IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 22, 2020, accepted May 2, 2020, date of publication May 7, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993118

A New TTZ Feature Extracting Algorithm to Decipher Tobacco Related Mutation Signature Genes for the Personalized Lung Adenocarcinoma Treatment

QIEN HE[1], ZHEWEI QIU[1,2], YIFAN TONG[1], AND KAI SONG[1]
[1]School of Chemical Engineering and Technology, Tianjin University, Tianjin 300350, China
[2]School of Life Sciences, Tsinghua University, Beijing 100089, China

Corresponding author: Kai Song (ksong@tju.edu.cn)

This work was supported by the National Key Research and Development Program of China under Grant 2018YFC0808600.

ABSTRACT The big percentage of lung adenocarcinomas (LUAD) arising in lifetime nonsmokers and the low sensitivities of known major tobacco biomarkers urgent the identification of real molecular signatures for corresponding personalized treatment. Moreover, cancer is presumed to have a symptomatology strongly dependent on modules of functionally-related genes rather than on a unique important gene. Our aims, therefore, are to identify signature genes by optimizing the tobacco exposure pattern (TEP) classification model and to uncover their interaction relationships at different molecular levels. A new method, TTZ, is proposed to extract features as input variables to TEP classification model. Based on the Z-curve method, TTZ is able to extract features not only from mutation frequencies but also from sequencing information of insertions and deletions. Two independent LUAD datasets, The Cancer Genome Atlas (TCGA) and Broad data, are downloaded to train and test the TEP classification model. Thirty-four genes are identified as tobacco related mutational signature genes with the accuracies of 93.55% and 92.65% for train and validation data, respectively. The inference of genetic and protein-protein interaction (PPI) networks uncover that *LAMA1*, *EGFR*, *KRAS* and *TNN* are the most connected core genes. Six signature genes are proved significantly involved in the cilium damage pathway, which is considered as one of the root causes of lung cancer. The identified signature genes may serve as potential drug targets for the precision medicine of LUAD. Most importantly, the TTZ feature extracting method can be easily extended to other disease or cancer related mutational signature identification issues.

INDEX TERMS Biological network inference, lung adenocarcinomas, mutation signature identification, tobacco exposure, Z-curve method.

I. INTRODUCTION

Lung cancer has been the leading cause of cancer-related mortality throughout the world for decades [1]. Cigarette smokers are proved to be 15-30 times more likely to get lung cancer or die from it than lifetime nonsmokers. It is linked to about 80% to 90% of lung cancers in United States. Even though tobacco smoking is the major risk for lung cancer, however, there are still 10-15% of cancer patients of western world who have no history of tobacco exposure [2], [3]. Most of them tend to suffer lung adenocarcinoma (LUAD) [4], [5]. More importantly, it's been shown that smoking during cancer therapy may influence radiotherapy and chemotherapy outcomes [6]. Therefore, when considering therapies for LUAD patients, the carcinogenic mechanisms of smokers are believed to differ from those of nonsmokers [7]–[9]. Unfortunately, more and more well-known major mutations have been proved high false positive or high false negative by accumulated research results. Taken the two well-known major mutations frequently present in LUADs, *KRAS* and *EGFR* mutations, as examples: Riely *et al.* [10] found that *KRAS* mutations in LUADs occurred at a frequency of only 25% in smokers but at a frequency of as high as 15% in nonsmoker; By contrast, in a large meta-analysis study, Ren *et al.* [11]

The associate editor coordinating the review of this manuscript and approving it for publication was Qin Ma.

found that *EGFR* mutations in NSCLC were approximately only 5 times more common in nonsmokers than in smokers. Consequently, the rising proportion of nonsmokers in LUAD urges the precision treatment for the patients with different tobacco exposure, which further urges the deep understanding of the difference in the carcinogenic mechanisms between smokers and nonsmokers.

Somatic DNA mutations are relatively stable and are believed to lead to initiation and progression of many types of cancer. It has been proved that tobacco exposure results in cancer risk by increasing the somatic mutation load, both in number and type [9], [12]. The average mutation frequency is more than 10-fold higher in smokers than in nonsmokers. Uncovering signature mutation genes for the difference in tobacco exposure pattern (TEP) holds high promise for the deep understanding of the difference in the carcinogenic mechanisms between smokers and nonsmokers. Consequently, it holds high promise for fasting development of personalize treatment for LUAD.

To uncover TEP related mutation signatures, various efforts have been made to incorporate cancer-specific mutation information into analysis. Most of these tools can be classified into three categories based on their basic principles. 1) Frequency-based methods: identifying signature genes that are more frequently mutated than the background mutation rate [9], [13]–[15]. 2) Subnetwork methods: identifying signature genes based on prior knowledge of pathways, proteins or genetic interactions [16], [17]. 3) Hotspot-based methods [18]–[20]. The term hotspot refers to hotspot mutation regions, which are driven by positive selection and especially located in functional domains or important residues for three-dimensional protein structures [21], [22].

However, despite the rapid progress in computational approaches to prioritize cancer mutational signature genes with the advent of next generation sequencing technologies, the ultimate goal of discovering a complete catalog of genes truly associated with TEP is far from being achieved. Signature gene lists predicted from these tools lack consistency [18]. Many tools are not optimally balanced between precision and sensitivity [23]. The apparently significant mutation genes tend to be highly enriched for genes encoding extremely large proteins because of their prominence in mutation burden caused by the sequence length. Most importantly, the sequence information of insertions and deletions has never been considered in any of these tools. Moreover, recent studies showed that cancer is presumed to have a symptomatology strongly dependent on modules of functionally-related genes rather than on a unique important gene [24], [25].

Therefore, we proposed a new mutational sequence feature extracting method, named TTZ-feature, to extract features from not only mutation frequencies but also from sequences of insertions and deletions to identify TEP mutational signature genes for LUAD. Afterwards, networks at genetic and proteinic levels were inferred to uncover the modules of functionally related signature genes. Then, pathway analysis was explored to verify them according to their functions.

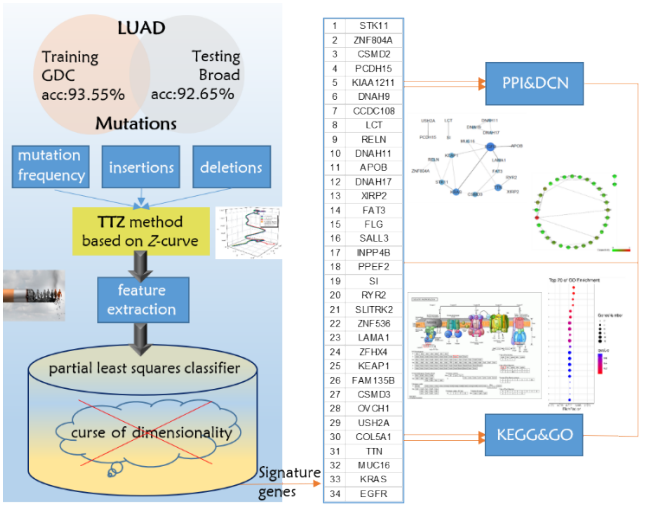

FIGURE 1. The study outline. We proposed a new mutational sequence feature extracting method, named TTZ-feature, to identify tobacco exposure mutation signature genes using PLS (Partial Least Squares) algorithm. Afterwards, subnetworks at different molecular levels were inferred to analyze their relationships. Then, KEGG and GO enrichment analysis were conducted to analyze their molecular functions and associated pathways.

TABLE 1. Summary of the sample datasets.

Number of samples	Cancer types	
	LUAD (TCGA)	LUAD (Broad)
Total number	564	159
Heavy smokers	154	45
Light smokers	195	45
Nonsmokers	63	23
Smoking status not available	152	46

Correspondingly, the aim of this study is two-fold: 1) to identify mutation signature genes highly related with TEP from hundreds of thousands of genome-wide genes; 2) To uncover the important relationships among identified genes from different molecular levels, i.e., gene expression level and protein level. Fig. 1 shows the study outline.

II. MATERIALS AND METHODS

A. LUAD DATASETS

Two independent datasets of LUAD were downloaded for training and testing the TEP classification model for signature gene identification. The somatic variants of the whole exome sequencing (WXS) of TCGA (Legacy Genomic Data Commons, https://portal.gdc.cancer.gov/projects) data were measured with MuTect Variant Calling Pipeline. Mutation profiles of total 22549 genes of 564 samples were available for analyzing.

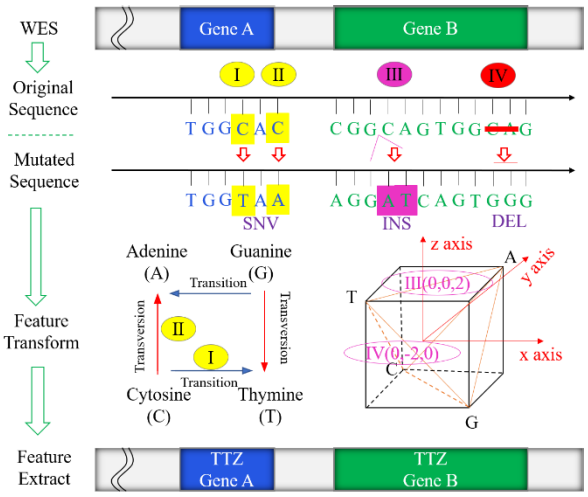

FIGURE 2. The flowchart of Feature extraction. The first row represents a WES sequence fragment on which are Gene A and Gene B. The second row shows a hypothetical sequence for Gene A and Gene B. There are two SNVs(I, II) in Gene A and INS together with DEL in Gene B. The left part of 'Feature Transforms' is the SNV patterns considered in the TTR feature. The right part is DEL and INS that can be mapped to Cartesian coordinate system according to the Z-curve theory. The coordinates of the sequence variation then can be transformed to ZC feature by the formula. The sum of all above features together with TNA is the TTZ feature calculated from the mutated sequence of a gene.

Another LUAD dataset consisting of somatic variants and clinical information named Broad dataset was downloaded from a published paper [8]. The mutation variations of its 183 LUAD samples were examined with a combination of WES or whole genome sequencing (WGS): 159 WES, 23 WES and WGS, and only 1 WGS. In this study, we focused only on the WES mutations. Consequently, mutation profiles of total 14809 genes were used. The information of these two datasets was summarized in Table 1.

B. THE PROPOSED TTZ FEATURE EXTRACTING METHOD BASED ON THE Z-CURVE ALGORITHM

To extract useful information as comprehensive as possible, features should contain information not only from mutation frequencies but also from the mutated sequences from insertions and deletions (indels). Therefore, the TTZ feature consists of three parts: ZC feature extracted from indels, TTR (transversion/transition ratio for a gene) and TNA (total number of alterations for a gene). The flow of feature extraction is shown in Fig. 2.

1) ZC FEATURE

Z-curve has been a geometrical approach for genome sequence analysis since proposed in 1990s. It's a three-dimensional curve or a point which represents a given DNA sequence and necessarily contains all the information that the corresponding DNA sequence carries. In our case, we creatively extract DEL and INS sequence information of each gene by applying the Z-curve theory. The resulting curve has a zigzag shape, hence the name Z-curve. The 3D curve or point of a given DNA sequence is calculated from the frequencies of the four bases occurring in it [26].

In the original Z-curve algorithm, the frequencies of nucleotides A, C, G and T occurring in a DNA fragment are denoted by a, c, g and t, respectively. Based on the Z-curve method, a, c, g and t are mapped onto a point in a 3D space, which are denoted by x, y, z [27].

$$\begin{cases} x = (a+g) - (c+t) \\ y = (a+c) - (g+t) \\ z = (a+t) - (c+g) \end{cases} \tag{1}$$

Consequently, compared with the wild sequence of a gene, we can get the Δx, Δy and Δz for the mutated sequence:

$$\begin{cases} \Delta x = (\Delta a + \Delta g) - (\Delta c + \Delta t) \\ \Delta y = (\Delta a + \Delta c) - (\Delta g + \Delta t) \\ \Delta z = (\Delta a + \Delta t) - (\Delta c + \Delta g) \end{cases} \tag{2}$$

where Δa, Δg, Δc, and Δt are the differences in the frequencies of bases a, g, c and t in the mutated and the corresponding wild sequence of this gene.

According to the quadratic form of x, y and z in [26].

$$x^2 + y^2 + z^2 = 4S - 1 \tag{3}$$

where S, the "genome order index" [28], is defined as

$$S = a^2 + c^2 + g^2 + t^2 \tag{4}$$

Thus, the relationship among x, y, z and a, t, c, g is:

$$x^2 + y^2 + z^2 + 1 = 4\left(a^2 + c^2 + g^2 + t^2\right) \tag{5}$$

To avoid the Dimensional curse problem (the number of variables is dozens of times of the number of samples), Δx, Δy and Δz are better to be combined into one variable.

Therefore, we defined the ZC-feature as (6):

$$ZC = \frac{\Delta x^2 + \Delta y^2 + \Delta z^2 + 1}{4} \tag{6}$$

Consequently, for deletion mutations (DEL) of a mutated gene:

$$\begin{cases} \Delta x_{ij} = -x_{iDELj} = \left(c_{iDELj} + t_{iDELj}\right) - \left(a_{iDELj} + g_{iDELj}\right) \\ \Delta y_{ij} = -y_{iDELj} = \left(g_{iDELj} + t_{iDELj}\right) - \left(a_{iDELj} + c_{iDELj}\right) \\ \Delta z_{ij} = -z_{iDELj} = \left(c_{iDELj} + g_{iDELj}\right) - \left(a_{iDELj} + t_{iDELj}\right) \end{cases} \tag{7}$$

where a_{iDELj}, g_{iDELj}, c_{iDELj} and t_{iDELj} are the frequencies of bases A, G, C and T in the jth deleted segment of the ith gene, respectively; x_{iDELj}, y_{iDELj} and z_{iDELj} are their Z-curve parameters.

Combining (6) and (7), the ZC-feature of the deletion can be written as:

$$ZC_{iINSj} = \frac{\Delta x_{iINSj}{}^2 + \Delta y_{iINSj}{}^2 + \Delta z_{iINSj}{}^2 + 1}{4} \tag{8}$$

correspondingly, for the inserted fragment (INS) of a mutated gene:

$$\begin{cases} \Delta x_{ij} = x_{iINSj} = \left(a_{iINSj} + g_{iINSj}\right) - \left(c_{iINSj} + t_{iINSj}\right) \\ \Delta y_{ij} = y_{iINSj} = \left(a_{iINSj} + c_{iINSj}\right) - \left(g_{iINSj} + t_{iINSj}\right) \\ \Delta z_{ij} = z_{iINSj} = \left(a_{iINSj} + t_{iINSj}\right) - \left(c_{iINSj} + g_{iINSj}\right) \end{cases} \quad (9)$$

then

$$ZC_{iINSj} = \frac{\Delta x_{iINSj}{}^2 + \Delta y_{iINSj}{}^2 + \Delta z_{iINSj}{}^2 + 1}{4} \quad (10)$$

where a_{iINSj}, g_{iINSj}, c_{iINSj} and t_{iINSj} are the frequencies of bases A, G, C and T in an inserted segment of the ith mutated gene, respectively; x_{iINSj}, y_{iINSj} and z_{iINSj} are their Z-curve parameters and ZC_{iINSj} is the corresponding ZC-feature.

Hence for the ith mutated gene with k deletions and l insertions, the corresponding ZC-feature should be:

$$ZC_i = \sum_{j=1}^{k} ZC_{iDELj} + \sum_{j=1}^{l} ZC_{iINSj} \quad (11)$$

Through this method, all kinds/lengths of mutation deletions and insertions can be transformed into a score which varies only with the mutated sequence.

2) TNA FEATURE

To quantify how many alterations happened to a mutated gene, TNA (total number of alterations) feature is defined as the total number of alterations of a mutated gene in a certain sample compared with its corresponding wild type. *Note: deletion, insertion or any other non-SNV alterations are all considered.*

3) TTR FEATURE

Transversion/transition ratio (TTR) feature is the ratio between single nucleotide variation types of transversions and transitions of a mutated gene. In total, there are 8 types tranversions and 4 types transitions. According to our previous study, the ratio of C > A/G > T (transversions) vs. C > T/G > A (transitions) is highly related with TEP. Therefore, TTR was particularly defined as the ratio of C > A/G > T (transversions) vs. C > T/G > A (transitions) happened in a mutated gene.

4) TTZ FEATURE

Taken together, for the ith mutated gene in the jth sample, TTZ-feature can be calculated as:

$$TTZ_{ij} = ZC_{ij} + TNA_{ij} + TTR_{ij} \quad (12)$$

where TTZ_{ij} is the TTZ-feature; TNA_{ij} is the total number of alterations; TTR_{ij} is the ratio between C > A/G > T (transversions) and C > T/G > A (transitions); ZC_{ij} is the feature extracted based on the Z-curve method.

From the definition of TTZ-feature, we can see that TNA and TTR consider both the SNV (single nucleotide variation) and non-SNV alteration frequencies of mutated genes while ZC considers the sequence information of non-SNV alterations. Therefore, TTZ-feature takes both frequency and sequence information of all types of mutations into consideration. Additionally, it doesn't need any information beyond DNA sequence information which makes it very practical for further applications.

5) PARTIAL LEAST SQUARES (PLS)

PLS is a widely used algorithm for modeling relationship between sets of observed variables by means of latent variables. It comprises regression and classification tasks as well as dimension reducing and modeling [29]. Instead of finding hyperplanes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables (i.e., classification labels) and the observed variables (TTZ features in our case) to a new lower space. Therefore, it performs very well for the analysis of high-dimension-small-sample data in bioinformatics. Additionally, the linearity characteristic of it makes it possible to identify important features according to their contributions to the classification model, which is the main aim of this study. Please see the Supplementary document for more details.

C. THE IDENTIFICATION OF MUTATIONAL SIGNATURE GENES USING THE DEEP SELECTING METHOD

Besides quantifying mutation information, another big challenge in identifying the mutational signature genes or the mutational biomarkers from thousands of genome-wide genes is the 'Curse of Dimensionality'. It means the number of variables is much bigger than the number of available samples (e.g. 13363 TTZ-features *vs*. 564 samples in TCGA LUAD dataset). Therefore, inspired by the concept of ''Deep Learning'' methods for extracting features step by step, the deep selecting method based on PLS algorithm was proposed to identify TEP signature genes. It consists of the following steps: 1) initiating a TEP classification model with TTZ features of whole exome genes as input variables; 2) sorting genes according to their contributions to the classification model; 3) removing certain number of the least important genes; 4) remodeling the classification model; 5) repeating steps 2-5 iteratively until the classification accuracy couldn't be improved any more. Then the remaining genes are considered as the signatures since using only TTZ features extracted from their mutated sequences can accurately predict the TEP of patients.

For the TEP classification model, the TCGA LUAD samples were used as the training samples and the Broad LUAD samples were used as the independent validation samples. The heavy smokers were taken as positive samples and nonsmokers were taken as negative samples. 5-fold cross-validation were performed to train the classification model. The details and corresponding pipeline are available in Supplementary material and Fig. S1.

D. CONSTRUCTION OF THE DIFFERENTIAL COEXPRESSION GENETIC NETWORK

Differential coexpression analysis is emerging as a complement to conventional gene coexpression analysis in response to environmental stresses or genetic changes. It's an efficient way not only to uncover the unique structural characteristics of a gene interaction network but also to provide new insights into the biological significance and the underlying gene correlation dynamics. The differential coexpression network of signature genes was constructed using their gene expression data as follows:

1) Calculate Pearson correlation coefficient of each pair of the identified signature genes with their expression values in heavy smokers and nonsmokers, respectively. Gene pairs whose p value $>= 0.05$ were removed.
2) The difference of coexpression of signature gene i and gene j under two different conditions h (heavy smokers) and n (nonsmokers) was measured as:

$$de = \left|P_h^{i,j} - P_n^{i,j}\right| * \max(\left|P_h^{i,j}\right|, \left|P_n^{i,j}\right|) \qquad (13)$$

where $P^{i,j}$ (differential edge) was the correlation coefficient between gene i and gene j. Here, we modified the method proposed by Hsu *et al.* [30] using the absolute value of $\left|P_h^{i,j} - P_n^{i,j}\right|$ to identify differentially coexpressed gene pairs. The de value of each gene pair was considered as ''weight''. The differential coexpression network was visualized using Cytoscape software [31].

E. INFERENCE OF PPI NETWORK OF SIGNATURE GENES

A weighted PPI network among signature genes was obtained from the STRING (Search Tool for the Retrieval of Interacting Genes) database (version 11.0) to search the known and predicted interactions between related proteins (https://string-db.org/) [32]. Using it, the interacted proteins or genes can be mapped to a weighted network where proteins or genes are denoted as nodes and the interactions are denoted as edges marked with a confidence score (cutoff, 0.4). The visualization of the network was accomplished by Cytoscape software.

F. GENE ONTOLOGY AND KYOTO ENCYCLOPEDIA OF GENES AND GENOMES PATHWAY ANALYSIS

The Gene Ontology (GO) knowledgebase is the worldwide largest source of information about the functions of genes. It is to develop a comprehensive, computational model of biological systems, ranging from the molecular to the organic level, across the multiplicity of species in the tree of life [33].

KEGG (Kyoto encyclopedia of genes and genomes) is a knowledgebase for systematic analysis of gene functions at the molecular-level in biological systems, from cells to organisms and ecosystems. It has been generated by genome sequencing and other high-throughput experimental technologies [34]. Both GO and KEGG pathway enrichment analysis for all signature genes were performed using the OmicShare tools, a free online platform for data analysis (www.omicshare.com/tools).

Except for KEGG, GO and PPI network, all other analyses were performed using MATLAB codes. Please refer to supplementary materials for more details.

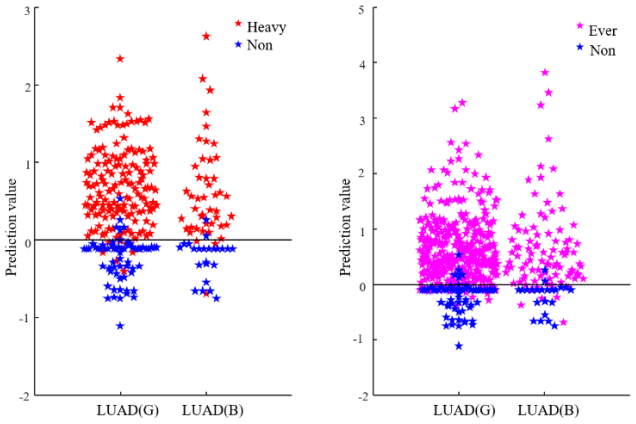

FIGURE 3. Classification plot of both datasets by the 34 significant tobacco-related gene classifier. Each dot on the plot represents a sample. The left figure shows the prediction of heavy (in red) and nonsmoker (in blue) tumors of TCGA (G) and Broad (B) datasets while the right one is the prediction result for ever and nonsmoker samples. The horizontal line in both subplots represents the classification boundary of the class categories (above it is Heavy/Ever group and below it is nonsmoker group).

III. RESULTS

A. THE IDENTIFICATION OF MUTATIONAL SIGNATURE GENES

According to the classification results obtained by the TTZ-features extracted from different number of genes, the best classification performance (shown in Fig. 3 and Table 2) was obtained by a set of 34 genes. The highest classification performance normally indicates their closest relationship to predict tumor's TEP. Therefore, these 34 genes were consequently considered as the potential tobacco-related mutational signatures. Their gene symbols, TTZ-features and molecular variations are listed in Table S1.

From Table 2, we could see that: for TCGA LUAD training dataset, sensitivity (SN), specificity (SP) and accuracy (ACC) are 94.16%, 92.06% and 93.55%, respectively; for Broad LUAD validation dataset, they are 93.33%, 91.30% and 92.65%, respectively. All these measurements are higher than 90%. More importantly, the differences between SNs and SPs for these two datasets are both small enough, only 2.10% and 2.03%, which means the false classification for both heavy smokers and nonsmoker samples are better than good enough.

To further test its performance for new samples, we extended the samples to all available ever (including current smokers) and nonsmokers. The corresponding classification results are also shown in Fig. 3 and Table 2. From Table 2, we could see that all performances are a bit worse than that of heavy/nonsmoker classification. But they are

TABLE 2. The performance of the classification model for the training and validation datasets using TTZ-features only of the final 34 mutational signatures.

Categories	Data set	SN	SP	ACC
Training (154 Heavy + 63 Non)	TCGA	94.16%	92.06%	93.55%
Validation I (45 Heavy + 23 Non)	Broad	93.33%	91.30%	92.65%
Validation II (349 Ever + 63 Non)	TCGA	91.98%	92.06%	91.99%
Validation III (90 Ever + 23 Non)	Broad	90.00%	91.30%	90.27%

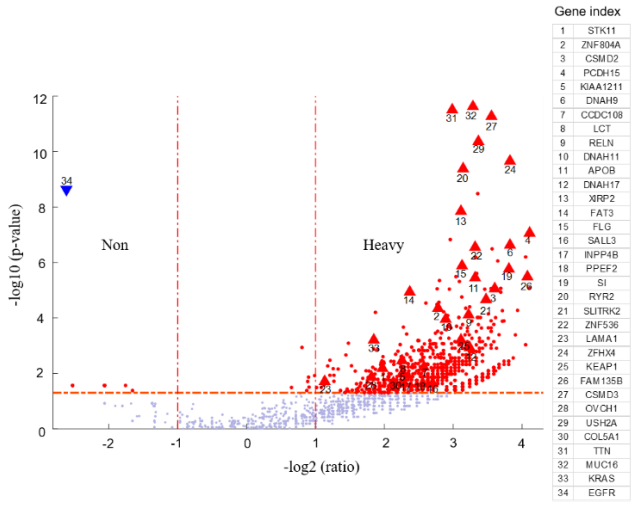

FIGURE 4. Volcano plot of TTZ-features of the 34 significant tobacco-related genes. Each dot on the plot is a single gene feature. Horizontal axis: fold change (in log2 scale); vertical axis: p-value (in log10 scale). The vertical lines highlight fold changes of {−1} and {+1}, while a horizontal line represents a *p*-value of 0.05. The red dot represents the genes with p-value < 0.05 and the triangle and round bigger markers represent the 34 signature genes.

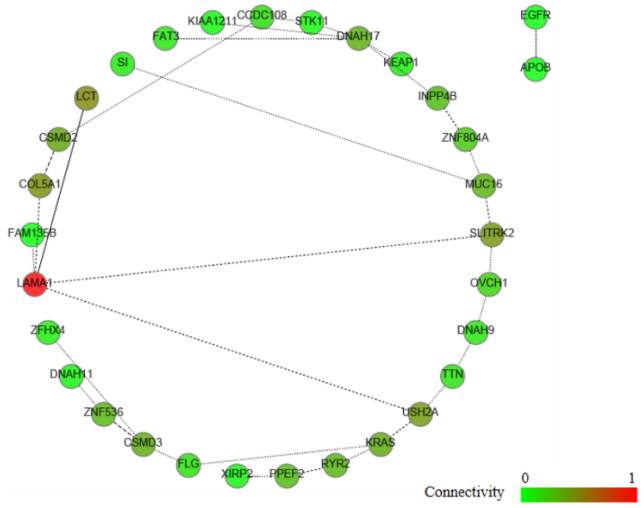

FIGURE 5. The differential coexpression genetic network of 31 genes (31 gene pairs). The soft threshold method was used to calculate the connectivity of the genes which was expressed using the color of nodes. The *de* value was considered as "weight" that was expressed using four line types, which are dotted line, short dash line, long dash line and solid line according to its value.

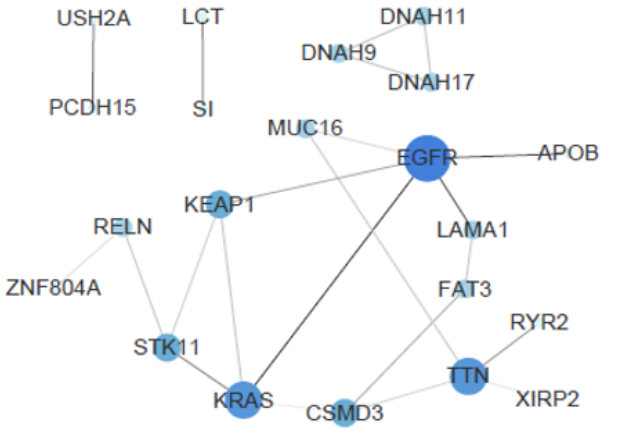

FIGURE 6. The PPI network of 21 signature genes. The network was drawn by the SRTING database and visualized by Cytoscape software. The shade of blue and size of nodes represent connectivity (degree) of genes, which represents the degree of relevance of gene pairs.

still greater than 90%, which indicates the generalization capability for new samples of this model is good enough too. Consequently, it proves the generalization capability for the identified TEP signature genes.

The volcano plot of the TTZ-features of the whole exome genes in TCGA LUAD dataset is shown in Fig. 4. The corresponding spots of these 34 genes are highlighted with bigger markers. It is very obvious that TTZ features of signature genes are all significantly different between heavy/nonsmokers. Of note, among them, only mean TTZ-feature in *EGFR* gene in nonsmokers is more than two-fold higher than that in heavy smokers. This result is completely consistent with the well-known knowledge that *EGFR* is a mutation signature for nonsmokers [35]. Additionally, *MUC16*, *TTN*, *CSMD3* are the three genes whose mutation patterns are most significantly different in heavy smokers *vs*. nonsmokers.

B. GENETIC NETWORK AND PPI NETWORK AMONG SIGNATURE GENES

Genetic network is shown in Fig. 5. There are 31 links (gene pairs) consisting of 31 genes in two coexpression networks that correspond with condition *h* (heavy smoker sample) and condition *n* (nonsmoker sample), respectively. Three signature genes (*PCDH15*, *RELN*, and *SALL3*) couldn't be involved in this network.

Figure 6 shows the PPI network among these 34 genes. It contains 21 nodes and 21 edges. It shows that 21 out

of 34 genes have interactions at the protein level. In the biggest PPI network, *EGFR* is the major one who has the most interactions with other signature genes at the protein level. Then *KARS* and *TTN* are the second major ones. *DNAH9*, *DNAH11* and *DNAH17* form a small subnetwork.

To further study the related molecular mechanisms in which the identified signature genes are directly involved, functional enrichment and pathway analysis were performed. GO secondary classification map can be seen in Fig. S2. Signature genes are mostly enriched in anatomical structure development, single-organism developmental process, binding and ion binding, extracellular region and other molecular functions. In addition, 34 signature genes are enriched in 92 KEGG pathways in total, of which PI3K-Akt signaling pathway, dorso-ventral axis formation, galactose metabolism are highly significant.

IV. DISCUSSION

Due to the importance of somatic mutations in cancer initiation and progression, various studies have recently been conducted to incorporate massive data analysis into cancer-specific mutation signature identification [9], [15], [17]. However, most methods used only tumor mutation frequencies, tumor mutation burden [36], the frequency ratios between different categories of SNVs [9], [13], [15], a binary entity (the presence or absence of a mutation) and other simple features as variables. The priceless information of the sequence information of indels has been wasted due to the fact that these methods haven't make any use of it. Consequently, there are only a few numbers of reports on indels based mutation signature genes and most of them were achieved by wet experiments or clinical observations. For example: the deletion of *EGFR* [35] and the insertion of *HER2* [37]. It's in a great necessity to develop a method to quantify mutation patterns both in number and sequence to uncover mutational signatures.

To overcome the abovementioned disadvantages, MutSigCV was proposed by Lawrence *et al.* [38]. The significant feature of it is the correction for patient-specific and gene-specific mutational heterogeneities by incorporating DNA replication timing and transcriptional activity. But the compensation factors about DNA replication timing and transcriptional activity need to be induced from other genes with similar properties (e.g. replication time, expression level), which makes it less practical. Additionally, MutSigCV is limited to recognize drivers to distinguish tumor samples from non-malignant samples, which makes it unsuitable for gene identification for personalized cancer treatment.

Therefore, we proposed a new feature extraction method, named TTZ-feature, to fill in the gap for the identification of TEP signature genes. Our TTZ-feature can consider mutation frequencies by TNA, SNV mutation spectrums by TTR and non-SNV segment sequence information by ZC-feature. More importantly, it does not need any other accessory information beyond sequence information. Therefore, it is very easily extended to other cancer related or disease related mutation signature gene identification.

PLS based deep selection algorithm was used to train the TEP classification model and to identify mutational signature genes with TTZ-features as input variables. By removing the least important genes iteratively, genes with the best classification performance were uncovered as the final TEP mutational signatures. The classification accuracies of both LUAD training dataset (TCGA) and independent validation dataset (Broad) are higher than 90%. Additionally, the balance between sensitivity and specificity are good enough (only 2.10% and 2.03%) too. These exciting high enough classification accuracies strongly proved the excellent performance of the proposed TTZ-feature.

The Volcano plot of TTZ-features of the whole exome of heavy smokers *vs*. nonsmokers in TCGA LUAD dataset is shown in Fig. 4 with the 34 signature genes highlighted. From this figure, we could see that the P-values of the TTZ-features of identified signature genes are almost the lowest (in $-\log 10$ scale) ones. It means that their mutation patterns are most significantly different between heavy smokers and nonsmokers.

Among them, *EGFR* is the only one whose TTZ-feature is much higher in nonsmokers than in heavy smokers. It is well known from wet experiments that *EGFR* mutation, which mainly targets nonsmokers and reformed smokers >15 years. The deletion mutations of *EGFR* have been reported strongly related with lung cancer [39]. According to Table S1, we could see that 17 out of 63 (27%) nonsmokers have *EGFR* deletion mutations. The average deletion frequency is 1.06 per sample and the average sequence length is 13.59nt in nonsmoker LUAD patients with *EGFR* mutation. The average ZC_{Del}-feature $= 5.77$ in nonsmoker samples, which is much higher than that in heavy smokers (ZC_{Del}-feature $= 0.006$). These results strongly proved that the ZC-feature part in TTZ-feature can take the sequence information of indels into the consideration. Therefore, it can uncover signature genes whose indels play important roles in cancer.

TP53 has been reported as the most frequently mutated gene in lung cancer [40]. It's average TTZ-feature in heavy smoker and nonsmoker is 0.33 and 0.18, respectively. The corresponding P-value between these two groups is 0.0012 which means the mutation patterns in heavy/nonsmokers are significantly different. But compared with the 34 signature genes, the difference of it in TTZ features between heavy and nonsmokers isn't significant enough. As a result, *TP53* isn't selected as a tobacco-related mutational biomarker. On the contrary, even *CSMD3* is reported as the second most frequently mutated gene (next to *TP53*) in lung cancer, it is identified as the tobacco-associated signature [41]. The selection of *CSMD3* and the deselection of *TP53* both strongly proved that TTZ method isn't a frequency depending feature.

To prove the contribution of ZC feature in TTZ, we compared the results obtained by using only TNA and TTR with or without ZC. From the receiver operating characteristic

curves (ROC) shown in Fig. S3, we can see that the performance of TTZ-feature model is a better one. This proves the contribution of sequence information extracted by ZC feature.

The overwhelming predominance in number of SNV mutation over indels of all training samples (for example, the average value of SNV per sample is 378.31 while of indels is 16.42) makes ZC-feature is easily to be buried by TNA and TTR features. Correspondingly, the improvement of ROCs with ZC compared ROCs without ZC feature isn't very extraordinary. Without enough and clear background knowledge or big enough samples of indels, it's hard to optimize the weights of these three parts of TTZ-feature up to now. It needs more research efforts on the regarding issue.

It should be expected that modules composed of functionally-related genes, rather than one or a few key genes, play symptomatology roles in cancer initiation and progression [24], [25]. The enrichment of the interaction relationships among identified signature genes at genetic and protein-protein levels strongly proved they work as several modules of functionally-related genes for the tobacco related LUAD. From the differential coexpression genetic network, we can see that only three genes (*PCDH15*, *RELN*, and *SALL3*) did not appear in the differential coexpression genetic network.

It is worth noting that the correlation between *LAMA1* and *LCT* and the correlation between *ZNF536* and *CSMD3* are significantly different in heavy smokers and nonsmokers. In other words, the regulatory relationships among these genes may be very sensitive to tobacco exposure pattern. This may provide a clue for the difference in the carcinogenesis mechanisms between smokers and nonsmokers.

To further explore the reliability of selected signature genes, we constructed their PPI network using STRING database. According to the database description, our network has significantly more interactions than expected. This means that our proteins have more interactions among themselves than what would be expected for a random set of proteins of similar size. It is well known that genes with higher betweenness centrality and degree have more features associated with malignancies, which are usually called 'hub' genes. In our results, the top three genes with highest betweenness centrality have the highest degree. They are *EGFR*, *KRAS* and *TNN*. Among them, *EGFR* and *KRAS* are two famous proto-oncogenes and their mutations can activate tumor proliferation. Numerous studies have shown that mutations in these genes are closely related to the development of lung cancer [42], [43]. Our results are highly consistent with previous studies, which shows that our methods and results are reasonable. On the other hand, some genes such as *TNN* have rarely appeared in previous literature on lung cancer research, which means they need to be paid more attention.

From PPI network shown in Fig. 6, we also discovered that *DNAH9*, *DNAH11* and *DNAH17* formed a small subnetwork. Studies have shown that the interaction of *DNAH9* with environmental tobacco smoke exposure can cause disease associated with abnormalities of pulmonary function [44], [45]. But the roles play by the co-functions of these three genes remain unclear. It may worth further studying.

GO and KEGG pathway analysis were performed for 34 signature genes (Figure S4). For cellular component, six signature genes (*DNAH9*, *DNAH11*, *DNAH17*, *USH2A*, *PCDH15* and *PPEF2*) were significantly associated with cilium damage (Fig. S4c). It's well known that cilium damage is one of the root causes of lung cancer. The smoke produced by smoking permeates all layers of the trachea and bronchus, which affects the cleaning movement function of cilia and eventually leads to the development of lung cancer. Additionally, the signal transduction, focal adhesion, ErbB signaling pathway, Wnt signaling pathway and VEGF signaling pathway etc. also appear in our pathway enrichment results.

From the fact that the EGFR and other mutation percentages of lung adenocarcinoma are quite different between Asian and Caucasian LUAD patients, therefore, it is necessary to describe the distribution of Asian and Caucasian in the datasets. However, the proportions of Asian LUAD patients in both of these datasets are very low (only 0.9%). Thus, it's impossible to get any statistically significant conclusion. Consequently, the results and conclusion obtained here are more specific to Caucasian LUAD patients.

V. CONCLUSION

Thirty-four genes were identified as tobacco related mutational signature genes for LUAD patients. Genetic network and PPI network analysis proved these genes co-operate as modules at different molecular levels. KEGG and GO analysis were then performed to verify their involvement in pathways and molecular functions to LUAD initiation and progression.

Our work provided a new method to extract molecular variation features from mutated sequences for identifying mutational signature genes using advanced statistical analysis methods. The satisfactory classification performance strongly proved the effectiveness of TTZ-feature as variables for data mining in molecular variation scope. Most importantly, it opens a novel way for disease-independent mutational mechanism research to improve precision medicine and to identify new drug targets for the development of personalized treatment.

ACKNOWLEDGMENT

Qien He and Zhewei Qiu contributed equally to this work.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA, A Cancer J. Clinicians*, vol. 66, no. 1, pp. 7–30, Jan./Feb. 2016.

[2] D. C.-L. Lam, L. Girard, R. Ramirez, W.-S. Chau, W.-S. Suen, S. Sheridan, V. P. C. Tin, L.-P. Chung, M. P. Wong, J. W. Shay, A. F. Gazdar, W.-K. Lam, and J. D. Minna, "Expression of nicotinic acetylcholine receptor subunit genes in non-small-cell lung cancer reveals differences between smokers and nonsmokers," *Cancer Res.*, vol. 67, no. 10, pp. 4638–4647, May 2007.

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA, A Cancer J. Clinicians*, vol. 61, no. 2, pp. 69–90, Mar./Apr. 2011.

[4] Y. Liu, Q. Lan, J. M. Siegfried, J. D. Luketich, and P. Keohavong, "Aberrant promoter methylation of p16 and MGMT genes in lung tumors from smoking and never-smoking lung cancer patients," *Neoplasia*, vol. 8, no. 1, pp. 46–51, Jan. 2006.

[5] C.-M. Choi, H. C. Kim, C. Y. Jung, D. G. Cho, J. H. Jeon, J. E. Lee, J. S. Ahn, S. J. Kim, Y. Kim, Y.-D. Choi, Y.-G. Suh, J.-E. Kim, B. Lee, Y.-J. Won, and Y.-C. Kim, "Report of the Korean association of lung cancer registry (KALC-R), 2014," *Cancer Res. Treat.*, vol. 51, no. 4, pp. 1400–1410, Oct. 2019.

[6] J. Norum and C. Nieder, "Tobacco smoking and cessation and PD-L1 inhibitors in non-small cell lung cancer (NSCLC): A review of the literature," *ESMO Open*, vol. 3, no. 6, Oct. 2018, Art. no. e000406.

[7] R. Govindan *et al.*, "Genomic landscape of non-small cell lung cancer in smokers and never-smokers," *Cell*, vol. 150, no. 6, pp. 1121–1134, Sep. 2012.

[8] M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, and C. Sougnez, "Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing," *Cell*, vol. 150, no. 6, pp. 1107–1120, Sep. 2012.

[9] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Børresen-Dale, and S. Boyault, "Signatures of mutational processes in human cancer," *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013.

[10] G. J. Riely, M. G. Kris, D. Rosenbaum, J. Marks, A. Li, D. A. Chitale, K. Nafa, E. R. Riedel, M. Hsu, W. Pao, V. A. Miller, and M. Ladanyi, "Frequency and distinctive spectrum of KRAS mutations in never smokers with lung adenocarcinoma," *Clin. Cancer Res.*, vol. 14, no. 18, pp. 5731–5734, Sep. 2008.

[11] J.-H. Ren, W.-S. He, G.-L. Yan, M. Jin, K.-Y. Yang, and G. Wu, "EGFR mutations in non-small-cell lung cancer among smokers and non-smokers: A meta-analysis," *Environ. Mol. Mutagen.*, vol. 53, no. 1, pp. 78–82, Jan. 2012.

[12] L. B. Alexandrov, Y. S. Ju, K. Haase, P. Van Loo, I. Martincorena, S. Nik-Zainal, Y. Totoki, A. Fujimoto, H. Nakagawa, T. Shibata, P. J. Campbell, P. Vineis, D. H. Phillips, and M. R. Stratton, "Mutational signatures associated with tobacco smoking in human cancer," *Science*, vol. 354, no. 6312, pp. 618–622, Nov. 2016.

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

[14] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering signatures of mutational processes operative in human cancer," *Cell Rep.*, vol. 3, no. 1, pp. 246–259, Jan. 2013.

[15] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, and L. Fulton, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069–1075, Oct. 2008.

[16] E. Cerami, E. Demir, N. Schultz, B. S. Taylor, and C. Sander, "Automated network analysis identifies core pathways in glioblastoma," *PLoS ONE*, vol. 5, no. 2, Feb. 2010, Art. no. e8918.

[17] W.-F. Guo, S.-W. Zhang, L.-L. Liu, F. Liu, Q.-Q. Shi, L. Zhang, Y. Tang, T. Zeng, and L. Chen, "Discovering personalized driver mutation profiles of single samples in cancer by network control strategy," *Bioinformatics*, vol. 34, no. 11, pp. 1893–1903, Jun. 2018.

[18] Y. Han, J. Yang, X. Qian, W.-C. Cheng, S.-H. Liu, X. Hua, L. Zhou, Y. Yang, Q. Wu, P. Liu, and Y. Lu, "DriverML: A machine learning algorithm for identifying driver genes in cancer sequencing studies," *Nucleic Acids Res.*, vol. 47, no. 8, p. e45, May 2019.

[19] I. R. Watson, K. Takahashi, P. A. Futreal, and L. Chin, "Emerging patterns of somatic mutations in cancer," *Nature Rev. Genet.*, vol. 14, no. 10, pp. 703–718, Oct. 2013.

[20] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes," *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, Sep. 2013.

[21] H. Abdi and L. J. Williams, "Partial least squares methods: Partial least squares correlation and partial least square regression," *Methods Mol. Biol.*, vol. 930, pp. 549–579, Jan. 2013.

[22] K. Song, J.-H. Bi, Z.-W. Qiu, R. Felizardo, L. Girard, J. D. Minna, and A. F. Gazdar, "A quantitative method for assessing smoke associated molecular damage in lung cancers," *Transl. Lung Cancer Res.*, vol. 7, no. 4, pp. 439–449, Aug. 2018.

[23] J. P. Hou and J. Ma, "DawnRank: Discovering personalized driver genes in cancer," *Genome Med.*, vol. 6, no. 7, p. 56, Jul. 2014.

[24] D. P. Cahill, K. W. Kinzler, B. Vogelstein, and C. Lengauer, "Genetic instability and darwinian selection in tumours," *Trends Biochem. Sci.*, vol. 24, no. 12, pp. M57–M60, Dec. 1999.

[25] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, Jan. 2000.

[26] R. Zhang and C.-T. Zhang, "A Brief Review: The Z-curve theory and its application in genome analysis," *Current Genomics*, vol. 15, no. 2, pp. 78–94, Apr. 2014.

[27] C.-T. Zhang and R. Zhang, "Analysis of distribution of bases in the coding sequences by a digrammatic technique," *Nucleic Acids Res.*, vol. 19, no. 22, pp. 6313–6317, Nov. 1991.

[28] C.-T. Zhang and R. Zhang, "A nucleotide composition constraint of genome sequences," *Comput. Biol. Chem.*, vol. 28, no. 2, pp. 149–153, Apr. 2004.

[29] Y. Tan, L. Shi, W. Tong, G. T. Gene Hwang, and C. Wang, "Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models," *Comput. Biol. Chem.*, vol. 28, no. 3, pp. 235–243, Jul. 2004.

[30] C.-L. Hsu, H.-F. Juan, and H.-C. Huang, "Functional analysis and characterization of differential coexpression networks," *Sci. Rep.*, vol. 5, no. 1, Oct. 2015, Art. no. 13295.

[31] P. Shannon, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[32] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering, "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.

[33] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: Tool for the unification of biology," *Nature Genet.*, vol. 25, pp. 25–29, May 2000.

[34] M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[35] Y. W. Choi, S. Y. Jeon, G. S. Jeong, H. W. Lee, S. H. Jeong, S. Y. Kang, J. S. Park, J.-H. Choi, Y. W. Koh, J. H. Han, and S. S. Sheen, "EGFR exon 19 deletion is associated with favorable overall survival after first-line gefitinib therapy in advanced non–small cell lung cancer patients," *Amer. J. Clin. Oncol.*, vol. 41, no. 4, pp. 385–390, Apr. 2018.

[36] S. Heeke and P. Hofman, "Tumor mutational burden assessment as a predictive biomarker for immunotherapy in lung cancer patients: Getting ready for prime-time or not?" *Transl. Lung Cancer Res*, vol. 7, no. 5, pp. 631–638, Dec. 2018.

[37] Y. Takase, K. Usui, K. Shimizu, Y. Kimura, T. Ichihara, T. Ohkawa, J. Atsumi, Y. Enokida, S. Nakazawa, K. Obayashi, Y. Ohtaki, T. Nagashima, Y. Mitani, and I. Takeyoshi, "Highly sensitive detection of a HER2 12-base pair duplicated insertion mutation in lung cancer using the Eprobe-PCR method," *PLoS ONE*, vol. 12, no. 2, Feb. 2017, Art. no. e0171225.

[38] M. S. Lawrence *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, Jul. 11 2013.

[39] S. Dogan, R. Shen, D. C. Ang, M. L. Johnson, S. P. D'Angelo, P. K. Paik, E. B. Brzostowski, G. J. Riely, M. G. Kris, M. F. Zakowski, and M. Ladanyi, "Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: Higher susceptibility of women to smoking-related KRAS-mutant cancers," *Clin. Cancer Res.*, vol. 18, no. 22, pp. 6169–6177, Nov. 2012.

[40] Z. Shajani-Yi, F. B. de Abreu, J. D. Peterson, and G. J. Tsongalis, "Frequency of somatic TP53 mutations in combination with known pathogenic mutations in colon adenocarcinoma, non–small cell lung carcinoma, and gliomas as identified by next-generation sequencing," *Neoplasia*, vol. 20, no. 3, pp. 256–262, Mar. 2018.

[41] P. Liu *et al.*, "Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing," *Carcinogenesis*, vol. 33, no. 7, pp. 1270–1276, Jul. 2012.

[42] O. Arrieta, A. F. Cardona, C. Martín, L. Más-López, L. Corrales-Rodríguez, G. Bramuglia, O. Castillo-Fernandez, M. Meyerson, E. Amieva-Rivera, A. D. Campos-Parra, H. Carranza, J. C. Gómez de la Torre, Y. Powazniak, F. Aldaco-Sarvide, C. Vargas, M. Trigo, M. Magallanes-Maciel, J. Otero, R. Sánchez-Reyes, and M. Cuello, "Updated frequency of EGFR and KRAS mutations in NonSmall-cell lung cancer in Latin America: The Latin-American consortium for the investigation of lung cancer (CLICaP)," *J. Thoracic Oncol.*, vol. 10, no. 5, pp. 838–843, May 2015.
[43] T. Kosaka, Y. Yatabe, R. Onozato, H. Kuwano, and T. Mitsudomi, "Prognostic implication of EGFR, KRAS, and TP53 gene mutations in a large cohort of Japanese patients with surgically treated lung adenocarcinoma," *J. Thoracic Oncol.*, vol. 4, no. 1, pp. 22–29, Jan. 2009.
[44] M.-H. Dizier, R. Nadif, P. Margaritte-Jeannin, S. J. Barton, C. Sarnowski, V. Gagné-Ouellet, M. Brossard, N. Lavielle, J. Just, M. Lathrop, J. W. Holloway, C. Laprise, E. Bouzigon, and F. Demenais, "Interaction between the DNAH9 gene and early smoke exposure in bronchial hyperresponsiveness," *Eur Respir J*, vol. 47, no. 4, pp. 1072–1081, Apr. 2016.
[45] M. C. Boelens, A. van den Berg, R. S. Fehrmann, M. Geerlings, W. K. de Jong, G. J. te Meerman, H. Sietsma, W. Timens, D. S. Postma, and H. J. Groen, "Current smoking-specific gene expression signature in normal bronchial epithelium is enhanced in squamous cell lung cancer," *J. Pathol.*, vol. 218, no. 2, pp. 182–191, Jun. 2009.

QIEN HE was born in Jilin City, Jilin, China, in 1996. He received the B.S. degree in process equipment and control engineering from Tianjin University, Tianjin, China, in 2019, where he is currently pursuing the M.S. degree in chemical process machinery. His research interests include computational cancer genomics, big data, and machine learning algorithms.

ZHEWEI QIU was born in Shijiazhuang, Hebei, China, in 1994. He received the B.S. and M.S. degrees in process equipment and control engineering from Northeast University and Tianjin University, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in biology with the Medical School, Tsinghua University. His research interests include cancer image recognition, computational biology, and machine learning algorithms.

YIFAN TONG was born in Tianjin, China, in 1997. He received the B.S. degree in process equipment and control engineering from Tianjin University, Tianjin, China, in 2019, where he is currently pursuing the M.Eng. degree in chemical process machinery. His research interests include big data, machine learning algorithms, and computer vision.

KAI SONG was born in Changchun, Jilin, China, 1975. She received the B.S. and Ph.D. degrees in control science and engineering from Zhejiang University, Zhejiang, in 1998 and 2005, respectively. From 2005 to 2007, she was an Assistant Professor with the Process Equipment and Control Engineering Department, School of Chemical Engineering and Technology, Tianjin University, China. Since 2007, she has been an Associate Professor with the Process Equipment and Control Engineering Department, School of Chemical Engineering and Technology, Tianjin University. She is the author of *Introduction of Synthetic Biology* (in Chinese, the first textbook about synthetic biology in China), and more than 80 articles. Her research interests include bioinformatics, synthetic biology, big data, and other applications of machine learning algorithms in biology, cancer, and process control. From February 2013 to October 2015, she was a Visiting Associate Professor with Dr. John Minna's lab at the Department of Clinical Science, UT Southwestern Medical Center, Dallas, Texas, USA. From then on, she started her research and published several articles about bioinformatics in cancer research cooperating with Dr. John Minna and Dr. Adi Gazdar. She's in charge of several projects supported by the National Natural Science Foundation of China and The National Key Research and Development Program of China.