40, $P= 0.61$) for TMB and dVAF respectively after adjusting for other clinical factors. These findings confirm that AFD can be used as a predictive factor for the overall survival of NSCLC patients(**Table 3**).

Identification of gene sets associated with overall patient survival

For NSCLC training set data, we used the KM analysis to establish the association between the gene expression and overall survival. We identified 409 genes associated with overall survival, we downloaded the TCGA data to screen and confirm the genes that really associated with survival in NSCLC, we identified 1177 genes. By overlapping the two datasets, 149 genes log-rank P value < 0.01 were identified to be associated with NSCLC survival. Of those, 31 genes have not been reported in NSCLC patients, which were used to conduct the next analyses to develop a prognostic signature model(**Supplementary Figure 3**).

Identification of a 7-gene prognostic signature

To identify NSCLC gene signature association with OS, we selected unreported 31 prognosis- related genes which associated with survival, followed by univariate Cox regression and lasso regression analysis for further selection, we identified 24 genes then 11 genes from both analysis respectively. Out of these genes, 7 genes were finally identified and established using a multivariate cox regression analysis to participate in overall survival, which was used to construct a prognostic model, and a multivariate cox regression analysis was conducted using a stepwise regression method(**Supplementary Figure 3**). The formula(1) is as follows:

$$\text{Risk} = (-0.3658 * \text{ExpGRIA1}) + (0.5701 * \text{ExpUCN2}) + (-0.601 * \text{ExpPKHD1L1}) + (0.2192 * \text{ExpRIMS2}) + (-0.3617 * \text{ExpPGM5}) + (-0.6036 * \text{ExpCLIC6}) + (1.1686 * \text{ExpCAVIN2})$$

The information related to 7 genes is shown in the **Table 4**. Finally, a set of 7 genes including (n=3) the risky gene (HR>1) and (n=4) the protective genes (HR<1) were examined. **Table 4** displayed the prognostic correlation of 7 genes with the NSCLC patients survival.

The validation of 7-gene prognostic signature

Based on the gene expression as well as regression coefficients of the 7 genes from the multivariate cox analysis, we built a prognostic model for predicting the prognosis using the risk score approach. A risk score for each patient was given in the prognostic model. The median risk score of (0.7334 and 0.9367) were used as the cutoff points to classify the patients into high-risk and low-risk groups in the training and LUAD-TCGA validation groups respectively. (**Figure 2. B&E**) shows the predictive power of overall survival through a 7-gene signature for patients. (**Figure 2 C&F**) showed the distribution of the gene risk score, the level of gene expression, and the survival status of patients in both data.

Patients who belong to the high-risk group were found to have a significantly shorter OS than patients belonging to the low-risk group, as shown in Kaplan-Meier curves, with 29.4% higher risk and 3.9% lower risk of death for high and low risk groups respectively (HR = 1.042, 95% CI, 1.02 to 1.06, P = 0.0002) (**Table 3 & Table 5**). The P-value of one-side stratified log-rank was 0.00037, confirming a significant difference between the high and low-risk groups, therefore, the clinical outcomes of patients in the low-risk group are found to be better than those in the high-risk group (**Figure 2-A**). The overall survival at 13 months was 98% (95% CI, 94.2 to 100) and 84.3% (95% CI 74.9 to 94.9) in the low and high-risk groups respectively; 68.6% (95% CI, 56.4 to 83.5) in the high-risk group at 31 months (**Table 5**).

For the TCGA validation group, Kaplan-Meier curves showed that overall survival was significantly longer in the low-risk group compared to the high-risk group, with 23.8% lower risk and 47.9% higher risk of death in the low-risk and high-risk groups, respectively (HR = 2.01, 95% CI, 1.57 to 2.557, P < 0.001) (**Table 5**). The single-sided stratified log-rank p-value was 0.0001, indicating the difference

between the two groups (**Figure 2-D**). The median overall survival at 36 months was 49.9 percent(95% CI, 42.95-58.1) for the high-risk group and 48.7 percent (95% CI, 38.29-61.8) for the low-risk group at 77 months (**Table 5**). Furthermore, to assess the 7-gene signature sensitivity and specificity for the overall survival prediction in both datasets, the time-dependent ROC curve was conducted . Markedly, in training and TCGA validation groups, the AUC value of 7-gene signatures was 0.883 and 0.70, respectively(**Figure 2-B&E**), suggesting a strong OS prediction efficiency.

Clinical independence of the 7-gene signature model

To evaluate the contribution of the 7-gene signature as an independent prognostic biomarker in the NSCLC training data and LUAD TCGA data set, univariate and multivariate Cox regression models were implemented. The 7-gene signature and other clinicopathological factors, including gender, age, stage, tumor size, smoking were included as covariates. **Table 3** shows the correlation between the OS and these factors. It was found out through univariate regression analysis that risk score, stage and tumor size were significantly associated with patients survival in NSCLC training set(**Figure 3. A**); risk score, T, N, M and stage were identified to have significantly correlation with OS of LUAD-TCGA validation set(**Figure 3.B**) and(**Table 3**). Interestingly, the corresponding multivariate cox regression analysis revealed that the 7-gene signature predictive ability was found significantly independent of other clinical factors for OS of NSCLC patients in training(HR= 1.036, 95% CI, 1.01033 to 1.063, p= 0.006) (**Figure 3.C**) and LUAD-TCGA validation(HR= 1.082, 95% CI, 1.011 to 1.158, p= 0.02) groups (**Figure 3.D**) (**Table 3**). These results indicate that our prognostic model of 7-gene signature is a highly prognostic independent biomarker and it presents independent predictive performance through clinical application.

Discussion

When considering prognosis, NSCLC is believed to be an extremely heterogeneous disease where survival time among patients differs based on their pathological stages. Traditional clinicopathological variables like TNM level, tumor size, sex, age, as well as tumor factors such as cell differentiation, vascular invasion, and vascularity have been used in a broad framework to predict patient outcomes for diagnosis and

treatment of NSCLC patients at early stages. Predicting outcomes was found to be insufficient due to the difference in effectiveness from among treatment strategy [41-45]. Consequently, an inspection of molecular prognostic markers that reliably represent the biological traits of tumors should be crucial for NSCLC patients treatment as well as for individualized prevention.

In the current study, we first developed a new algorithm to determine the genetic deviation of tumor allele frequency from the allele frequency of normal samples(WBC) which is called allele frequency deviation(AFD). And by AFD we can fairly detect any genetic alteration of tumor sample that deviated from the normal sample at each single site of the WES. AFD is an independent factor which is significantly associated with the overall survival of patients. Second, we analyzed the DEGs of 102 NSCLC cancer samples from Fudan University and identified a novel 7 unreported genes that were significantly associated with OS. This 7-gene signature was found to be capable of identifying NSCLC patients who may be at high or low risk during the prognosis. This signature is substantially independent of other clinical factors and could be crucial in deciding the treatment of NSCLC patients at an early stage.

Previous studies have shown that TMB is significantly correlated with the Immune Checkpoint Inhibitor (ICI) such as PD-L1 and PD-1, as well as other biomarkers like EGFR and TP53 [46-48]. In our research, we evaluated the relationship between AFD and TMB as well as dVAF and found that there was no correlation between AFD and the other two factors. Further, the AUC of the prediction for patients survival in both AFD and TMB was high and almost the same while AUC in dVAF was lower, suggesting that AFD had a substantial capacity not less than the capacity of TMB to predict OS. In addition, these results are consistent with our findings in the Kaplan-Meiren analysis of NSCLC patients, where AFD showed more significance than TMB in OS prediction and patients were also defined into high risk group and low-risk group, the patients with high AFD value had shorter OS compared to the low AFD group. On the contrary, in both univariate and multivariate analysis, TMB tended to be a non-independent prognostic factor for predicting NSCLC survival, and there was no significant association between TMB and OS, which is consistent with previous studies [49,50]. Interestingly, AFD displayed a prognostic ability in both univariate and multivariate cox regression analysis, and emerged as an independent prognostic factor.

A number of studies have reported that tumor size is a prognostic factor used to predict patient progression and outcomes [51]. A Previous study related to AFD demonstrated the effectiveness of AFD in predicting the benefit and response of cervical cancer patients to treatment and the predicted evidence of metastases was better than tumor size [27]. AFD performed in our study was independent of the tumor size, patients with high AFD had worse progression compared to patients with low AFD values. As a result, we concluded that AFD has the potential ability to predict patient survival, consequently suggesting the use of AFD for clinical purpose instead of tumor size due to its accuracy. In contrast to our finding, *Rajiv Raja et al.* [23] has reported a close correlation of dVAF with clinical outcomes in NSCLC and UC. Our result has shown that AFD is strongly correlated with patients survival better than dVAF, where dVAF has not shown statistically significant predictability to predict NSCLC patient's survival.

Recently, molecular biomarkers and gene signatures occupied a great deal of interest from researchers and are used in clinical practice for many aspects of cancer including tumorigenesis, progression and prognosis [52]. Gene signatures [53] as well as TMB and bTMB [54,55] classify patients as high-risk and low-risk groups. In this aspect, *Shuguang Zuo et al.* [56] identified a 6-gene signature, however their AUC was 0.749, 0.685 and 0.667 in the three independent datasets GSE31210, GSE37745 and GSE50081 respectively. *Ru He et al.* [57] identified an 8-gene signature, and their AUC was 0.726, 0.725, 0.701 and 0.650 in GSE31210, GSE50081, GSE37745 and TCGA respectively. Furthermore, *Hui Xie et al.* [58] identified a 6-gene signature based on integrated analysis and weight gene coexpression network. The area under curve(AUC) was 0.99 and 0.82 or 0.77 and 0.75 in predicting one to ten years survival of TCGA-LUAD and GSE11969 datasets respectively. On the contrary, the AUC of our 7-gene signature was higher with using 7 genes, this makes it appropriately suitable for clinical application.

The seven genes in our signature consist of UCN2, RIMS2 and CAVIN2 as risk factors, and GRIA1, PKHD1L1, PGM5 and CLIC6 as protective factors. It has been reported that CLIC6 is a member of the intracellular chloride channels consisting one of the dopamine receptor-mediated signaling pathways and has changed its expression in breast cancer [59,60]. And there are no previous reports related to the prognosis of patients' cancer outcomes. *Chen Zheng et al.* [61] has been reported that PKHD1L1 may be a PTC-associated tumor suppressor gene and maybe a potential molecular

biomarker useful as a therapeutic target in the coming years. PGM5 has been reported as a diagnostic and prognostic biomarker independently associated with the survival of patients with liver cancer [62] and colorectal cancer [63]. *Sloane K Tilley, et al.* [64] reported that increased expression and hypermethylation of GRIA1 was correlated with survival patients with Basal-Like Bladder Cancer and was used as a Prognostic Biomarker. Another report for *Guodong Yang, et al.* [65] showed that GRIA1 is one of the top 10 target genes in the protein-protein interaction network present in the five-miRNA signature model used as a novel prognosis biomarker and therapeutic target for CRC patients. *Silvia Codenotti, et al.* [66] reported that Cavin2 is a useful marker for discriminating the degree of differentiation in liposarcoma(LPS) tumors. Another study conducted by *Bayader Annabi, et al.* [67] highlights the role of cavin2 in the regulation of each inflammatory and angiogenic for TNF-activated MSC. Where there are no previous reports related to the prognosis of cancer outcomes in patients. *Stephane Esnault, et al.* [68] reported that ucn2 had the downstream function of inflammation, tissue remodeling, and lipid synthesis in human lung fibroblasts(HLFs). Where no previous survival prediction reports have been reported for cancer patients. RIMS2 has been reported to be mutated in melanoma [69] and there are no other reports related to the prediction of cancer patient's outcomes.

7-gene signature and AFD are still a new biomarker and factor respectively and has not yet been used as a prognostic marker for the prediction of clinical outcomes in lung cancer or any other type of cancer, moreover, the 7 genes are novel genes in lung cancer have not been reported . That is said our study is the first to demonstrate that 7-gene signature is an independent prognostic biomarker and AFD is also an independent prognostic factor in NSCLC, this is due to that AFD is more comprehensive to identify not only the selected mutations but also any genetic alteration at every single site of the chromosome. In addition, these results may indicate a more fundamental role in AFD efficacy in early cancer detection and accurate survival prediction. Moreover, our findings suggest that the 7-gene signature can be used as a new NSCLC prediction biomarker that affects the prognosis of the tumor. However, our study has limitations, first, inadequate sample size, since we have only analyzed the data obtained from Fudan University and only 102 out of 204 patients have clinical follow-up information, which impacts subgrouping the data while conducting the analysis, therefore, the stage was not considered as parameter in case of AFD analysis. Second, confirmation of applying bioinformatics on genes need to be verified experimentally due to

insufficient results. These limitation can be avoided through conducting studies with larger samples. Third, AFD prognostic factor and a 7-gene signature model were applied in lung cancer more research is needed to validate them in different types of cancer.

Conclusion

In our study, we developed a new algorithm to calculate AFD as well as developed a 7-gene signature prognostic model, both of which have high AUC values, and they are independent of other clinical features. The results suggested that each of AFD and 7-gene signature powerfully predicts overall survival of NSCLC patients. Therefore, it is recommended that each AFD and gene model score to be used as a molecular diagnostic test to evaluate the prognosis risk in lung cancer patients.

List of abbreviations

AFD: Allele Frequency Deviation

AUC: Area Under Curve

bTMB: blood Tumor Mutation Burden

CAVIN2: Caveolae Associated Protein 2

CLIC6: Chloride Intracellular Channel 6

CI: Confidence Interval

CRC: Colorectal Cancer

dVAF: Change of Variants Allele Frequency

DEGs: Differentially Expressed Genes

FDR: False Discovery Rate

GRIA1: Glutamate Ionotropic Receptor AMPA Type Subunit 1

HLFs: human lung fibroblasts

HR: Hazard Ratio

LPS: liposarcoma

LUAD: Lung Adenocarcinoma

LUSC: Lung Squamous Cell Carcinoma

NSCLC: Non-Small Lung Cancer

OS: Overall Survival

PGM5: Phosphoglucomutases

PKHD1L1: Polycystic Kidney and Hepatic Disease 1-Like 1

RIMS2: Regulating Synaptic Membrane Exocytosis 2

SCLC: Small Cell Lung Cancer

TCGA: The Cancer Genome Atlas

TMB: Tumor Mutation Burden

UCN2: Urocortin 2

WES: Whole-exome Sequencing.

Ethics declarations

Ethical approval and consent to participate

The data analysis process was conducted according to the ethical standards (Fudan University Shanghai Cancer Center Institutional Review Board No. 090977-1). Informed consents of patients or their relatives were obtained while donating a samples to the tissue bank of Fudan University Shanghai Cancer Center [70].

Consent for publication

Not applicable.

Competing Interests

The authors declare no competing interests.

Data Availability

The raw data used and/or analysed during the current study are available in the European Genome-phenome Archive (EGA) with the accession code EGAS00001004006.

Funding

This work was supported by National Natural Science Foundation of China (No. 81872655, 81820108024, 31801100, 81672784, 81602200, 81630005 and 81472637), the Pandeng Scholar Program from the Department of Education of Liaoning Province (to Dr. Zhiguang Li), FONDECYT 1180241, CONICYT-FONDAP 15130011, IMII P09/016-F (GIO) and startup funds from Dalian Medical University (to Dr. Zhiguang Li). The funders had no role in the design of the study, the collection and analysis of the data, the decision to publish, or the preparation of the manuscript.

Authors contributions:

AA analyzed the data and interpreted the results for both data (Whole Exome Sequencing data and RNA Sequencing data); LS, YY generated the data; AA, YWL and YW were responsible for the developing an algorithm; QH analyzed the RNA sequencing data; YZ and XZ helped with data analysis; RH and SA wrote some part of codes in R language; JZ, CH, SS, DG, WB, HM, YS, RL and YL made contributions to the final revision; DL, ZL and QL guided the research, revised the manuscript and final approval of the manuscript; AA wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge the support of National Natural Science Foundation of China (No. 81872655, 81820108024, 31801100, 81672784, 81602200, 81630005 and 81472637), the Pandeng Scholar Program from the Department of Education of Liaoning Province (to Dr. Zhiguang Li), FONDECYT 1180241, CONICYT-FONDAP 15130011, IMII P09/016-F (GIO) and startup funds from Dalian Medical University (to Dr. Zhiguang Li).

REFERENCES

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin.* 2015;65(2):87-108. doi:10.3322/caac.21262.
2. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. *Nature.* 2018;553(7689):446-454. doi:10.1038/nature25183.
3. Hou S, Zhou S, Qin Z, et al. Evidence, Mechanism, and Clinical Relevance of the Transdifferentiation from Lung Adenocarcinoma to Squamous Cell Carcinoma. *Am J Pathol.* 2017;187(5):954-962. doi:10.1016/j.ajpath.2017.01.009.

4. 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. doi:10.3322/caac.21492.
5. Lin HT, Liu FC, Wu CY, Kuo CF, Lan WC, Yu HP. Epidemiology and Survival Outcomes of Lung Cancer: A Population-Based Study. *Biomed Res Int.* 2019;2019:8148156. Published 2019 Dec 28. doi:10.1155/2019/8148156.
6. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* 2019;69(1):7-34. doi:10.3322/caac.21551.
7. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(1):7-30. doi:10.3322/caac.21590.
8. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization Classification of Lung Tumors: Impact of Genetic, Clinical and Radiologic Advances Since the 2004 Classification. *J Thorac Oncol.* 2015;10(9):1243-1260. doi:10.1097/JTO.0000000000000630.
9. Gandhi L, Rodríguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *N Engl J Med.* 2018;378(22):2078-2092. doi:10.1056/NEJMoa1801005.
10. Soria JC, Ohe Y, Vansteenkiste J, et al. Osimertinib in Untreated EGFR-Mutated Advanced Non-Small-Cell Lung Cancer. *N Engl J Med.* 2018;378(2):113-125. doi:10.1056/NEJMoa1713137.
11. Zhang WC, Shyh-Chang N, Yang H, et al. Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis [published correction appears in *Cell.* 2012 Mar 2;148(5):1066. Mitchell, Wayne [added]]. *Cell.* 2012;148(1-2):259-272. doi:10.1016/j.cell.2011.11.050.
12. Chen R, Khatri P, Mazur PK, et al. A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer Res.* 2014;74(10):2892-2902. doi:10.1158/0008-5472.CAN-13-2775.
13. Baeuerle PA, Gires O. EpCAM (CD326) finding its role in cancer [published correction appears in *Br J Cancer.* 2007 May 7;96(9):1491]. *Br J Cancer.* 2007;96(3):417-423. doi:10.1038/sj.bjc.6603494.
14. Papadaki C, Sfakianaki M, Lagoudaki E, et al. PKM2 as a biomarker for chemosensitivity to front-line platinum-based chemotherapy in patients with metastatic non-small-cell lung cancer. *Br J Cancer.* 2014;111(9):1757-1764. doi:10.1038/bjc.2014.492.
15. Zeng D, Wu X, Zheng J, et al. Loss of CADM1/TSLC1 Expression Is Associated with Poor Clinical Outcome in Patients with Esophageal Squamous Cell Carcinoma. *Gastroenterol Res Pract.* 2016;2016:6947623. doi:10.1155/2016/6947623.
16. Barron CC, Bilan PJ, Tsakiridis T, Tsiani E. Facilitative glucose transporters: Implications for cancer detection, prognosis and treatment. *Metabolism.* 2016;65(2):124-139. doi:10.1016/j.metabol.2015.10.007.
17. Lau SK, Boutros PC, Pintilie M, et al. Three-gene prognostic classifier for early-stage non small-cell lung cancer. *J Clin Oncol.* 2007;25(35):5562-5569. doi:10.1200/JCO.2007.12.0352.
18. Gandara DR, Paul SM, Kowanetz M, 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):1441-1448. doi:10.1038/s41591-018-0134-3.

19. Wang C, Liang H, Lin C, et al. Molecular Subtyping and Prognostic Assessment Based on Tumor Mutation Burden in Patients with Lung Adenocarcinomas. *Int J Mol Sci.* 2019;20(17):4251. Published 2019 Aug 30. doi:10.3390/ijms20174251.
20. Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science.* 2015;348(6230):124-128. doi:10.1126/science.aaa1348.
21. Chae YK, Davis AA, Raparia K, et al. Association of Tumor Mutational Burden With DNA Repair Mutations and Response to Anti-PD-1/PD-L1 Therapy in Non-Small-Cell Lung Cancer. *Clin Lung Cancer.* 2019;20(2):88-96.e6. doi:10.1016/j.clcc.2018.09.008.
22. Owada-Ozaki Y, Muto S, Takagi H, et al. Prognostic Impact of Tumor Mutation Burden in Patients With Completely Resected Non-Small Cell Lung Cancer: Brief Report. *J Thorac Oncol.* 2018;13(8):1217-1221. doi:10.1016/j.jtho.2018.04.003.
23. Raja R, Kuziora M, Brohawn PZ, et al. Early Reduction in ctDNA Predicts Survival in Patients with Lung and Bladder Cancer Treated with Durvalumab. *Clin Cancer Res.* 2018;24(24):6212-6222. doi:10.1158/1078-0432.CCR-18-0386.
24. Xie H, Xie C. A Six-Gene Signature Predicts Survival of Adenocarcinoma Type of Non-Small-Cell Lung Cancer Patients: A Comprehensive Study Based on Integrated Analysis and Weighted Gene Coexpression Network. *Biomed Res Int.* 2019;2019:4250613. Published 2019 Dec 4. doi:10.1155/2019/4250613.
25. Sun R, Meng X, Wang W, et al. Five genes may predict metastasis in non-small cell lung cancer using bioinformatics analysis. *Oncol Lett.* 2019;18(2):1723-1732. doi:10.3892/ol.2019.10498.
26. Zuo S, Wei M, Zhang H, et al. A robust six-gene prognostic signature for prediction of both disease-free and overall survival in non-small cell lung cancer. *J Transl Med.* 2019;17(1):152. Published 2019 May 14. doi:10.1186/s12967-019-1899-y.
27. Tian J, Geng Y, Lv D, et al. Using plasma cell-free DNA to monitor the chemoradiotherapy course of cervical cancer. *Int J Cancer.* 2019;145(9):2547-2557. doi:10.1002/ijc.32295.
28. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010;26(5):589-595. doi:10.1093/bioinformatics/btp698.
29. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.
30. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568-576. doi:10.1101/gr.129684.111.
31. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.
32. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164. doi:10.1093/nar/gkq603.

33. Chalmers ZR, Connelly CF, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 2017;9(1):34. Published 2017 Apr 19. doi:10.1186/s13073-017-0424-2.
34. Berland L, Heeke S, Humbert O, et al. Current views on tumor mutational burden in patients with non-small cell lung cancer treated by immune checkpoint inhibitors. *J Thorac Dis.* 2019;11(Suppl 1):S71-S80. doi:10.21037/jtd.2018.11.102.
35. Hendriks LE, Rouleau E, Besse B. Clinical utility of tumor mutational burden in patients with non-small cell lung cancer treated with immunotherapy. *Transl Lung Cancer Res.* 2018;7(6):647-660. doi:10.21037/tlcr.2018.09.22.
36. Niyomnaitham S, Parinyanitikul N, Roothumnong E, et al. Tumor mutational profile of triple negative breast cancer patients in Thailand revealed distinctive genetic alteration in chromatin remodeling gene. *PeerJ.* 2019;7:e6501. Published 2019 Feb 25. doi:10.7717/peerj.6501.
37. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635.
38. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656.
39. Ali G, Bruno R, Poma AM, et al. Whole transcriptome targeted gene quantification provides new insights on pulmonary sarcomatoid carcinomas. *Sci Rep.* 2019;9(1):3536. Published 2019 Mar 5. doi:10.1038/s41598-019-40016-8.
40. Guo JC, Wu Y, Chen Y, et al. Protein-coding genes combined with long noncoding RNA as a novel transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. *Cancer Commun (Lond).* 2018;38(1):4. Published 2018 Apr 9. doi:10.1186/s40880-018-0277-0.
41. Alatorre CI, Carter GC, Chen C, et al. A comprehensive review of predictive and prognostic composite factors implicated in the heterogeneity of treatment response and outcome across disease areas. *Int J Clin Pract.* 2011;65(8):831-847. doi:10.1111/j.1742-1241.2011.02703.x.
42. Crinò L, Weder W, van Meerbeeck J, Felip E; ESMO Guidelines Working Group. Early stage and locally advanced (non-metastatic) non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol.* 2010;21 Suppl 5:v103-v115. doi:10.1093/annonc/mdq207.
43. Rami-Porta R, Bolejack V, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for the Revisions of the T Descriptors in the Forthcoming Eighth Edition of the TNM Classification for Lung Cancer. *J Thorac Oncol.* 2015;10(7):990-1003. doi:10.1097/JTO.0000000000000559.
44. Tas F, Ciftci R, Kilic L, Karabulut S. Age is a prognostic factor affecting survival in lung cancer patients. *Oncol Lett.* 2013;6(5):1507-1513. doi:10.3892/ol.2013.1566.
45. Radkiewicz C, Dickman PW, Johansson ALV, Wagenius G, Edgren G, Lambe M. Sex and survival in non-small cell lung cancer: A nationwide cohort study. *PLoS One.* 2019;14(6):e0219206. Published 2019 Jun 27. doi:10.1371/journal.pone.0219206.

46. Samstein RM, Lee CH, Shoushtari AN, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat Genet.* 2019;51(2):202-206. doi:10.1038/s41588-018-0312-8.
47. Li WY, Zhao TT, Xu HM, et al. The role of EGFR mutation as a prognostic factor in survival after diagnosis of brain metastasis in non-small cell lung cancer: a systematic review and meta-analysis. *BMC Cancer.* 2019;19(1):145. Published 2019 Feb 13. doi:10.1186/s12885-019-5331-z.
48. Jiao XD, Qin BD, You P, Cai J, Zang YS. The prognostic value of TP53 and its correlation with EGFR mutation in advanced non-small cell lung cancer, an analysis based on cBioPortal data base. *Lung Cancer.* 2018;123:70-75. doi:10.1016/j.lungcan.2018.07.003.
49. Wu HX, Wang ZX, Zhao Q, et al. Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Ann Transl Med.* 2019;7(22):640. doi:10.21037/atm.2019.10.116.
50. Marina Garassino, MD & Corey J. Langer, MD. In International Association for the Study of Lung Cancer (IASLC) 2019 World Conference on Lung Cancer (WCLC) in Barcelona on “Tumor Mutational Burden Disappoints as Biomarker for Treatment Response in Exploratory Analyses of Nonsquamous NSCLC” (2019).
51. Zhang J, Gold KA, Lin HY, et al. Relationship between tumor size and survival in non-small-cell lung cancer (NSCLC): an analysis of the surveillance, epidemiology, and end results (SEER) registry. *J Thorac Oncol.* 2015;10(4):682-690. doi:10.1097/JTO.0000000000000456.
52. Zhu CQ, Tsao MS. Prognostic markers in lung cancer: is it ready for prime time?. *Transl Lung Cancer Res.* 2014;3(3):149-158. doi:10.3978/j.issn.2218-6751.2014.06.09.
53. Wang J, Chen X, Tian Y, et al. Six-gene signature for predicting survival in patients with head and neck squamous cell carcinoma. *Aging (Albany NY).* 2020;12(1):767-783. doi:10.18632/aging.102655.
54. Kim J, Kim B, Kang SY, et al. Tumor Mutational Burden Determined by Panel Sequencing Predicts Survival After Immunotherapy in Patients With Advanced Gastric Cancer. *Front Oncol.* 2020;10:314. Published 2020 Mar 13. doi:10.3389/fonc.2020.00314.
55. Gandara DR, Paul SM, Kowanetz M, 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):1441-1448. doi:10.1038/s41591-018-0134-3.
56. Zuo S, Wei M, Zhang H, et al. A robust six-gene prognostic signature for prediction of both disease-free and overall survival in non-small cell lung cancer. *J Transl Med.* 2019;17(1):152. Published 2019 May 14. doi:10.1186/s12967-019-1899-y.
57. He R, Zuo S. A Robust 8-Gene Prognostic Signature for Early-Stage Non-small Cell Lung Cancer. *Front Oncol.* 2019;9:693. Published 2019 Jul 31. doi:10.3389/fonc.2019.00693.
58. Xie H, Xie C. A Six-Gene Signature Predicts Survival of Adenocarcinoma Type of Non-Small-Cell Lung Cancer Patients: A Comprehensive Study Based on Integrated Analysis and Weighted Gene Coexpression Network. *Biomed Res Int.* 2019;2019:4250613. Published 2019 Dec 4. doi:10.1155/2019/4250613.
59. Low SK, Chin YM, Ito H, et al. Identification of two novel breast cancer loci through large-scale genome-wide association study in the Japanese population. *Sci Rep.* 2019;9(1):17332. Published 2019 Nov 22. doi:10.1038/s41598-019-53654-9.

60. Ko JH, Ko EA, Gu W, Lim I, Bang H, Zhou T. Expression profiling of ion channel genes predicts clinical outcome in breast cancer. *Mol Cancer*. 2013;12(1):106. Published 2013 Sep 22. doi:10.1186/1476-4598-12-106.
61. Zheng C, Quan R, Xia EJ, Bhandari A, Zhang X. Original tumour suppressor gene polycystic kidney and hepatic disease 1-like 1 is associated with thyroid cancer cell progression. *Oncol Lett*. 2019;18(3):3227-3235. doi:10.3892/ol.2019.10632.
62. Jiao Y, Li Y, Jiang P, Han W, Liu Y. PGM5: a novel diagnostic and prognostic biomarker for liver cancer. *PeerJ*. 2019;7:e7070. Published 2019 Jun 11. doi:10.7717/peerj.7070.
63. Sun Y, Long H, Sun L, et al. PGM5 is a promising biomarker and may predict the prognosis of colorectal cancer patients. *Cancer Cell Int*. 2019;19:253. Published 2019 Oct 1. doi:10.1186/s12935-019-0967-y.
64. Tilley SK, Kim WY, Fry RC. Analysis of bladder cancer tumor CpG methylation and gene expression within The Cancer Genome Atlas identifies GRIA1 as a prognostic biomarker for basal-like bladder cancer. *Am J Cancer Res*. 2017;7(9):1850-1862. Published 2017 Sep 1.
65. Yang G, Zhang Y, Yang J. A Five-microRNA Signature as Prognostic Biomarker in Colorectal Cancer by Bioinformatics Analysis. *Front Oncol*. 2019;9:1207. Published 2019 Nov 12. doi:10.3389/fonc.2019.01207.
66. Codenotti S, Vezzoli M, Poliani PL, Cominelli M, Monti E, Fanzani A. Cavin-2 is a specific marker for detection of well-differentiated liposarcoma. *Biochem Biophys Res Commun*. 2017;493(1):660-665. doi:10.1016/j.bbrc.2017.08.135.
67. Annabi B, Zgheib A, Annabi B. Cavin-2 Functions as a Suppressive Regulator in TNF-induced Mesenchymal Stromal Cell Inflammation and Angiogenic Phenotypes. *Int J Stem Cells*. 2017;10(1):103-113. doi:10.15283/ijsc16032.
68. Esnault S, Bernau K, Torr EE, Bochkov YA, Jarjour NN, Sandbo N. RNA-sequencing analysis of lung primary fibroblast response to eosinophil-degranulation products predicts downstream effects on inflammation, tissue remodeling and lipid metabolism. *Respir Res*. 2017;18(1):188. Published 2017 Nov 10. doi:10.1186/s12931-017-0669-8.
69. Zhang D, Xia J. Somatic synonymous mutations in regulatory elements contribute to the genetic aetiology of melanoma. *BMC Med Genomics*. 2020;13(Suppl 5):43. Published 2020 Apr 3. doi:10.1186/s12920-020-0685-2.
70. Chen H, Carrot-Zhang J, Zhao Y, et al. Genomic and immune profiling of pre-invasive lung adenocarcinoma. *Nat Commun*. 2019;10(1):5472. Published 2019 Nov 29. doi:10.1038/s41467-019-13460-3

Figure legends :

Figure1. Performance of the AFD and other two factors in NSCLC patients. **(a)** Kaplan-Meier survival curve distribution of AFD (left panel) and ROC curve and AUC for the 5-year survival of NSCLC samples (right panel). **(b)** Kaplan-Meier survival curve of TMB (left panel) and ROC curve for the 5-year survival of NSCLC samples (right panel). **(c)** Kaplan-Meier survival curve of dVAF (left panel) and ROC curve for the 5-year survival of NSCLC samples (right panel).

Figure 2. Relation between the overall survival among patients and the seven-gene signature in two datasets. **(A)** Kaplan-Meier survival curve of the training group samples. **(B)** The receiver operating characteristic(ROC) curve for the three years survival of training group. **(C)** correlation between the risk score and expression of the 7-gene signature in training samples. **(D)** Kaplan-Meier survival curve for the LUAD-TCGA validation group samples. **(E)** ROC curve for the three years survival of LUAD-TCGA validation group. **(F)** correlation between the risk score and the expression of the 7-gene signature in LUAD-TCGA validation samples.

Figure 3. Forest plots of HRs for overall survival in NSCLC training group(a, b) and (c,d) LUAD-TCGA validation group.

Additional files:

Supplementary Figure 1: Whole exome sequencing analysis flowchart

Supplementary Figure 2: Calculation of Allele Frequency Deviation

- (a) Qualified distribution for every sites of allele frequency (AF) in normal cells should be lies around wild type (0%) , heterozygous(50%) and homozygous (100%)
- (b) Diagonal line on each point that have the same AF in both tumor and normal samples.
- (c) Transposition of X and Y coordinates by -45° .
- (d-f) Vertical lines across X and Y axis

Supplementary Figure 3. RNA sequencing analysis flowchart

Supplementary Figure 4. Spearman Correlation between the three factors
(a) correlation between AFD and TMB (b) correlation between AFD and dVAF (c)
correlation between TMB and dVAF

Figures

Figure 1

Performance of the AFD and other two factors in NSCLC patients. (a) Kaplan-Meier survival curve distribution of AFD (left panel) and ROC curve and AUC for the 5-year survival of NSCLC samples (right panel). (b) Kaplan-Meier survival curve of TMB (left panel) and ROC curve for the 5-year survival of

NSCLC samples (right panel). (c) Kaplan-Meier survival curve of dVAF (left panel) and ROC curve for the 5-year survival of NSCLC samples (right panel).

Figure 2

Relation between the overall survival among patients and the seven-gene signature in two datasets. (A) kaplan-Meier survival curve of the training group samples. (B) The receiver operating characteristic(ROC) curve for the three years survival of training group. (C) correlation between the risk score and expression

of the 7-gene signature in training samples. (D) kaplan-Meier survival curve for the LUAD-TCGA validation group samples. (E) ROC curve for the three years survival of LUAD-TCGA validation group. (F) correlation between the risk score and the expression of the 7-gene signature in LUAD-TCGA validation samples.

Figure 3

Forest plots of HRs for overall survival in NSCLC training group(a, b) and (c,d) LUAD-TCGA validation group.

Supplementary Files

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

- [SupplementaryFigure4.pdf](#)
- [supplementaryfigure3.pdf](#)
- [Supplementaryfigure2.pdf](#)
- [supplementaryfigure1.pdf](#)
- [Table5.docx](#)

- [TABLE2.doc](#)
- [TABLE3.doc](#)
- [TABLE4.doc](#)
- [TABLE1.doc](#)