Computational and Structural Biotechnology Journal 24 (2024) 404–411

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

A clinical-information-free method for early diagnosis of lung cancer from the patients with pulmonary nodules based on backpropagation neural network model

Xin Yang [a,1,2], Changchun Wu [a,2], Wenwen Liu [a], Kaiyu Fu [b], Yuke Tian [c], Xing Wei [d,3], Wei Zhang [c], Ping Sun [e,f], Huaichao Luo [g,*], Jian Huang [a,h,4,**]

[a] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Department of Obstetrics and Gynecology, West China Second University Hospital of Sichuan University, Chengdu 610041, China
[c] Department of medical oncology, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu 610041, China
[d] Department of Thoracic Surgery, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu 610041, China
[e] Department of Health Management Center & Institute of Health Management, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu 611731, China
[f] Chinese Academy of Sciences Sichuan Translational Medicine Research Hospital, Chengdu 610072, China
[g] Department of Clinical Laboratory, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu 610041, China
[h] School of Healthcare Technology, Chengdu Neusoft University, Chengdu, Sichuan 611844, China

ARTICLE INFO

Key words:
Lung cancer
Early diagnosis
Backpropagation neural network
TCRβ repertoire
Characteristic TCR clone

ABSTRACT

Lung cancer is the main cause of cancer-related deaths worldwide. Due to lack of obvious clinical symptoms in the early stage of the lung cancer, it is hard to distinguish between malignancy and pulmonary nodules. Understanding the immune responses in the early stage of malignant lung cancer patients may provide new insights for diagnosis. Here, using high-through-put sequencing, we obtained the TCRβ repertoires in the peripheral blood of 100 patients with Stage I lung cancer and 99 patients with benign pulmonary nodules. Our analysis revealed that the usage frequencies of TRBV, TRBJ genes, and V-J pairs and TCR diversities indicated by D50s, Shannon indexes, Simpson indexes, and the frequencies of the largest TCR clone in the malignant samples were significantly different from those in the benign samples. Furthermore, reduced TCR diversities were correlated with the size of pulmonary nodules. Moreover, we built a backpropagation neural network model with no clinical information to identify lung cancer cases from patients with pulmonary nodules using 15 characteristic TCR clones. Based on the model, we have created a web server named "Lung Cancer Prediction" (LCP), which can be accessed at http://i.uestc.edu.cn/LCP/index.html.

1. Introduction

Lung cancer is the most common type of cancer worldwide, with 5-year survival rate lower than 14% [1]. According to GLOBOCAN 2020, there were 19.3 million new lung cancer cases and 9.96 million deaths in 2020 [2]. Due to its high incidence and mortality, lung cancer is leading cause of cancer deaths globally [3,4]. Lack of early obvious clinical symptoms contribute greatly to the high mortality [5]. Thus, early diagnosis and treatment are crucial [6]. Pulmonary nodules are common manifestations of lung diseases. If a pulmonary nodule is

* Corresponding author.
** Corresponding author at: School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China.
E-mail addresses: luo1987cc@163.com (H. Luo), hj@uestc.edu.cn (J. Huang).
[1] https://orcid.org/0009–0003-3319–5094
[2] These authors contribute equally to this study.
[3] https://orcid.org/0000–0002-3035–5633
[4] https://orcid.org/0000–0003-3282–8892

https://doi.org/10.1016/j.csbj.2024.05.010
Received 9 November 2023; Received in revised form 15 April 2024; Accepted 7 May 2024
Available online 11 May 2024

greater than 6 mm, there is a high risk of lung cancer [7]. In recent years, CT has been widely used in high-risk populations for lung cancer diagnosis, reducing lung cancer mortality by up to 20% [8,9]. However, in the diagnostic results of CT, most positive results were false positives [10]. In addition, CT may increase the risk of radiation-induced cancer, accounting for approximately 0.5–5.5% of screened population [11–13]. Currently, there is no feasible approach for accurate and non-invasive lung cancer diagnosis at early stages. Therefore, an accurate and non-invasive diagnostic method is urgently needed for early detection of lung cancer.

In recent years, we have a deeper understanding of tumor immunology and the interactions between tumors and the immune system. The immune system is divided into two different subsystems: the innate immune system and the adaptive immune system. The adaptive immune system defends the body against different foreign antigens using two main types of antigen receptors: T-cell receptors (TCRs) and antibodies [14]. TCRs can recognize different epitopes presented by major histocompatibility (MHC) class I or II proteins on the cell surface [15]. In humans, TCRs consist of α and β chains. TCRβ chain is produced by the random recombination of variable (V), diversity (D) and joining (J) gene segments named 'VDJ recombination'. VDJ recombination generates the highly variable complementary determining region 3 (CDR3) which is the key to recognizing the antigenic peptide specifically [16]. This process can produce an incredible diversity of TCRs with a theoretical bound on the numbers of unique variants of $\sim 10^{13}$ for TCRβ chain [17–19]. CDR3 polymorphisms of TCRβ is the major component of TCR diversity, hence T cells could target any endogenous or exogenous antigen [20]. After recognizing disease-associated antigens, T cells could be specifically activated and expand, yielding a unique TCR repertoire. Lung cancer is an immune-related disease involving the whole immune system, especially adaptive immune system [21–23]. Thus, the analysis of TCR repertoire can be used for the study of immune response to lung cancers.

High-throughput sequencing has been increasingly developed and applied to disease and oncology research in recent years [24–26]. Recent studies showed that the adaptive immune response to tumor cells could serve as a postoperative prognostic marker and have been used for cancer diagnosis [27,28].

In this study, TCRβ CDR3 sequencing was performed on the peripheral blood of 99 patients with benign pulmonary nodules and 100 stage I lung cancer patients. Analyzing TCRβ CDR3 sequencing results revealed significant changes in the immune repertoire between lung cancers and benign pulmonary nodules. We further trained a backpropagation neural network model for the diagnosis of early-stage lung cancer in patients with pulmonary nodules.

2. Materials and methods

2.1. Sample collection

This study was approved by the medical ethical committee of Sichuan Cancer Hospital (SCCHEC-02–2021-037). Patients with a suspected lung cancer nodule diagnosed in the Thoracic Surgery department at Sichuan Cancer Hospital were enrolled. All patients met the following criteria: [1] pathologically and immunohistochemically confirmed diagnosis by expert pathologists; [2] a lack of acute infection or chronic active inflammatory disease. Clinical information (age, gender), CT scan information (nodule size), and pathology information (pathological stage) recorded. After obtaining pathological results, the subjects with benign lung nodule and malignant lung cancers were enrolled for TCR sequencing. Peripheral blood was collected before surgery and stored at − 80 °C until use.

In Sichuan Cancer Hospital, the pulmonary nodules were detected by SIEMENS Definition Flash CT (SIEMENS Healthineers Co., Ltd, Erlangen, GER) and PHILIPS Brilliance iCT (Koninklijke Philips N.V., Amsterdam, NED). Clinical information was collected from the hospital management system.

2.2. TCR sequencing

High-throughput sequencing of TCRβ genes was performed using previously described methods [29]. In brief, we extracted total genome DNA from peripheral blood. The TCRβ DNA was amplified using multiplex PCR. Sequencing libraries were loaded onto the Illumina NavaSeq6000 System.

2.3. TCR sequence analyses

Sequence quality was monitored by cross analyses of potential contamination. Potential TCR Vβ, Dβ, and Jβ germline gene assignment was conducted using a locally operating IgBLAST program from NCBI. TCRβ CDR3 amino acid sequences between 6 and 28 aa in length were retained while non-functional sequences were removed. For each sample, 30,000 random TCR gene sequences were selected for subsequent analyses.

2.4. TCR repertoire diversity analyses

We analyzed the diversity index D50, Shannon index, Simpson index, and the frequency of the largest clone in each sample to reflect TCR repertoire diversity. Better diversity corresponds to higher D50, Shannon and Simpson indices, and lower frequency of the largest clone. D50 is defined as the ratio of the numbers TCR clones that account of half of the total TCR sequences in one sample [29]. The formulas of Shannon index and Simpson index were as follows (i.e., eq. [1] and eq. [2]):

$$\text{Shannon index} = -\sum_{i=1}^{n} pi \times lnpi \tag{1}$$

$$\text{Simpson index} = 1 - \sum\nolimits_{i=1}^{n} pi\hat{}2 \tag{2}$$

2.5. Statistical analyses

Statistical analysis and graph drawing were performed using R programming language (R 4.0.3). Mann-Whitney U test was used to compare differences between two groups. A p-value < 0.05 was considered statistically significant. The receiver operating characteristic (ROC) curve was used to illustrate the diagnostic ability of a binary system. A higher area under the curve (AUC) indicates better diagnostic ability. PPV (Positive Predictive Value) represents the proportion of true positive results among all the positive results obtained from the test. A higher PPV value suggests a higher likelihood of a positive result being accurate. NPV (Negative Predictive Value), on the other hand, represents the proportion of true negative results among all the negative results obtained from the test. A higher NPV value suggests a higher likelihood of a negative result being accurate.

2.6. Backpropagation neural network

A backpropagation neural network was trained using the nnet R package. The initial random weights were set to 0.1, weight decay to 5e-4, the maximum iterations (i.e., epochs) to 100, the maximum allowable number of weights to 100,000, and 4 units were utilized in the single hidden layer. We used default values for all other parameters.

In the preliminary stage, we selected 99 samples (i.e., 50 MG patients, 49 BG patients) for neural network model construction. 70% of the 99 samples were randomly selected as the training set (i.e., 35 BG patients, 35 MG patients) and the rest (i.e.,14 BG patients, 15 MG patients) as the validation set. Subsequently, we collected an additional set of 100 samples as an independent validation set to evaluate the model's performance. Importantly, these 100 samples were not involved in the

model's construction and fine-tuning process.

2.7. Characteristic TCR clones

The 99 samples (i.e., 50 MG patients, 49 BG patients) were used for TCR feature identification, and the same set of samples was used for constructing the neural network model. First, we compared the expansion value of each TCR CDR3 sequence in both groups (i.e., 50 MG patients, 49 BG patients), retaining TCRs with significant expansion differences. We then calculated the sharing rate of each clone and retained clones that were expressed in more than 20% of the samples in both groups. Next, we calculated the average expansion value of these TCRs in both groups.

In the MG (i.e.,50 patients), if the average expansion value was more than 3 times higher than in the BG (i.e.,49 patients), the corresponding TCR CDR3 sequence was defined as a characteristic TCR clone of the MG. The same applies to the BG. The TCRMatch tool was employed for predicting antigen epitopes [30].

2.8. Implementation of the web server

LCP (Lung Cancer Prediction) is a web service that utilizes HTML on the frontend to create the structure and content of web pages, CSS to define the style and layout, and JavaScript for webpage interactivity and dynamic effects. On the backend, it employs the PHP language to handle data submitted through web forms and obtains results through backend programs written in R language. Ultimately, the website is hosted on Apache, enabling users to access the website over the internet.

3. Results

The different usages patterns of Vβ and Jβ genes in the benign and the malignant samples.

To investigate the T cell immune response in patients with benign pulmonary nodule or stage I lung cancer, we used high-throughput sequencing approach to characterize TCRβ repertoires in the peripheral blood of benign and malignant samples (Fig. S1). In total, we obtained almost 1.4×10^8 TCRβ sequences from 49 benign pulmonary nodule patients and 50 stage I lung cancer patients (Table S1).

In our study, among the 47 functional human TRBV and the 13 functional human TRBJ genes, 6 distinct Vβ and 2 distinct Jβ genes were identified (Table S2-S3). We found that the usage frequency of 1 distinct Vβ gene (TRBV20–1) was significantly elevated in the malignant samples compared with that in the benign samples; In contrast, the usage frequencies of 5 distinct Vβ genes (TRBV7–7, TRBV10–2, TRBV11–1, TRBV11–2, TRBV24–1) were significantly reduced in the malignant samples compared to those in the benign samples (Fig. 1a, $p < 0.05$). In addition, the usage frequencies of 2 distinct Jβ genes (TRBJ1–6, TRBJ2–2) were significantly reduced in the malignant samples compared with those in the benign samples (Fig. 1b, $p < 0.05$). Furthermore, nearly half of the MG were clustered together in the expansion of 6 distinct Vβ genes (Fig. 1**c**).

Similarly, we compared the usage frequencies of 611 V-J pairs in the benign and the malignant samples, and the result showed that there were 75 V-J pairs with significant differences, among which 13 pairs (TRBV3–1/TRBJ1–1, TRBV3–1/TRBJ1–5, TRBV5–8/TRBJ1–5, TRBV6–4/TRBJ2–5, TRBV6–8/TRBJ1–3, TRBV7–9/TRBJ1–5, TRBV9/TRBJ2–1, TRBV10–2/TRBJ1–6, TRBV11–2/TRBJ2–2, TRBV18/TRBJ1–4, TRBV20–1/TRBJ2–6, TRBV27/TRBJ2–6) had significantly increased frequency in the malignant samples compared with the benign ones; The frequencies of left 62 pairs were significantly decreased in the malignant samples (Fig. 1d; Table S4). In addition, most of the MG were clustered together in the expansion of 75 distinct V-J pairs (Fig. 1e).

Next, we compared the V-J combinations of the benign and the malignant samples using 3-D V-J plots. These V-J combination plots showed high single columns in both samples, indicating imbalanced TCR repertories. However, such status was more severe in the malignant samples (Fig. 1f-j) than that in benign ones (Fig. 1k-o).

The TCR diversities were significantly reduced in the malignant samples compared with those in the benign ones.

When lung cancer emerges, relevant antigens may stimulate proliferation of some T cells with unique TCRβ. In this study, we calculated D50, Shannon index, Simpson index, and the frequency of the largest TCR clone in each peripheral blood sample. Significant differences were found between the benign and the malignant samples.

The results showed that D50s, Shannon indexes, and Simpson indexes in the malignant samples (mean ± standard error of the mean [SEM]; D50s: 0.1125 ± 0.066, Shannon indexes: 7.8827 ± 0.8898, Simpson indexes: 0.9839 ± 0.0269) were significantly lower than those in the benign ones (D50s: 0.1466 ± 0.0562, $p = 0.0162$; Fig. 2a; Shannon indexes: 8.3405 ± 0.577, $p = 0.0156$; Fig. 2b; Simpson indexes: 0.9932 ± 0.0052, $p = 0.0051$; Fig. 2c). In the BG, 3 individuals (6%) had the largest clones comprising more than 10% of total CDR3 sequences. However, 14 individuals (28%) had the largest clones comprising more than 10% of total CDR3 sequences in the MG. TCRβ genes obtained from the malignant samples ($8.66\% \pm 6.62\%$) showed significantly higher frequencies of the largest TCR clone compared with the benign ones ($5.73\% \pm 2.45\%$, $p = 0.0164$; Fig. 2d). These results suggested that the T cell repertoire diversity had greatly reduced after lung cancer occurred.

3.1. Larger pulmonary nodules, lower TCR repertoire diversity

T cell repertoire diversity was lower in the MG versus the BG. In clinical practice, the size of pulmonary nodules is a more important and more frequently used factor in the diagnosis of lung cancer. We analyzed the correlation between TCR repertoire diversity and the size of pulmonary nodules.

Although no significant linear negative correlation was observed between nodule size and D50s ($r = -0.0637$, $p = 0.531$; Fig. 3a) or Shannon indexes ($r = -0.1724$, $p = 0.088$; Fig. 3b), we found that Simpson indexes negatively correlated with nodule size ($r = -0.3193$, $p = 0.0013$; Fig. 3c) as did the frequencies of the largest clone ($r = 0.2566$, $p = 0.0103$; Fig. 3d). These results revealed that larger nodule sizes correlated with lower TCR diversity.

15 characteristic TCR clones were identified and applied for the training of backpropagation neural network.

Here, we obtained 1160974 unique TCR CDR3 clones, including 625916 unique clones in the BG and 579489 unique clones in the MG. Of these, 44431 unique clones were shared between groups (Fig. S2). 5 and 10 TCR CDR3 clones were identified as the BG and MG characteristic sequences, respectively. All MG and BG were clustered together separately based on the expansion of these 15 characteristic TCR clones (Fig. 4a).

Neoantigens play a key role in the recognition of tumor cells by T cells. Using the TCRMatch tool, we identified the specific epitopes targeted by these 15 characteristic TCR clones. In 11 of 15 characteristic TCR clones, their corresponding binding epitopes belong to SARS CoV-2; In other 3 characteristic TCR clones, their corresponding binding epitopes belong to influenza A virus, human herpesvirus 5, and yellow fever virus 17D. The sequence (i.e., PHQDSSPAAPLHPGAAGGRSQP) was the only one without any annotation information (Table S5).

Above 15 TCR CDR3 clones were used to build a neural network model (Fig. 4b). The results showed that validation-AUCs (AUC: 0.7314 ± 0.0851) ranged from 0.35 to 0.98 in 1000 repetitions of building the backpropagation neural network model indicating fair model performance (Fig. 4c). The model with the mean value of validation-AUCs was chosen as the final model (training set: AUC = 0.99; Fig. 4d; validation set: AUC = 0.74; Fig. 4e). In the final model, the positive predictive value (PPV) was 80%, and the negative predictive value (NPV) was 71.4%. Notably, the independent validation set exhibits an AUC level similar to that of the validation set (independent validation set: AUC =

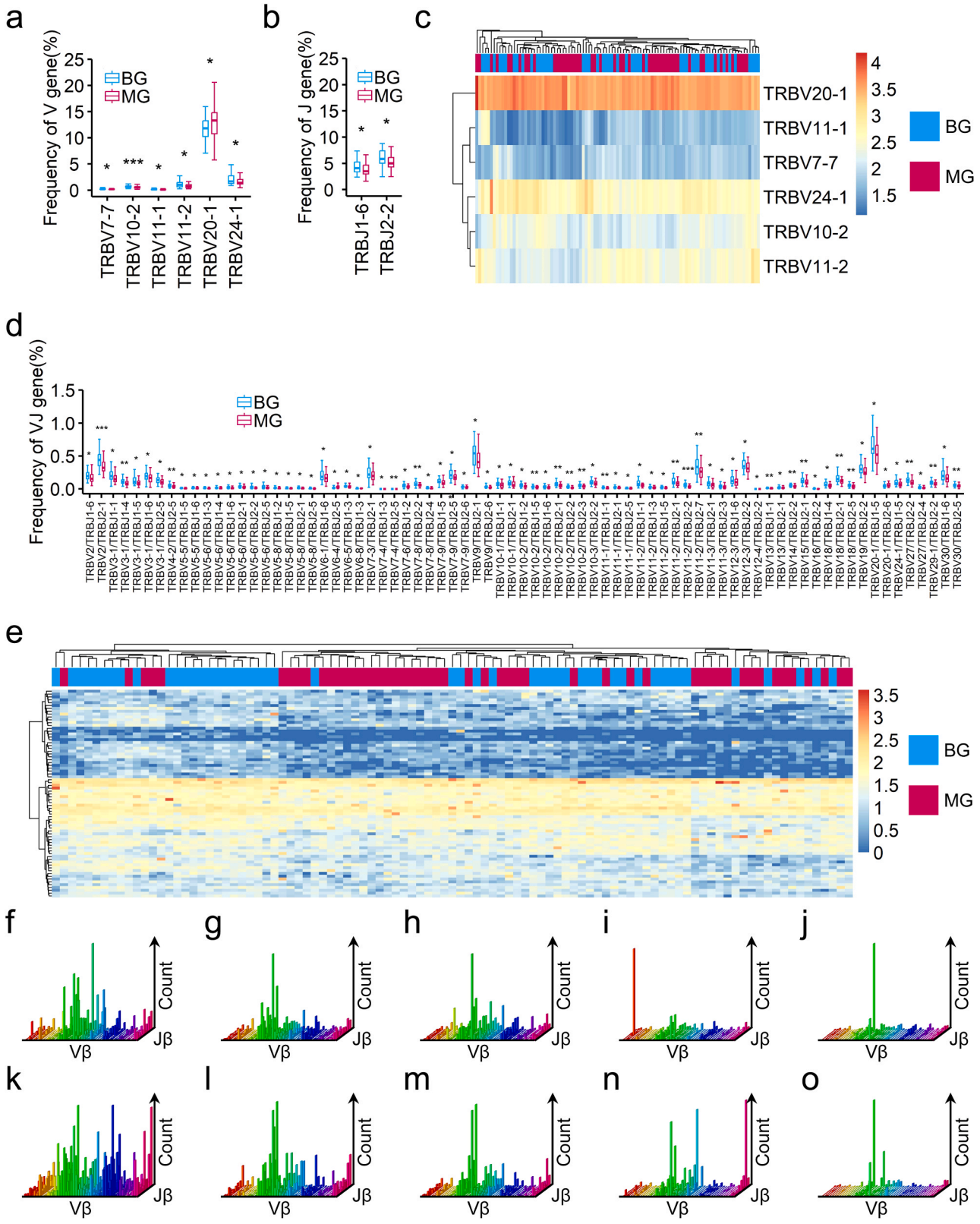

Fig. 1. The different usages patterns of Vβ and Jβ genes in the benign and the malignant samples The usage frequencies of (a) 6 significantly different TRBV and (b) 2 significantly different TRBJ genes in the benign (n = 49) and the malignant pulmonary nodule patients (n = 50). The error bars indicate the standard deviation. (c) The heatmap of expansion frequencies of the 6 significantly different TRBV genes in the benign and the malignant patients. (d) The usage frequencies of 75 significantly different V-J pairs in the benign and the malignant patients. (e) The heatmap of expansion frequencies of the 75 significantly different V-J pairs in the benign and the malignant patients. (f-o) The V-J combinations were analyzed using the 3D plots, f to j, representative examples of the benign patients, k to o, representative examples of the malignant patients. X and Y axes depict functional human TRBV and TRBJ alleles, respectively. Z-axis indicates the counts of sequence reads. * , $p < 0.05$, * *, $p < 0.01$, * ** , $p < 0.001$, Mann-Whitney U test.

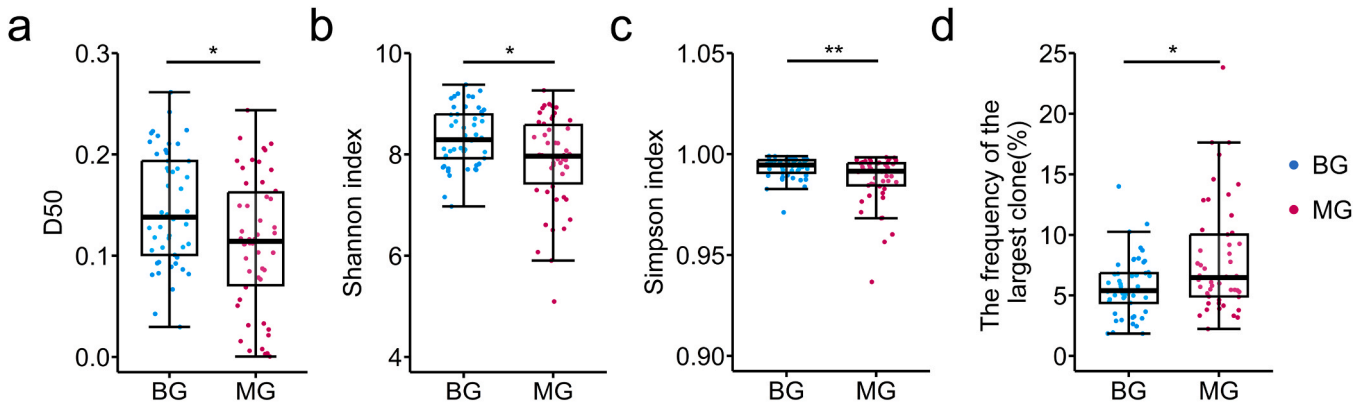

Fig. 2. The TCR diversities were significantly reduced in the malignant samples compared with those in the benign ones Comparison of (a) D50, (b) Shannon index, (c) Simpson index, and (d) the frequency of the largest clone in the malignant patients versus the benign patients. * , $p < 0.05$, * *, $p < 0.01$, * ** , $p < 0.001$, Mann-Whitney U test.

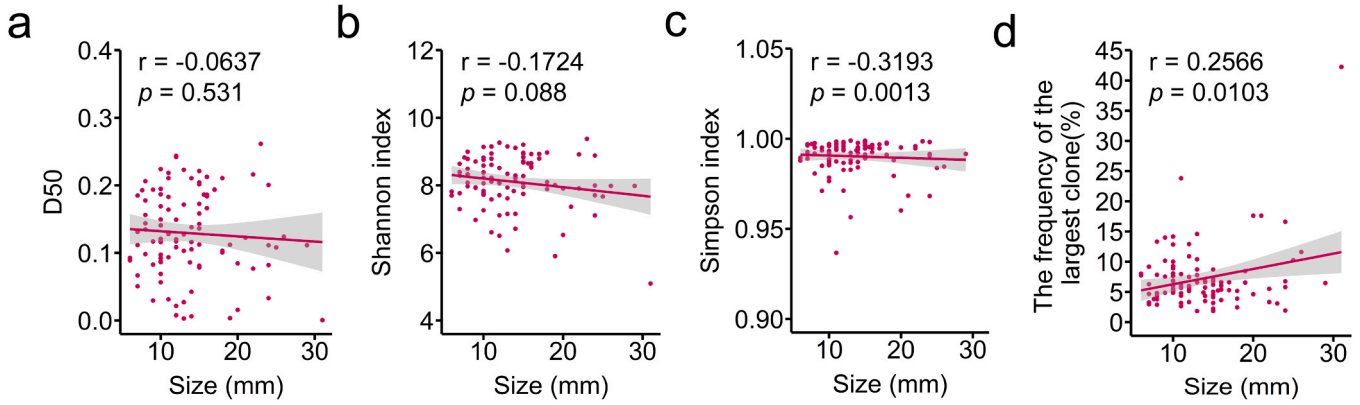

Fig. 3. Larger pulmonary nodules, lower TCR repertoire diversity The correlation between the size of pulmonary nodule and TCR diversity indexes (**a**: D50, **b**: Shannon index, **c**: Simpson index, **d**: the frequency of the largest clone), Pearson correlation test.

0.72; Fig. 4f).

Based on the model, a web server, called "Lung Cancer Prediction" (LCP), has been developed to provide public access to the back-propagation neural network model. You can access it at http://i.uestc.edu.cn/LCP/index.html.

4. Discussion

To explore the T cell immune responses in the transition from benign pulmonary nodules to stage I lung cancer, we analyzed the peripheral TCR repertoires of benign and malignant samples. Our results showed significantly changed usage frequencies of specific TRBV, TRBJ genes and V-J pairs, and significantly reduced TCR repertoire diversities in the MG compared with those in the BG. We built a backpropagation neural network model without any clinical information to identify potential lung cancer patients from patients with pulmonary nodules using 15 characteristic TCR clones. The model showed a fair performance.

Low-dose computed tomography (LDCT) is commonly regarded as the current standard screening method for lung cancer, proved to benefit the diagnosis of lung cancer through multiple large randomized clinical trials. However, a significant false positive rate is observed, leading to unnecessary invasive biopsies being performed [31]. Adaptive immunity, mainly relying on antigen-specific T/B cells, plays an important role in fighting against cancer. According to Burnet's clonal selection theory, when a disease occurred, the neoantigens could stimulate the corresponding unique T cell to proliferate [32]. Specific TCR repertoire is related to unique disease. Using high-throughput sequencing to characterize T cell repertoires may provide a suitable approach to analyze T cell responses to different diseases [33–35].

In our study, comparing TCR repertoires between the benign and the malignant groups showed significantly different usage frequencies of multiple TRBV, TRBJ genes and V-J pairs, indicating these genes may involve in anti-viral or anti-tumor immune responses in the malignant patients.

Typically, The V and J genes are the major constituent of TCR repertoire. Diversity of T-cell repertoire ensures that the level of cellular immunity in the body can adequately respond to a complex antigen environment. Recently, the Lung CT Screening Reporting and Data System (Lung-RADS) is widely used to evaluate and manage pulmonary nodules depending on nodule size [36,37], as the risk of lung cancer increases with the size of pulmonary nodule [38]. For D50s and Shannon indexes, no significant linear negative correlation with nodule size was observed, which may be due to their sensitivities to changes in evenness and richness. However, we found that the TCR repertoire diversities in pulmonary nodule patients declined with size of the pulmonary nodule, which indicated that there was production of large clones during the immune process of pulmonary nodule enlargement. Furthermore, our results showed that the TCR repertoire diversities in the malignant samples were significantly reduced compared with those in the benign samples, such as the diversity indexes including D50s, Shannon indexes, Simpson indexes, and the frequencies of the largest clone. Such finding revealed that there was a reduction of TCR variety and increase of large clone in malignant patients.

In clinical practice, the initial manifestations of lung cancer are complex, diverse and lack of specific symptoms, making diagnosis even more difficult. In recent research, it has been proved that cell-free DNA

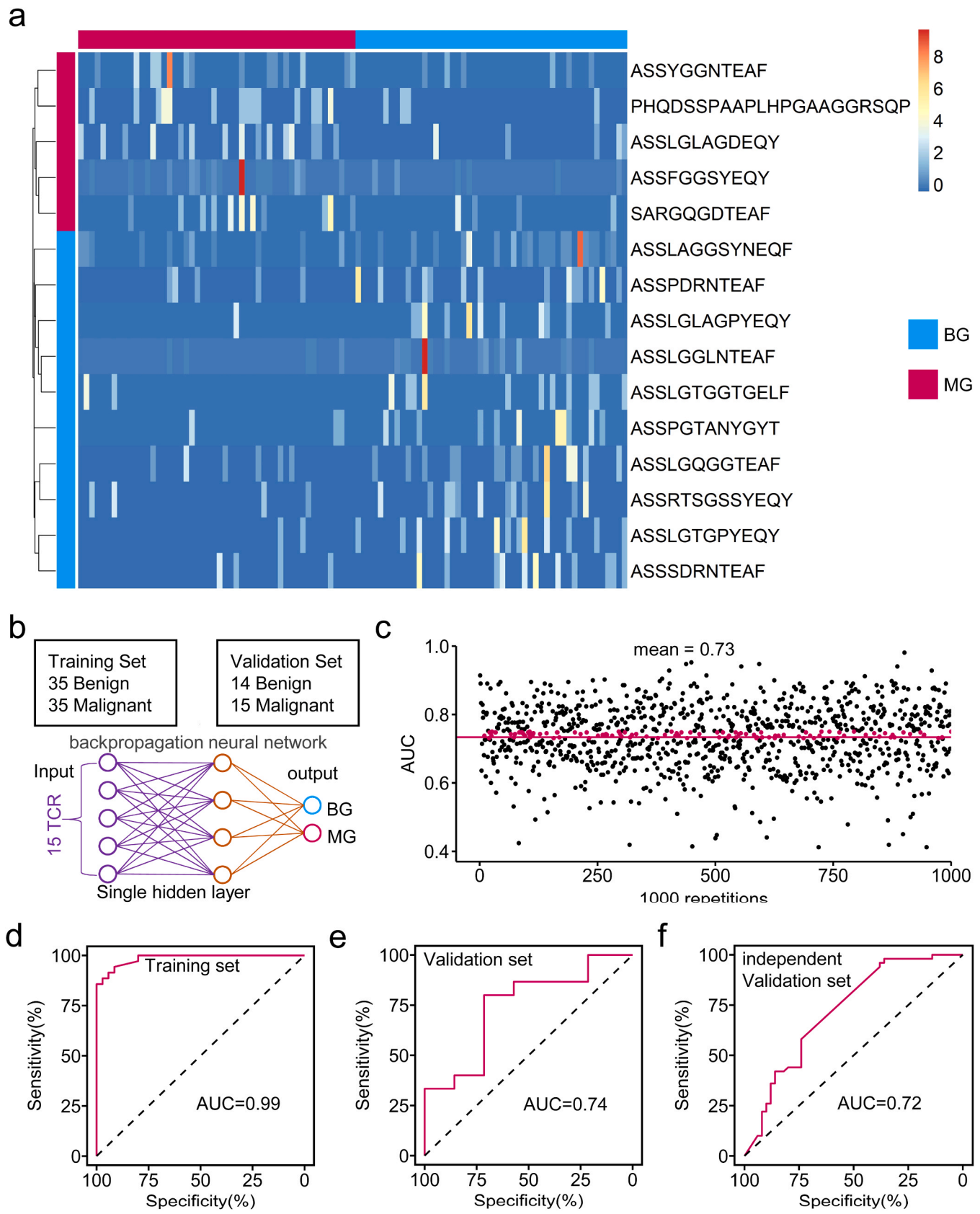

Fig. 4. 15 characteristic TCR clones were identified and applied for the training of backpropagation neural network (**a**) The heatmap of expansion frequencies of the 15 characteristic TCR clones in the benign and the malignant patients. (**b**) The experimental design of the backpropagation neural network. (**c**) The Scatter plot showed validation-AUCs in 1000 repetitions of building the neural network model and their mean value. The black points present the AUC values greater than 0.75 or less than 0.73, while the highlight points present the AUC values greater than 0.73 and less than 0.75. In the final backpropagation neural network model, the AUC of (**d**) the training set, (**e**) the validation set and (**f**) the independent validation set.

(i.e., cfDNA) methylation was the most promising genomic feature for cancer signal detection through comparing ten different classifiers. Unfortunately, analyzable tumor tissue comprised only 17.14% of the total analyzable stage I lung cancer cases in the research. Furthermore, using whole-genome methylation, the sensitivity of cancer signal detection classifier was bellowed 10% when evaluated at 98% specificity [39]. However, another study has demonstrated that the elevated levels of cfDNA detected in cancer patients did not originate from neoplastic cells or adjacent normal epithelial cells derived from the tumor's tissue of origin [40]. This may be the reason for the inaccuracy of cfDNA as a biomarker. In our previous research, using three vital features (Ground glass nodule, Shannon index, and evenness index), we developed a TCR-based model named TCRnodseek to distinguish malignant nodules from benign ones, which performed well in validation group (AUC = 0.80) [41]. However, clinical imaging features such as ground glass nodules require CT imaging. In recent research, a meta-analysis was conducted using data from 111.6 million adult participants across three continents (Asia, Europe, and America). The findings of the meta-analysis revealed a significant increase in cancer risks associated with CT scans in adults. Thus, there is a clinical demand for the development of equally effective or slightly less effective but non-CT-based approaches for the monitoring and management of pulmonary nodules.

In our study, by comparing the expansion of TCR clone in the benign and the malignant groups, 15 characteristic TCR clones (5 benign characteristic TCR clones and 10 malignant characteristic TCR clones) were identified, suggesting that there were skewed TCR repertoires in groups. Moreover, our investigation revealed that the majority of these 15 characteristic TCR clones possess binding epitopes that are associated with SARS-CoV-2, indicating many lung nodules were induced by SARS-CoV-2 infection in current years. More importantly, we created a clinical-information-free method for early diagnosis of lung cancer from the pulmonary nodule patients based on backpropagation neural network model using 15 characteristic TCR clones. Notably, our method involves the use of peripheral blood for detection, which falls under the category of non-invasive testing without the need for CT scans. In this model, the AUC of validation patients, the PPV, and the NPV were 0.74, 80%, and 71.4%, respectively. Additionally, the independent validation set exhibits an AUC level similar to that of the validation set, indicating a fair performance.

In conclusion, analyzing the TCR repertoire diversities in the peripheral blood of pulmonary nodule patients is beneficial to early diagnosis of lung cancer. The significantly changed usage frequencies of TRBV genes and reduced TCR diversities, as indicated by D50s, Shannon indexes, Simpson indexes, and the frequencies of the largest TCR clone reflected active T cell immune responses during the progression of viral infection leading to early stage lung cancer. The reduced TCR diversities were correlated with the size of pulmonary nodules. Importantly, we build a backpropagation neural network model without clinical information to identify the potential lung cancer from pulmonary nodules patients using only 15 characteristic TCR clones. Based on the model, we have created a web server named "Lung Cancer Prediction" (LCP), which can be accessed at http://i.uestc.edu.cn/LCP/index.html.

Ethics approval and consent to participate

The study was approved by the medical ethical committee of Sichuan Cancer Hospital (SCCHEC-02–2021-037).

Consent for publication

Not applicable.

Author contributions

XY, JH, HL designed and directed the study; XY, CW performed the sequence analyses; HL organized patient recruitment and sample collection; ALL the authors worked together to write the manuscript.

Funding

This study was supported by Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (ZYGX2022YGRH004), the National Natural Science Foundation of China (62071099; 62371112), the Sichuan Medical Association Research project (S20087), and Sichuan Cancer Hospital Outstanding Youth Science Fund (YB2021033).

CRediT authorship contribution statement

Wenwen Liu: Conceptualization, Validation. **Kaiyu Fu:** Conceptualization, Methodology, Validation. **Yuke Tian:** Conceptualization, Validation. **Xing Wei:** Conceptualization, Validation. **Wei Zhang:** Conceptualization, Validation. **Ping Sun:** Conceptualization, Validation. **Huaichao Luo:** Data curation, Methodology, Validation. **Jian Huang:** Data curation, Funding acquisition, Methodology, Writing – review & editing. **Xin Yang:** Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Changchun Wu:** Software.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data availability

Raw data of this project have been uploaded to https://bigd.big.ac.cn/gsa (HRA001754 and HRA002253).

Acknowledgments

We thank all the doctors, nurses, technicians, patients for their help in this study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.010.

References

[1] Jayia PK, Mishra PK, Shah RR, Panayiotou A, Yiu P, Luckraz H. Preoperative assessment of lung cancer patients: evaluating guideline compliance (re-audit). Asian Cardiovasc Thorac Ann 2015;23(3):299–301.
[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71(3):209–49.
[3] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.
[4] Feng RM, Zong YN, Cao SM, Xu RH. Current cancer situation in China: good or bad news from the 2018 Global Cancer Statistics? Cancer Commun (Lond) 2019;39(1):22.
[5] Liang J, Ye G, Guo J, Huang Q, Zhang S. Reducing false-positives in lung nodules detection using balanced datasets. Front Public Health 2021;9:671070.
[6] Ost D, Goldberg J, Rolnitzky L, Rom WN. Survival after surgery in stage IA and IB non-small cell lung cancer. Am J Respir Crit Care Med 2008;177(5):516–23.
[7] Liu Y, Wang H, Li Q, McGettigan MJ, Balagurunathan Y, Garcia AL, et al. Radiologic features of small pulmonary nodules and lung cancer risk in the national lung screening trial: a nested case-control study. Radiology 2018;286(1):298–306.
[8] National Lung Screening Trial, Research T, Aberle DR, Adams AM, Berg CD, Black WC, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
[9] National Lung Screening Trial, Research T, Church TR, Black WC, Aberle DR, Berg CD, et al. Results of initial low-dose computed tomographic screening for lung cancer. N Engl J Med 2013;368(21):1980–91.

[10] de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med 2020;382(6):503–13.
[11] Huda W. Radiation doses and risks in chest computed tomography examinations. Proc Am Thorac Soc 2007;4(4):316–20.
[12] de Jong PA, Mayo JR, Golmohammadi K, Nakano Y, Lequin MH, Tiddens HA, et al. Estimation of cancer mortality associated with repetitive computed tomography scanning. Am J Respir Crit Care Med 2006;173(2):199–203.
[13] Sodhi KS, Lee EY. What all physicians should know about the potential radiation risk that computed tomography poses for paediatric patients. Acta Paediatr 2014; 103(8):807–11.
[14] Dudley DJ. The immune system in health and disease. Baillieres Clin Obstet Gynaecol 1992;6(3):393–416.
[15] Rudolph MG, Stanfield RL, Wilson IA. How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol 2006;24:419–66.
[16] Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res 2009; 19(10):1817–24.
[17] Nielsen SCA, Boyd SD. Human adaptive immune receptor repertoire analysis-Past, present, and future. Immunol Rev 2018;284(1):9–23.
[18] Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alphabeta T cell receptor diversity. Science 1999;286 (5441):958–61.
[19] Nikolich-Zugich J, Slifka MK, Messaoudi I. The many important facets of T-cell repertoire diversity. Nat Rev Immunol 2004;4(2):123–32.
[20] Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity 2000;13(1):37–45.
[21] Luo B, Chu X, Yu P, Tian J. The TCR repertoire diversity and its application in the prevention and treatment of lung cancer. Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 2022;38(10)):939–43.
[22] Han J, Duan J, Bai H, Wang Y, Wan R, Wang X, et al. TCR Repertoire Diversity of Peripheral PD-1(+)CD8(+) T Cells Predicts Clinical Outcomes after Immunotherapy in Patients with Non-Small Cell Lung Cancer. Cancer Immunol Res 2020;8(1):146–54.
[23] Wang X, Zhang B, Yang Y, Zhu J, Cheng S, Mao Y, et al. Characterization of Distinct T Cell Receptor Repertoires in Tumor and Distant Non-tumor Tissues from Lung Cancer Patients. Genom Proteom Bioinforma 2019;17(3):287–96.
[24] Ma J, Sun G, Zhu P, Liu S, Ou M, Chen Z, et al. Determination of the complexity and diversity of the TCR beta-chain CDR3 repertoire in bladder cancer using high-throughput sequencing. Oncol Lett 2019;17(4):3808–16.
[25] Wang T, Wang C, Wu J, He C, Zhang W, Liu J, et al. The Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues. Cancer Immunol Res 2017;5(2):148–56.
[26] Chen YT, Hsu HC, Lee YS, Liu H, Tan BC, Chin CY, et al. Longitudinal High-Throughput Sequencing of the T-Cell Receptor Repertoire Reveals Dynamic Change and Prognostic Significance of Peripheral Blood TCR Diversity in Metastatic Colorectal Cancer During Chemotherapy. Front Immunol 2021;12:743448.
[27] de la Cruz-Merino L, Grande-Pulido E, Albero-Tamarit A, Codes-Manuel de Villena ME. Cancer and immune response: old and new evidence for future challenges. Oncologist 2008;13(12):1246–54.
[28] Gooden MJ, de Bock GH, Leffers N, Daemen T, Nijman HW. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. Br J Cancer 2011;105(1):93–103.
[29] Zhuo Y, Yang X, Shuai P, Yang L, Wen X, Zhong X, et al. Evaluation and comparison of adaptive immunity through analyzing the diversities and clonalities of T-cell receptor repertoires in the peripheral blood. Front Immunol 2022;13:916430.
[30] Chronister WD, Crinklaw A, Mahajan S, Vita R, Kosaloglu-Yalcin Z, Yan Z, et al. TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. Front Immunol 2021;12:640725.
[31] Kan CFK, Unis GD, Li LZ, Gunn S, Li L, Soyer HP, et al. Circulating Biomarkers for Early Stage Non-Small Cell Lung Carcinoma Detection: Supplementation to Low-Dose Computed Tomography. Front Oncol 2021;11:555331.
[32] Burnet FM. Self-recognition" in colonial marine forms and flowering plants in relation to the evolution of immunity. Nature 1971;232(5308):230–5.
[33] Han Y, Liu X, Wang Y, Wu X, Guan Y, Li H, et al. Identification of characteristic TRB V usage in HBV-associated HCC by using differential expression profiling analysis. Oncoimmunology 2015;4(8):e1021537.
[34] Cui JH, Lin KR, Yuan SH, Jin YB, Chen XP, Su XK, et al. TCR repertoire as a novel indicator for immune monitoring and prognosis assessment of patients with cervical cancer. Front Immunol 2018;9:2729.
[35] Sherwood AM, Emerson RO, Scherer D, Habermann N, Buck K, Staffa J, et al. Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. Cancer Immunol Immunother 2013;62(9):1453–61.
[36] Gierada DS, Rydzak CE, Zei M, Rhea L. Improved Interobserver Agreement on Lung-RADS classification of solid nodules using semiautomated CT volumetry. Radiology 2020;297(3):675–84.
[37] Sundaram V, Gould MK, Nair VS. A comparison of the pancan model and lung-RADS to assess cancer probability among people with screening-detected, solid lung nodules. Chest 2021;159(3):1273–82.
[38] MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the fleischner society 2017. Radiology 2017;284(1):228–43.
[39] Jamshidi A, Liu MC, Klein EA, Venn O, Hubbell E, Beausang JF, et al. Evaluation of cell-free DNA approaches for multi-cancer early detection. Cancer Cell 2022;40 (12):1537–49. e12.
[40] Mattox AK, Douville C, Wang Y, Popoli M, Ptak J, Silliman N, et al. The origin of highly elevated cell-free DNA in healthy individuals and patients with pancreatic, colorectal, lung, or ovarian cancer. Cancer Discov 2023.
[41] Luo H, Zu R, Huang Z, Li Y, Liao Y, Luo W, et al. Characteristics and significance of peripheral blood T-cell receptor repertoire features in patients with indeterminate lung nodules. Signal Transduct Target Ther 2022;7(1):348.