



iATMEcell: identification of abnormal tumor microenvironment cells to predict the clinical outcomes in cancer based on cell–cell crosstalk network

Yuqi Sheng[†], Jiashuo Wu[†], Xiangmei Li[†], Jiayue Qiu, Ji Li, Qinyu Ge, Liang Cheng  and Junwei Han 

Corresponding authors: Junwei Han, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China.

E-mail: hanjunwei@ems.hrbmu.edu.cn; Qinyu Ge, College of Biological Science & Medical Engineering, Southeast University, Nanjing 210096, China.

E-mail: geqinyu@seu.edu.cn; Liang Cheng, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China.

E-mail: liangcheng@hrbmu.edu.cn

[†]Yuqi Sheng, Jiashuo Wu, Xiangmei Li contributed equally to this work.

Abstract

Interactions between Tumor microenvironment (TME) cells shape the unique growth environment, sustaining tumor growth and causing the immune escape of tumor cells. Nonetheless, no studies have reported a systematic analysis of cellular interactions in the identification of cancer-related TME cells. Here, we proposed a novel network-based computational method, named as iATMEcell, to identify the abnormal TME cells associated with the biological outcome of interest based on a cell–cell crosstalk network. In the method, iATMEcell first manually collected TME cell types from multiple published studies and obtained their corresponding gene signatures. Then, a weighted cell–cell crosstalk network was constructed in the context of a specific cancer bulk tissue transcriptome data, where the weight between cells reflects both their biological function similarity and the transcriptionally dysregulated activities of gene signatures shared by them. Finally, it used a network propagation algorithm to identify significantly dysregulated TME cells. Using the cancer genome atlas (TCGA) Bladder Urothelial Carcinoma training set and two independent validation sets, we illustrated that iATMEcell could identify significant abnormal cells associated with patient survival and immunotherapy response. iATMEcell was further applied to a pan-cancer analysis, which revealed that four common abnormal immune cells play important roles in the patient prognosis across multiple cancer types. Collectively, we demonstrated that iATMEcell could identify potentially abnormal TME cells based on a cell–cell crosstalk network, which provided a new insight into understanding the effect of TME cells in cancer. iATMEcell is developed as an R package, which is freely available on GitHub (<https://github.com/hanjunwei-lab/iATMEcell>).

Keywords: tumor microenvironment, cell–cell crosstalk network, network propagation algorithm, prognostic biomarker

Introduction

Tumor microenvironment (TME) cells contain many different non-cancerous cell types in addition to cancer cells, such as immune cells, stromal cells, epithelial cells, etc., which have been widely implicated in tumorigenesis, prognosis and response to immunotherapy [1, 2]. For instance, an increase of CD8⁺ T cells and CD4⁺ T cells has been correlated with improved clinical outcomes and response to immunotherapy in various cancers, such as stomach cancer, melanoma, urothelial cancer, lung cancer and breast cancer [3]. Tumor-associated macrophages and regulatory T cells have both been linked to pro-tumor activities [4, 5]. B cells and natural killer (NK) cells have been demonstrated to have a good or negative impact on cancer patients' prognosis in

different cancers [6]. Furthermore, depending on tumor histology, the significance of different TME cells in different tissues varies widely. Certain cells are associated with improved survival in some cancer types. While in other cancer types, they may act with the opposite effects. For example, M1 macrophages are positively associated with longer survival times and most positive clinical outcomes in colorectal cancer, ovarian cancer and breast cancer; however, in some cancers such as renal cell carcinoma and melanoma, the presence of M1 macrophages is associated with a poor prognosis due to the interaction between M1 macrophages and M2 macrophages [7]. The biological mechanisms of the TME driving these responses are not yet fully understood. Thus, identifying abnormal cells associated with disease states (such as

Yuqi Sheng is a master at College of Bioinformatics Science and Technology, Harbin Medical University. Her research interests focus on bioinformatics.

Jiashuo Wu is a doctor at the College of Bioinformatics Science and Technology, Harbin Medical University. His research interests focus on computational system biology.

Xiangmei Li is a doctor at the College of Bioinformatics Science and Technology, Harbin Medical University. Her research interests focus on bioinformatics.

Jiayue Qiu is a master at College of Bioinformatics Science and Technology, Harbin Medical University. His research interests focus on computational system biology.

Ji Li is a master at College of Bioinformatics Science and Technology, Harbin Medical University. His research interests focus on bioinformatics.

Qinyu Ge is an associate professor at the College of School of Biological Science & Medical Engineering, Southeast University. His research interests focus on practical biochips, high-throughput sequencing and bioinformatics.

Liang Cheng is a professor and principal investigator at College of Bioinformatics Science and Technology, NHC Key Laboratory of Molecular Probes and Targeted Diagnosis and Therapy, Harbin Medical University.

Junwei Han is a professor and principal investigator at College of Bioinformatics Science and Technology, Harbin Medical University. His research interests focus on bioinformatics and computational system biology.

Received: November 5, 2022. Revised: January 13, 2023. Accepted: February 9, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

patient survival) is essential to understanding the effect of TME on cancer progression and therapeutic responses.

Recently, using gene transcriptome data from bulk tumors, numerous computational approaches have been developed to determine the relative infiltration levels of different TME cells. For example, CIBERSORT mainly applies the deconvolution algorithm to estimate the TME cell fractions; single sample gene set enrichment analysis (ssGSEA) calculates the enrichment scores of cell-type-specific marker gene sets to infer the cell abundance; xCell combines gene set enrichment and deconvolution techniques to count the number of different cell types [8–10]. According to these methods, we could analyze the TME cells with an abnormal infiltration level linked to malignancy and discover novel immunotherapeutic biomarkers. Despite this, no systematic examination of cellular interactions in the identification of cancer-related TME cells.

Some studies have revealed that the cell–cell interactions in the TME may drive cancer progression and influences therapeutic efficacy [11]. Recently, network modeling has been applied to cell–cell interaction analysis, mainly by calculating the expression and pairing of receptors and ligands in different cell types, or based on the concept of gene co-expression to construct networks and thus infer the interactions between different cells. For example, CellPhoneDB mainly obtains information about the interaction between different cell types through the expression of receptors of one cell type and ligands of another cell type [12]. iTALK takes cell populations as the object of interaction, calculates the expression of receptor ligands in each cell subpopulation, and uses this as an indicator of interaction to study the cell–cell communication between subpopulations [13]. Although these methods are useful for exploring cellular interactions in TME, they were only applicable to single-cell ribonucleic acid sequencing (scRNA-seq) and could not be applied to existing large-scale bulk tumor datasets, which represent a vast and mostly unexplored resource in cancer TME studies. Moreover, such methods often ignore the biological functional similarities between cells. Some studies confirmed that the interactions between different cells in the TME were very similar to normal physiological processes and aimed at providing essential materials for tumor growth [11]. Constructing cellular interaction networks based on the biological functions involved in the cells to deeply investigate the role of TME cells may provide new perspectives for the development of new drug targets and cancer therapy. Thus, identifying abnormal TME cells by considering cell–cell interactions in the context of a specific cancer type may provide some new insight into the mechanisms of TME cells.

Here, we proposed a novel computational approach, called iATMEcell, to identify abnormal TME cells associated with the biological outcome of interest (e.g. dead/alive) based on a weighted cell–cell interaction (hereafter called cell–cell crosstalk) network. iATMEcell mainly has two features: (i) a weighted cell–cell crosstalk network was constructed based on the biological functions shared by cells; (ii) abnormal TME cells was identified in the context of a specific bulk tumor transcriptome data. In iATMEcell, we first manually collected TME cell types from multiple published studies and obtained their corresponding gene signatures. Then, a weighted cell–cell crosstalk network was constructed based on the gene transcriptome data for a pair of binary conditions (e.g. dead/alive), where the weight between cells reflects both their biological function similarity and the transcriptional dysregulated activities of gene signatures shared by them. We then used a network propagated algorithm to calculate the centrality scores of cells to identify the abnormal cells linked to cancer. The statistical

significance of the cell centrality score was evaluated with a bootstrap-based randomization method. We applied iATMEcell to the TCGA-Bladder Urothelial Carcinoma (BLCA) dataset and identified NK cells to be significantly abnormal in BLCA. Based on the signature genes of NK cells, a risk score model was constructed, which could efficiently classify patients into high-risk and low-risk groups. Moreover, we applied the method to multiple cancer datasets, respectively, using the abnormal cells to create biomarkers and uncovering unanticipated pan-cancer similarities. Overall, identification of abnormal TME cells based on the cell crosstalk network is of great importance for understanding the mechanism of TME, which may provide a new research direction for the TME research. iATMEcell has been developed as an R package, which is freely available on GitHub (<https://github.com/hanjunwei-lab/iATMEcell>).

Materials and methods

Data acquisition and preprocessing

We constructed a cell–cell crosstalk network in the context of cancer bulk tissue transcriptome data. To construct the network, we need three data types: (i) TME cell type-specific gene signature sets retrieved from the published studies and existing TME cell estimation methods; (ii) biological function data from the Gene Ontology (GO) database; (iii) gene transcriptional data with two different conditions (e.g. alive/dead or normal/diseased).

Cell type-specific signature gene sets

The composition of the TME varies between tumor types, but hallmark features include immune cells, stromal cells, blood vessels and extracellular matrix. In this study, TME cells mainly include both stromal cells (endothelial cells, fibroblasts cells, etc.) and immune cells, which include adaptive response cells (B cells, T cells, etc.) and innate response cells [NK cells, macrophages, Dendritic cells (DCs), etc.] (Figure 1A). In the clinical practice, TME cells work together to protect us from infection and cancer [3]. To obtain cell type-specific gene signatures, we collected the gene signature sets from 12 sources, including the published cell signature sets (Bindea et al. [14], Charoentong et al. [15], Danaher et al. [16], Davoli et al. [17], He et al. [18], Rooney et al. [19], Tirosh et al. [20]) and the TME cell estimation methods (MCP-counter [21], EPIC [22], ImmuCellAI [9], TIDE [23] and xCell [10]). To make the cell type-specific gene sets as complete as possible, we integrated the above data sources, and if a cell had a corresponding gene set in more than one source, their union set was used. A total of 86 cell type-specific gene signature sets were obtained, including 60 immune cells and 26 stromal cells (Supplementary Table S1).

Biological function data

Biological function data was derived from GO biological processes. In the ‘gene ontology’ term, a biological process represents a specific objective that the organism is genetically programmed to achieve [24]. The biological process gene sets were downloaded from C5 GO gene sets in the Molecular Signatures Database database [25]. We then manually curated the GO gene sets associated with human immune function, which were deposited in our ‘iATMEcell’ package.

Gene expression profiles

The cell crosstalk network was constructed in the context of cancer bulk tissue transcriptome data. We collected 10 cancer types that have been proposed to be suitable for immunotherapy from TCGA, including BLCA, skin cutaneous melanoma (SKCM),

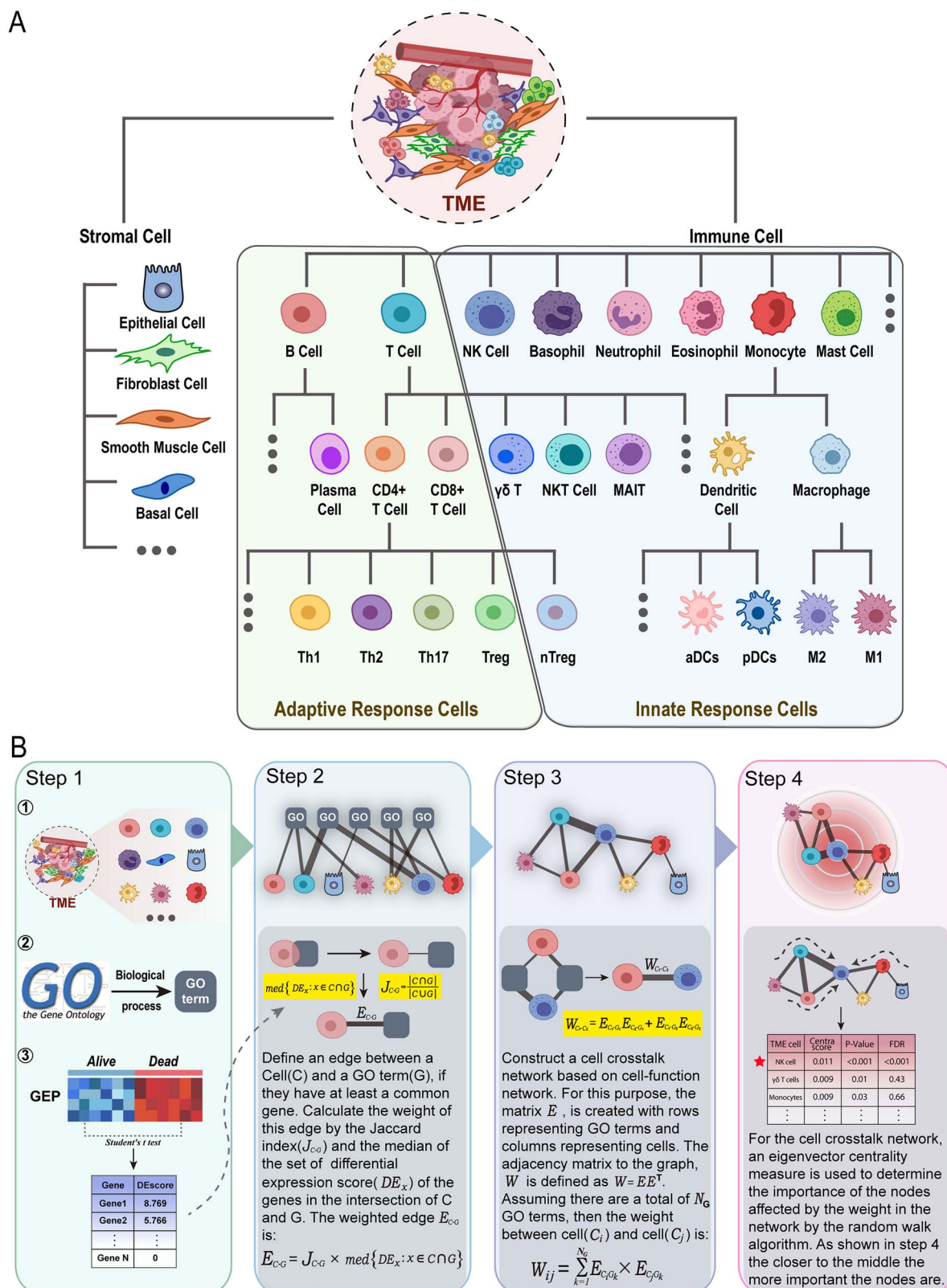


Figure 1. (A) TME cell type classification. (B) Schematic diagram of the iATMEcell method.

lung adenocarcinoma (LUAD), stomach adenocarcinoma (STAD), colon adenocarcinoma (COAD), liver hepatocellular carcinoma (LIHC), breast invasive carcinoma (BRCA), esophageal carcinoma (ESCA), cervical squamous cell carcinoma, endocervical adenocarcinoma (CESC) and kidney renal clear cell carcinoma (KIRC).

The normalized RNA-seq (Fragments per Kilobase per Million, FPKM) data and clinical information for these cancers were downloaded from the GDC TCGA data portal (<https://portal.gdc.cancer.gov>). For the gene expression data, the FPKM profiles of each gene were transformed by $\log_2(\text{FPKM}+1)$, which were then

normalized by z-score. Thus, the normalized gene expression values approximately obey normal distribution. Moreover, to validate our results, we collected two other independent bladder cancer cohorts receiving immunotherapy, the IMvigor210 [26] and GSE176307 [27] cohorts. The gene expression data and detailed clinical information were downloaded from <http://research-pub.gene.com/IMvigor210CoreBiologies/> and the GEO database (access no. GSE176307).

Calculating the gene differential expression level

We conducted a statistical comparison of gene expression values between case and control groups (e.g. dead and alive) (step1 in Figure 1B). In this case, we use Student's t-test method to calculate the differential expression level for each gene. Next, we convert the t-test P-value of each gene into a z-score using $z = \varphi^{-1}(1 - p)$, where φ^{-1} is the inverse normal cumulative density function. The z-score of each gene was defined as a differential expression score (DEscore), and a larger DEscore indicates a greater difference in gene expression between the two groups of samples.

Constructing a cell-GO bipartite network

To construct a cell-cell crosstalk network, we first constructed a cell-GO network (step 2 in Figure 1B), which is a bipartite network and the two sets of nodes are cells and GO terms. We think that two cells are functionally similar if they participate in at least a common GO term. To do this, we defined a weighted edge between each pair of cells (C) and GO term (G), if the intersection of genes between C and G is nonempty. In parallel, we assigned the edge weight based on two indicators: (i) the degree of overlap between C and G; (ii) the degree of transcriptional dysregulation of the intersection genes between C and G. The formula of the weight is as follows:

$$E_{C,G} = J_{C,G} \times \text{med} \{DEscore_x | x \in C \cap G\} \quad (1)$$

where $J_{C,G}$ is the Jaccard index between C and G, and $\text{med} \{DEscore_x | x \in C \cap G\}$ is the median DEscore of the intersection genes between C and G.

The Jaccard coefficient is a standard measure of similarity between sets, in this case, the value of $J_{C,G}$ indicates the extent to which cell C is involved in the biological function G. Thus, in our approach, the edge weight between C and G is jointly determined by the participation degree of cell C in the biological function G and differentially expressed levels of genes that common to C and G. In the context of a specific disease gene expression data, a higher weight suggests a cell that is more involved in a GO term in the disease.

Converting the bipartite network to cell-cell crosstalk network

We then constructed the cell-cell crosstalk network through the cell-GO network (step 3 in Figure 1B). The cell-GO network is a bipartite network, that can be represented algebraically in the form of an incidence matrix $E = [E_{ij}]$, where the rows of the matrix represent cells and the columns represent GO terms, and its elements reflect the weights between each pair of cells and GO terms. Two cells share more neighbor GO nodes in the bipartite network; they tend to perform comparable biological functions and communicate with one another. We thus constructed the cell-cell crosstalk network by defining a weighted adjacency matrix:

$$W = E \times E^T \quad (2)$$

The value of $W_{i'}$ denoted weight of each edge in the cell crosstalk network:

$$W_{i'} = \sum_{j=1}^{N_G} E_{C_i G_j} \times E_{C_{i'} G_j} \quad (3)$$

where the N_G is the total number of GO terms; the edge weight ($W_{i'}$) between cell i (C_i) and cell i' ($C_{i'}$) is the sum of the weight of the two cells with common GO terms, which means the edge weight between two cells is equal to the sum of their contributions to the transcriptional differences of all biological functions (GO terms) they share. Based on this principle, when and only when two cells share at least one GO term, they are linked in this network. As a result, a weighted cell crosstalk network is constructed, with self-interactions deleted. This process was previously used to construct a pathway crosstalk network [28], we used it to construct a cell crosstalk network.

Identifying the significant abnormal cells by using a network propagation algorithm

From steps 1 to 3, we construct a weighted cell crosstalk network based on the cellular functional similarity. The weighted edges reflect the evidence that a cell may be potentially altered, which can be expected to be reinforced by the evidence of its neighbors. To determine the significant abnormal cells that are influenced in the TME, we used a network propagation algorithm to calculate the eigenvector centrality score, which is a measure to evaluate the influence extent of nodes in a network. The random walk with restart (RWR) algorithm specifically embodies this notion of importance and is used to calculate the centrality scores of cells in the crosstalk network. In this algorithm, we define a probability transition matrix for a random walker, P , by row-normalizing the adjacency matrix W :

$$P_{i'} = \frac{W_{i'}}{\sum_{i=1}^{N_C} W_{i'}} \quad (4)$$

where N_C is the number of cells in the cell crosstalk network; $P_{i'}$ is the probability that, starting at cell C_i , the next step will be to cell $C_{i'}$. Thus, the edge weight will bias the random walker in such a way that the walker proceeds along bigger weighted edges with greater probability than smaller ones. The formula for the RWR algorithm is:

$$e^{t+1} = (1 - \alpha) P e^t + \alpha P e^0 \quad (5)$$

where P is the row-normalized adjacency matrix of W , e^t is the vector of nodes at time step t . In the study, the RWR algorithm was applied to the weighted cell network to identify the important nodes that are more likely to be influenced if it is connected to a lot of other neighbors and the edges have great weight. Thus, e^0 , an initial probability vector, is created by assigning to each node the same value and making the sum equal to 1. The parameter α is the restart probability, and the default value is set at 0.9. It has been demonstrated to have only a slight effect on the results when it was set from 0.1 to 0.9 [29]. After an infinite number of walks, the probability e^t will converge to a stable state e , whose i th element e_i is defined as the eigenvector centrality scores of the cell C_i . The eigenvector centrality score of cell is influenced not only by the number of neighbors but also by the weights on the edges linked to it. However, for some cells with a larger degree, it is possible to have their significance exaggerated by the original eigenvector centrality scores.

To rectify this, we use a bootstrap-based randomization procedure to estimate the statistical significance of cell centrality scores. In detail, we use bootstrap resampling of our initial DEscore at the gene level to become a set of transcriptional hypothetical data. We generate a set of hypothetical DEscore data, $DEscore_x^*$, for all genes x and apply the above algorithm (step 1 through 4) to obtain the hypothetical eigenvector centrality scores of all cells, e^* . This process was repeated 1000 times to produce a set of centrality score vectors $\{e^{*1}, e^{*2}, \dots, e^{*1000}\}$. According to the law of large numbers, all centrality scores for each cell in this collection can be considered to follow a normal distribution. We calculated a P -value for cell i by:

$$p - value = 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{C_i} \exp\left(-\frac{(C_i - \mu)^2}{2\sigma^2}\right) dC_i \quad (6)$$

where μ and σ are the average value and standard deviation of the random centrality scores and C_i is the origin eigenvector centrality score of cell i . The resulting ranking of cells follows the ranking of P -value, with the lower P -value indicating the greater statistically significant difference of cells. The P -value was adjusted by the false discovery rate (FDR) method proposed by Benjamin and Hochberg [30]. At present, the iATMEcell method has been developed as a freely available R package on the GitHub repository (<https://github.com/hanjunwei-lab/iATMEcell>).

Constructing the prognostic model based on the signature genes of abnormal cells

In order to test the impact of the significant abnormal TME cells on cancer prognosis, we conducted a survival analysis on the signature genes of abnormal cells. For each abnormal cell, we assessed the prognostic role of the cell signature genes based on univariate Cox regression analysis, and prognostic related genes were selected at a threshold of $P < 0.05$. Subsequently, due to the presence of multicollinearity among the gene variables, the prognosis-related genes were further downsampled using the least absolute shrinkage and selection operator (LASSO) regression analysis [31]. The LASSO regression algorithm uses the L1 parametric shrinkage penalty to penalize some variables that do not contribute much to the dependent variable, thus retaining the significant variables. This analysis was performed with R software based on the R package 'Glmnet'. The identified genes by the LASSO regression were used to construct a cell-specific risk score model based on their corresponding coefficients, and the risk score for each sample was calculated as follows:

$$\text{Risk score} = \sum_{k=1}^n \beta_k \times GE_k \quad (7)$$

where n is the number of selected prognosis-related cell signature genes in the LASSO regression, GE_k is the gene expression value of gene k , and β_k is the coefficient of gene k generated from the LASSO regression analysis. All patients were divided into high- and low-risk groups by the median risk score. The log-rank test and Kaplan–Meier survival analysis were used to test if the high-risk group and low-risk group show significant difference. Moreover, the 'survival' and 'survival ROC' R packages were used to generate the receiver operating characteristic curve (ROC) to evaluate the performance of prognostic classification of the risk score model.

Results

Identification of the abnormal TME cells in BLCA

TME contains multiple complex cell populations, each of which may be involved in tumor progression. Identifying abnormal TME cells can promote further understanding of the mechanisms of TME on tumor development. In this paper, the key purpose of iATMEcell is to identify abnormal TME cells by constructing a weighted cell–cell crosstalk network under bulk transcriptome data. In iATMEcell, TME cells will be ranked by their FDRs of eigenvector centrality scores calculated from bulk tissue transcriptome data under two different conditions (e.g. alive/dead).

To illustrate the effect of iATMEcell, we applied it to the BLCA dataset from the GDC TCGA database, which includes 406 samples ($N_{\text{live}} = 227$, $N_{\text{dead}} = 179$). With $FDR \leq 0.25$, iATMEcell identified five statistically significant abnormal TME cells (Table 1), which include NK cells, CD8+ Effector memory T Cell (CD8+ Tem), gamma delta ($\gamma\delta$) T cells, Monocytes, M1 Macrophages. Some evidence for the biological significance of these potentially abnormal cells has been found in several literatures. For example, NK cell was identified to be the most statistically significant by iATMEcell, and which was a type of cytotoxic lymphocyte critical to the innate immune system [32]. The relationship between NK cells activity and inhibition of tumorigenesis has been demonstrated in mouse models [33]. Meanwhile, Sun *et al.* found that tumor expression of activated NK cell receptors was more favorable for BLCA prognosis [34]. Hartana *et al.* demonstrated tumor immune escape mechanisms that suppress CD8+ T cells cytotoxicity in urothelial bladder cancer [35]. Pan *et al.* found that T cell receptor-positive $\gamma\delta$ T cells exhibited NK cell-like phenotypic characteristic and showed that $\gamma\delta$ T cells has a powerful benefit in BLCA treatment [36], and which indicates the relationship between $\gamma\delta$ T cells and NK cells.

Construction of the prognostic model based on the abnormal cells

To investigate the prognostic effect of NK cells in BLCA, we performed the survival analysis based on the signature genes of NK cells. The signature genes of NK cells were obtained from 10 sources (Supplementary Table S1), including 195 genes in total. For identifying prognosis-related genes, univariate Cox regression analysis and the LASSO regression analysis along with 10-fold cross validation was performed on the gene expression and overall survival (OS) data. Thus, 23 significant genes were obtained (Figure 2A and Supplementary Figure S1A), of which 14 were prognostic protective factors (Hazard ratio (HR) < 1) and nine were risk factors (HR > 1) (Figure 2B). With these genes, the NK cells-based gene risk score model was constructed using a formula derived from the expression of the genes weighted by their LASSO regression coefficients (Supplementary Figure S1B): $\text{risk score} = \sum_{k=1}^{23} \beta_k \times \text{expression value of gene } k$. According to the median values of risk scores, the BLCA patients were categorized into the high-risk and low-risk groups. The Kaplan–Meier survival curves showed that patients in the low-risk group had significantly longer OS than that in the high-risk group (Figure 2C; log-rank P -value < 0.0001). Moreover, the expression values of these genes show significant differences between high-risk and low-risk groups (Figure 2D). Furthermore, time-dependent ROC curve analyses were used to evaluate the prognostic power of the risk score model. The area under the ROC curve (AUC) for 1-, 3-, and 5-years OS were 0.75, 0.77 and 0.79, respectively (Figure 2E). Finally, we compared the abundance of NK cells infiltration using the xCell method [10], and interestingly,

Table 1. Significant abnormal TME cells identified by iATMEcell (FDR \leq 0.25)

Rank	Cell	Cell full name	Size ^a	P-value	FDR
1	NK cells	Natural Killer Cells	195	0.001	0.09
2	CD8+ Tem	Effector memory CD8+ T cell	329	0.003	0.12
3	$\gamma\delta$ T cells	Gamma Delta T cells	254	0.01	0.25
4	Melanocytes	Melanocytes	241	0.01	0.25
5	Monocytes	Monocytes	353	0.01	0.25

^aThe number of signature genes for a cell.

Table 2. Univariable and multivariable Cox analysis of the risk score and clinicopathological factors (Age, Sex, TMB and Stage) for OS in BLCA

	Univariable analysis			Multivariable analysis		
	HR	95% CI	P-value	HR	95% CI	P-value
TCGA BLCA dataset						
Risk score (high versus low)	3.3	2.6–4.2	<0.01	2.8	2.2–3.7	<0.01
Age (\geq 65 versus <65)	1.8	1.3–2.4	<0.01	1.6	1.1–2.2	<0.01
Sex (male versus female)	0.88	0.64–1.2	0.44	1	0.7–1.4	0.99
TMB	0.92	0.88–0.96	<0.01	0.94	0.9–0.98	<0.01
T stage (T0–T4)	1.7	1.4–2.1	<0.01	1.3	1.1–1.6	0.012

we found that the NK cell infiltration abundance was significantly lower in the high-risk group samples than that of the low-risk group (Figure 2F). These results demonstrated that the iATMEcell method could identify significant abnormal TME cells associated with patients' survival states (dead and alive) and the risk score model constructed with the cell signature genes may serve as a potential prognostic biomarker. In addition, we found that the risk score is an independent prognostic factor after adjusting for age, sex, TMB and T stage by the multivariable Cox regression analysis [HR = 2.8, 95% confidence interval (CI), 2.2–3.7, $P < 0.01$, Table 2].

To explore the biological functions that may be potentially induced by the NK cells, we performed the GSEA analysis to identify differentially activated pathways between high-risk and low-risk groups. The ridgeline plot shows the expression distribution of the core enriched genes for the top 30 significantly differential signaling pathways (Figure 3A). Interestingly, we found that the phosphatidylinositol 3'-kinase (PI3K)-Akt signaling pathway, mitogen-activated protein kinase (MAPK) signaling pathway and extracellular matrix (ECM)-receptor interaction, etc. were mainly enriched in the high-risk group, and antigen processing and presentation, RIG-I-like receptor signaling pathway, etc. were mainly enriched in the low-risk group (Figure 3B). These results are in line with our expectations that most of these pathways have important links to BLCA. For example, the PI3K-Akt signaling pathway is an intracellular signal transduction pathway that promotes metabolism, proliferation, cell survival, growth and angiogenesis in response to extracellular signals. Stefanos *et al.* showed that PI3K-Akt pathway activation was crucial for bladder cancer initiation and progression [37]. Liu *et al.* established the central role of the MAPK pathway in bladder tumorigenesis [38]. Moreover, antigen processing and presentation are the cornerstones of adaptive immunity, and enrichment of antigen processing and presentation related genes may activate some adaptive immune cells to regulate immune response [39]. This explains why the low-risk group patients have a relatively better prognosis.

Moreover, cancer immunotherapy by immune checkpoint blockade has emerged as an important therapeutic approach to treat BLCA. We thus compared the expression distribution of immune checkpoint genes (ICGs) between high-risk and

low-risk groups. We found that four ICGs: CD96, PDCD1, CTLA4 and PDCD1LG2 presented significant differentially expressed. Specifically, CD96, PDCD1 and CTLA4 expression values showed higher in the low-risk group than that in the high-risk group, while the expression of PDCD1LG2 was the opposite (t-test, P -value <0.05, Figure 3C), which is consistent with previous reports that PDCD1LG2 (known as PD-L2) plays an important role in negative regulation of the adaptive immune response [40]. These results indicate that the low-risk group of patients may be more suitable for immunotherapy.

Validation of the prognostic model of abnormal cells in the IMvigor210 cohort

To validate the prognosis effect of the NK cell-based risk score model constructed based on TCGA-BLCA, we used the IMvigor210 cohort as a validation set [26]. The IMvigor210 cohort includes a total of 298 BLCA patients treated with anti-PD-L1, which provides expression data, OS data and immunotherapy response data. We applied the risk score model to the IMvigor210 cohort, and 15 genes were retained in the dataset as the expression values of some genes are missing. According to the median risk score, the patients were divided into high-risk and low-risk groups. The Kaplan-Meier survival curve analysis demonstrates significantly better OS in the low-risk group compared to the high-risk group (Figure 4A; log-rank P -value = 0.0031). The heatmap of these gene expressions showed a significant difference between high-risk and low-risk groups (Figure 4B), which was consistent with the results in the TCGA-BLCA dataset. Moreover, a scatter plot based on the risk score of the IMvigor210 cohort indicated that patients in the high-risk group had a higher risk score than those in the low-risk group (Figure 4C). With the same trend as the training set, the NK cell infiltration abundance was significantly lower in the high-risk group samples than in the low-risk group (Figure 4D). Furthermore, we analyzed the prognostic efficiency of the risk model by operating a ROC curve analysis, and the AUCs for 1- and 2-years OS were 0.6 and 0.78, respectively (Supplementary Figure S2).

Then, we applied the risk score model to test if it could predict the immunotherapy response. According to the Response

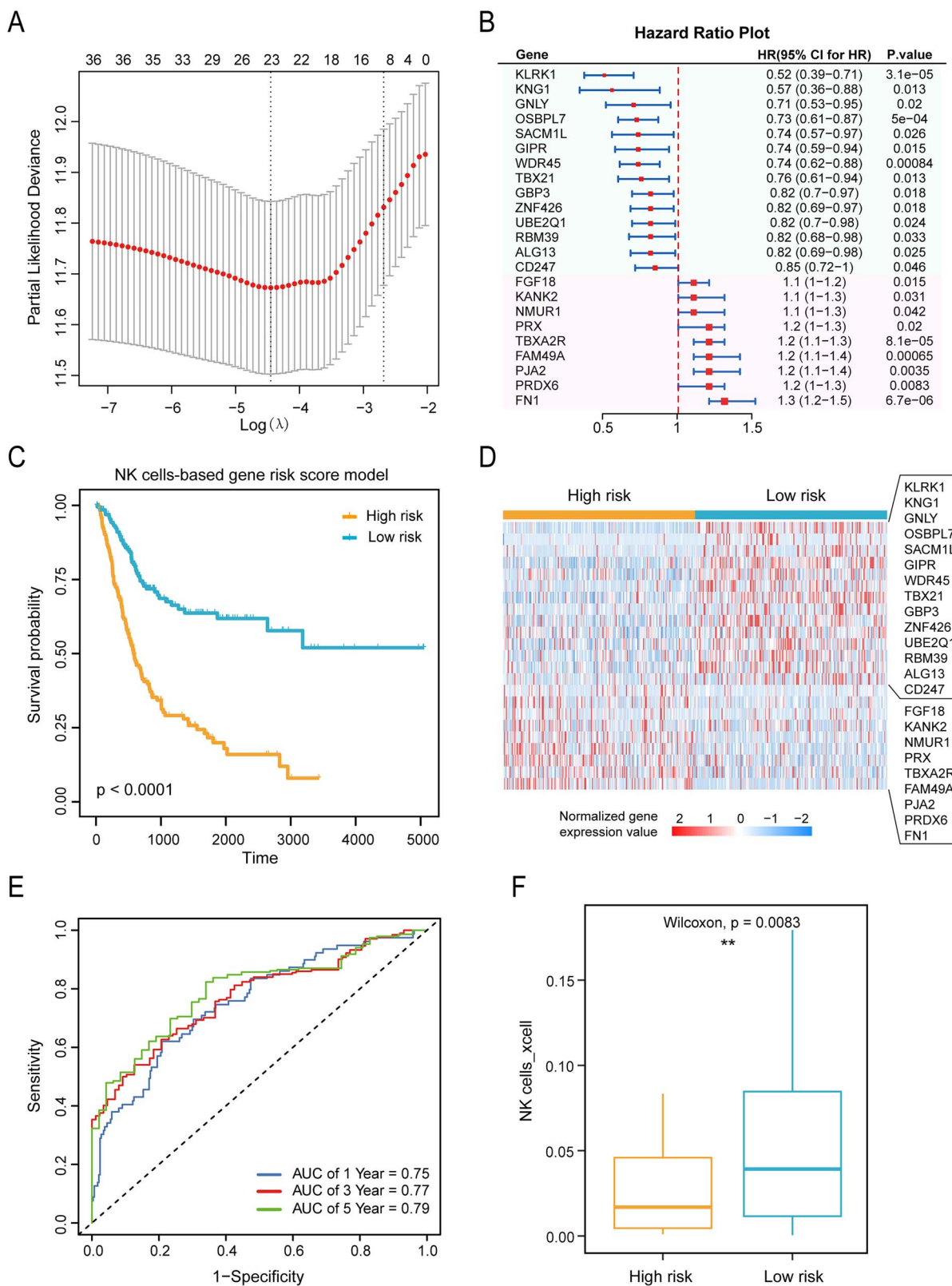


Figure 2. NK cells-based gene risk score model. **(A)** LASSO regression analysis of NK cells signature genes. **(B)** Forest plot shows HR, the 95% CI and P-values of the 23 genes selected by the LASSO regression analysis. **(C)** Kaplan-Meier survival curves of patients classified into high-risk and low-risk groups using the risk score model constructed based on NK cells. P-value was estimated using the log-rank test. **(D)** Heatmap of the normalized expression values of the 23 genes. **(E)** Time-dependent ROC curves for prognosis of the NK cells-based risk score model for 1-, 3- and 5- years OS in the TCGA BLCA dataset. **(F)** Box plots of NK cells infiltration abundance distributions for the high-risk and low-risk groups. The P-values were calculated with the Wilcoxon rank sum test.

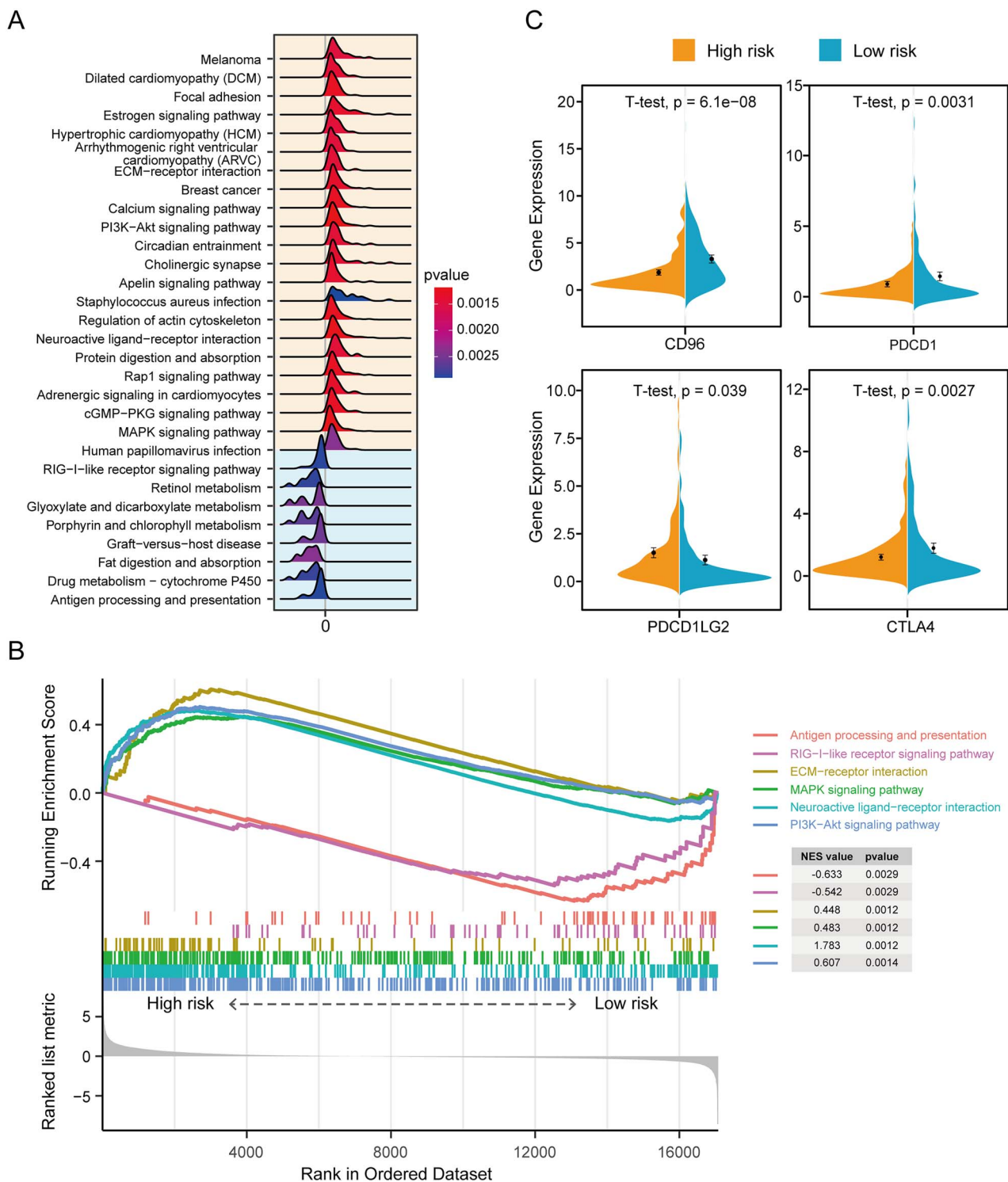


Figure 3. Biological pathways and genes analyses between high-risk and low-risk groups. **(A)** Ridgeline plot for the expression distribution of core genes for the top 30 significant pathways between high-risk and low-risk groups. The yellow area pathway is mainly enriched in the high-risk group and the blue area pathway is mainly enriched in the low-risk group. **(B)** GSEA plot of differentially activated pathways between high-risk and low-risk groups. **(C)** Split violin plots of four immune checkpoint gene expression distributions for the high-risk and low-risk groups. The P-values were calculated with the t-test.

Evaluation Criteria in Solid Tumors v1.1, the patients were characterized as response [complete response (CR)/partial response (PR)] or nonresponse [stable disease (SD)/progressive disease (PD)]. Interestingly, a significantly higher response rate was displayed in the low-risk group compared with the high-risk group (Figure 4E, Fisher's exact test, P -value=0.036), and the survival time of the

patients who responded well to the immunotherapy had a longer survival time (Figure 4F). Furthermore, the risk score was an independent prognostic factor after adjusting for sex, TMB and T staging (the data did not include age information) by multivariate Cox regression analysis (HR=1.27, 95% CI: 1.03–1.56, P -value=0.02, Table 3).

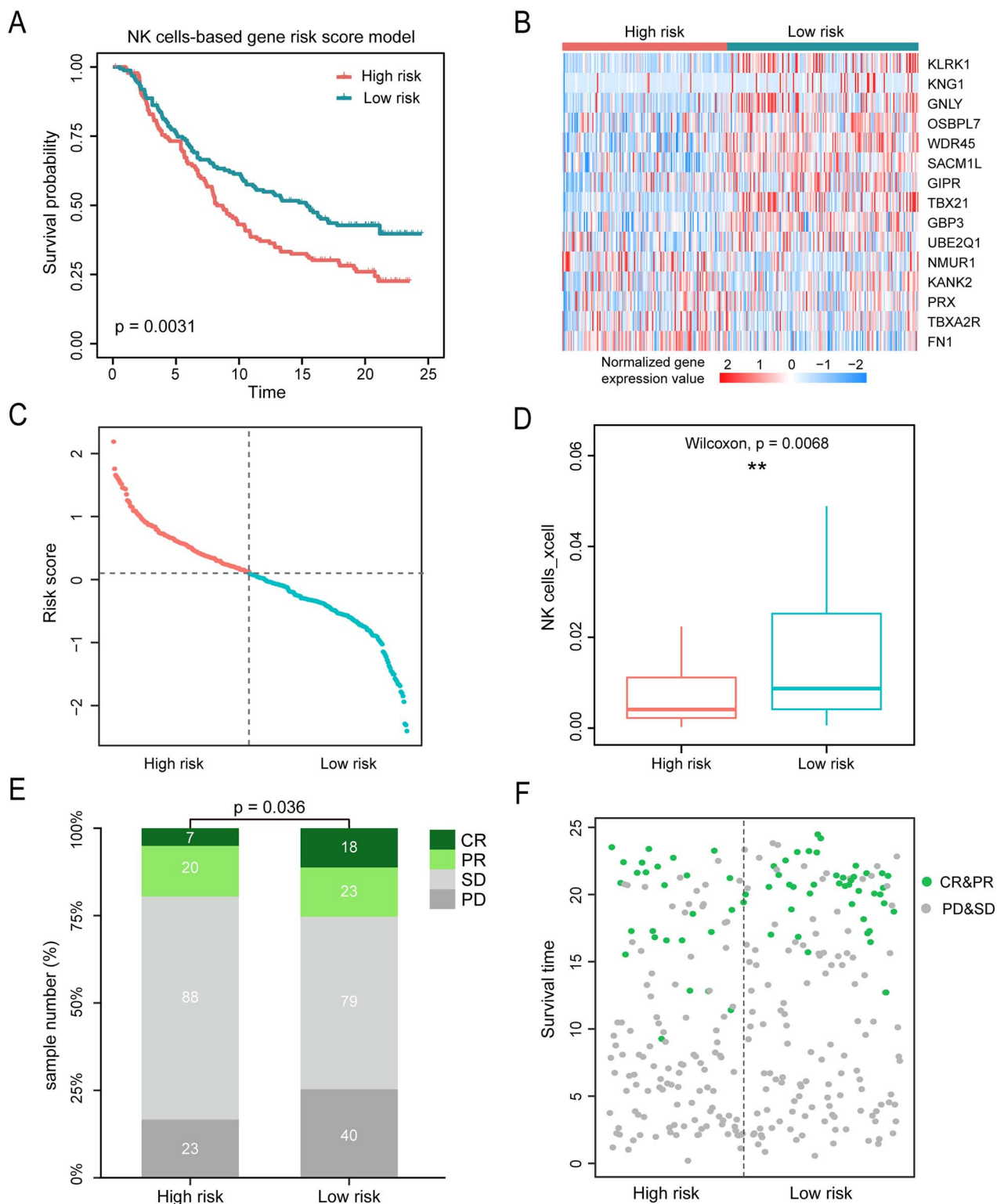


Figure 4. Validation of the risk score model in the IMvigor210 cohort. **(A)** Kaplan–Meier survival curves of patients classified into high-risk and low-risk groups using the NK cells-based gene risk score model. **(B)** Heat map of the normalized expression values of the 15 genes in the IMvigor210 cohort. **(C)** Risk score distribution in the high-risk and low-risk groups. **(D)** Box plots of NK cells infiltration abundance distributions for the high-risk and low-risk groups in the IMvigor210 cohort. The P -values were calculated with the Wilcoxon rank sum test. **(E)** The stacked bar chart shows the sample number of drug responses in the high-risk and low-risk groups. **(F)** Scatter plots show the distribution of survival times for the responders (including CR and PR) and non-responders (including SD and PD) in the high-risk and low-risk groups.

Table 3. Univariable and multivariable Cox analysis of the risk score and clinicopathological factors for OS in the IMvigor210 cohort

	Univariable analysis			Multivariable analysis		
	HR	95% CI	P-value	HR	95% CI	P-value
IMvigor210 cohort						
Risk score	1.4	1.1–1.7	<0.01	1.3	1.05–1.61	0.02
Sex (male versus female)	0.92	0.61–1.4	0.68	1.11	0.73–1.67	0.94
TMB	0.96	0.93–0.98	<0.01	0.96	0.93–0.98	<0.01
T stage (T1–T4)	1	0.87–1.2	0.94	0.97	0.84–1.13	0.72

Construction of an integrated risk score model based on the significant abnormal cells

Generally, the TME cells are not isolated, but crosstalk with each other to perform important biological processes. In iATMEcell, the abnormal cells were identified based on the cell–cell crosstalk network. To test the joint effect of abnormal cells, we thus constructed an integrated risk score model based on five significant abnormal cells identified in the TCGA–BLCA cohort (FDR ≤ 0.25 , Table 1). We first constructed the risk score model for each significant cell (see Materials and Methods) separately, and then an integrated risk score model was constructed by combining the risk scores of these cells with the LASSO regression. The forest plot shows that the risk scores of these cells were all prognostic risk factors (Supplementary Figure S3A, HR > 1). The Kaplan–Meier survival curves and time-dependent ROC curve showed that the integrated model had excellent prognostic efficacy (Supplementary Figure S3B; log-rank P-value < 0.0001) and the 1-, 3- and 5-years OS were 0.81, 0.80 and 0.81, respectively (Supplementary Figure S3C). We further validated the integrated model in two independent cohorts of bladder cancer receiving immunotherapy. In the IMvigor210 cohort, the overall survival was significantly better in the low-risk group compared with the high-risk group (Supplementary Figure S3D; log-rank P-value < 0.0001), and the AUCs for 1- and 2-years OS were 0.77, and 0.83 respectively (Supplementary Figure S3E). For the prediction of immunotherapy response, the response rate was significantly higher in the low-risk group compared with the high-risk group (Supplementary Figure S3F, Fisher's exact test, P-value < 0.001). Similarly, in the GSE176307 cohort, the integrated model also obtained good prognosis value (Supplementary Figure S3G, log-rank, P-value < 0.0001; Supplementary Figure S3H, the AUCs for 1- and 2-years OS were 0.82 and 0.92) and immunotherapy prediction value (Supplementary Figure S3I, Fisher's exact test, P-value = 0.019). The predicted efficacy of the integrated model indicates that the significant abnormal cells may jointly influence the development of cancer.

Identification of abnormal TME cells across multi-cancer types

To further explore the effect of TME cells in different cancers, we collected 10 cancer types that have been proposed to be suitable for immunotherapy from TCGA, including BLCA, SKCM, LUAD, STAD, COAD, LIHC, BRCA, ESCA, CESC and KIRC. The iATMEcell method was applied to these cancer types respectively to quantitatively determine the similarity of abnormal cells across different cancer types. To provide a general comparison, the top 10 most significant abnormal cells for each cancer were enrolled (Supplementary Table S2). Through comparison, five TME cells, including DCs, M1 Macrophages, NK cells, $\gamma\delta$ T cells and Monocytes, were shared across at least nine cancer types (Figure 5A). Interestingly,

these cells belong to innate immune response cells, which have been documented to play important roles in the induction of T-cell immunity [41].

Moreover, in each cancer type, we constructed the risk score models for the overlapped immune cells, respectively (see Materials and Methods). Through univariate Cox regression analysis, the risk score models of the five cells were found to be associated with patient survival in their corresponding cancer types (HR > 1, P-value < 0.01) (Figure 5B). The forest plot showed detailed HR information of these cells across these cancers (Figure 5C and Supplementary Figure S4). These overlapped abnormal immune cells may provide a new research direction for learning the mechanism of TME and highlight the repurposing potential of anti-cancer drugs targeting the cells.

Discussion

Infiltrated differences of TME cells (immune cells, stromal cells, epithelial cells, etc.) are correlated with clinical outcomes in a variety of malignancies, such as SKCM, LUAD, STAD and COAD [1, 2]. Thus, identification of abnormal TME cells may help to insight into the crucial role of TME in cancer progression and therapeutic responses. Some computational approaches, such as CIBERSORT, ssGSEA and xCell, were developed to infer the relative infiltration levels of TME cells. With these methods, many TME cells were identified to be associated with the survival and immunotherapeutic response of cancer patients. Nonetheless, no researchers have reported a systematic analysis of cell–cell interactions in the identification of TME cells associated with cancer. The cell–cell interactions in the TME have been proposed to be related to cancer progression and influence therapeutic efficacy [11]. Here, iATMEcell was developed to identify abnormal TME cells associated with the cancer process by constructing a cell–cell crosstalk network. In the method, the cell crosstalk network is a weighted network, where the weight between cells reflects both their biological function similarity and the transcriptional dysregulated activities of gene signatures shared by them. The RWR algorithm was used to identify the abnormal cells linked to disease status (such as patient survival).

We applied iATMEcell to TCGA BLCA data to evaluate the performance of the method. In BLCA, NK cells were identified as the most significantly dysregulated TME cells. NK cells belong to granulocytic lymphocytes, which are part of the human immune system, and it can rapidly lyse certain tumor cells. Thus, developing its anti-cancer function has been the focus of cancer research in recent years. We constructed a risk score model based on the signature genes of NK cells by using LASSO regression. According to the risk score model, the BLCA patients were obviously divided into high-risk and low-risk groups (Figure 2). The prognostic effect of the risk score model was then validated in an

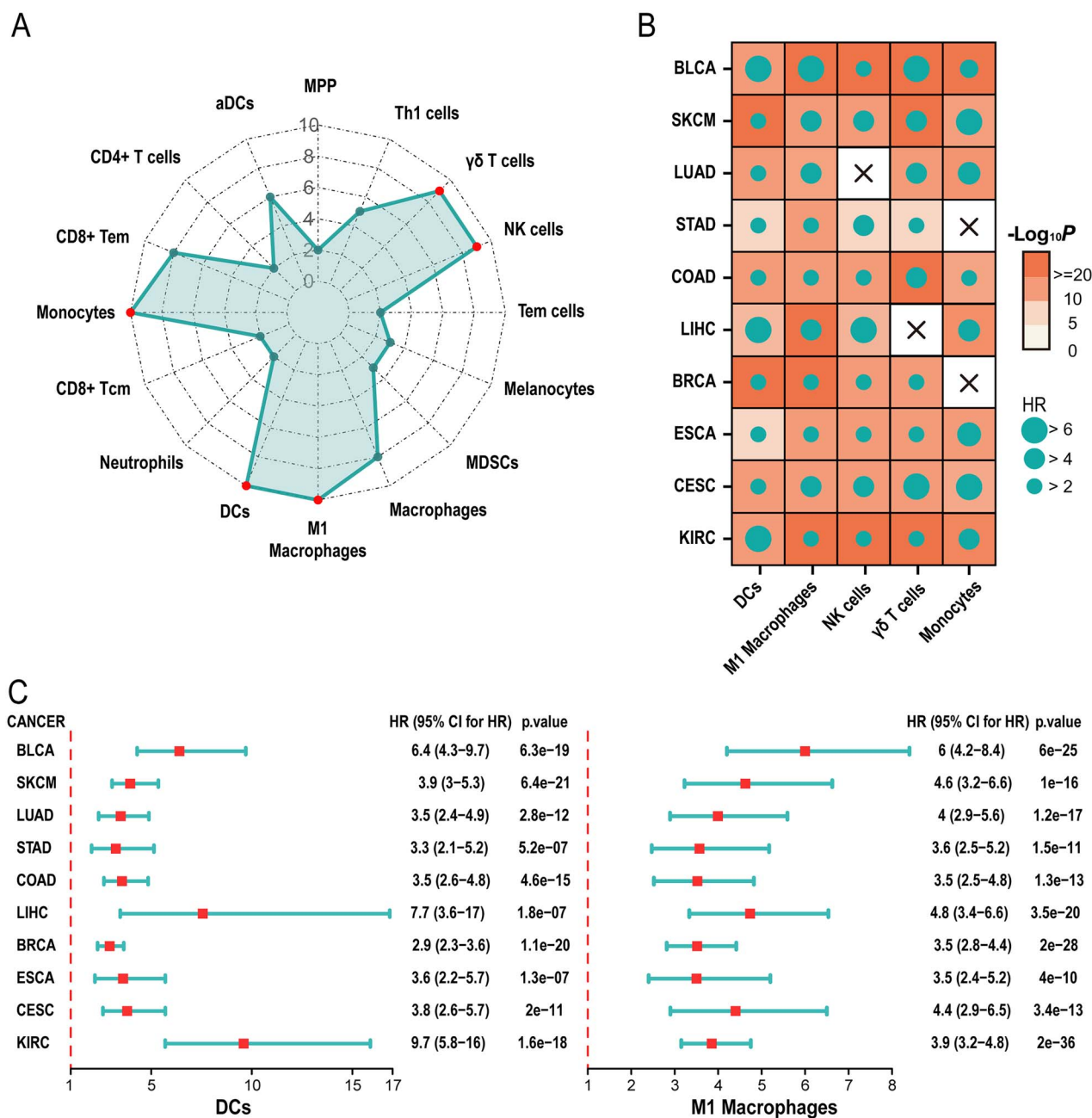


Figure 5. Identification of abnormal TME cells across multi-cancers. (A) The radar plot shows the number of abnormal cells overlapping in 10 cancer types. Red dots indicate the cells were overlapped in at least nine cancer types. (B) Dot plot of univariate HRs and P-values for the five highly overlapped immune cells associated with patient outcomes in their respective cancers. (C) Forest plot showing HR and 95% CI for the risk score models of DCs and M1 Macrophages across 10 cancers. MPP: Multipotent progenitor cell; MDSCs: Myeloid-derived suppressor cell.

independent IMvigor210 cohort, which includes 298 BLCA patients treated with anti-PD-L1. More interestingly, we found that the patients in the low-risk score group had a significantly higher response rate to immunotherapy than that in the high-risk groups (Figure 4E). These results indicate that the risk score model of NK cells may be potentially used as not only a prognostic signature but also a biomarker for immunotherapy response. Moreover, we constructed an integrated risk model based on five significant abnormal TME cells in the training set and validated it in two independent cohorts. The results showed that the integrated model improved predictive value of clinical outcomes in BLCA (Supplementary Figure S3).

We further applied iATMEcell to 10 cancer types that have been proposed to be suitable for immunotherapy. By comparing the results of these cancers, we found that five types of cells were identified in at least nine cancers, namely DCs, M1 Macrophages, NK cells, $\gamma\delta$ T cells and Monocytes (Figure 5A). Current studies confirmed that these immune cells were emerging as protagonists of immunotherapy in the treatment of cancer [42]. For example, $\gamma\delta$ T cells, which represent only 5% of all T cells in our body, but they play an important role in the fight against cancer development. Studies have shown that infiltration of $\gamma\delta$ T cells in tumors is the best predictor of good patient prognosis and exploiting the function of $\gamma\delta$ T cells to develop a new chimeric antigen

receptor (CAR) T-cell therapy has the potential to create a super-armed cell with significantly increased cytotoxicity against cancer [43]. The identification of similar abnormal TME cells across multiple cancers has important research implications for cancer immunotherapy and provides new perspectives on potential anti-cancer immunotherapies for drug repurposing.

The iATMEcell method has been developed to identify abnormal cells in the TME based on the cell–cell crosstalk network. The limitation of the method is that the TME cells and the cell-specific signature genes may be incomplete. As cancer research progresses, more cell information will be incorporated by our method for further analysis. To provide users with easy access to our method, iATMEcell has been implemented as a freely available R package on GitHub (<https://github.com/hanjunwei-lab/iATMEcell>). Overall, the iATMEcell method provides new insight into cancer research and it will become a new tool for identifying potential therapeutic targets for cancer.

Key Points

- Tumor microenvironment (TME) is a dynamic system containing many different non-cancerous cells in addition to cancer cells. Interactions between TME cells shape the unique growth environment, sustaining tumor growth and causing the immune escape of tumor cells. In the study, we developed a novel iATMEcell method to identify abnormal TME cells associated with the clinical outcomes based on a cell–cell crosstalk network.
- In iATMEcell, we constructed a cell–cell crosstalk network based on cell function similarities augmented with measurements of transcriptional dysregulation in the context of bulk tumor dataset, and then it used a network propagation algorithm to identify significantly dysregulated TME cells.
- iATMEcell could effectively identify potentially abnormal TME cells based on a cell–cell crosstalk network, which provided a new insight into understanding the effect of TME cells on prognosis and immunotherapy of cancer patients.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

National Natural Science Foundation of China (grant nos. 62072145), the Natural Science Foundation of Heilongjiang Province (grant no. LH2019C042).

Data availability statement

Availability of data and materials Biological function data was derived from GO biological processes, which were downloaded from C5 GO gene sets in the Molecular Signatures Database (MSigDB) database (<https://www.gsea-msigdb.org/gsea/index.jsp>). The normalized RNA-seq data and clinical information for the 10 cancers used in the study were downloaded from the GDC TCGA data portal (<https://portal.gdc.cancer.gov/>). The BLCA patients treated with anti-PD-L1 were from the IMvigor210 cohort,

which were downloaded from <http://research-pub.gene.com/IMvigor210CoreBiologies/>. The cell type-specific gene signature sets and core code were stored in the iATMEcell package, which is available on GitHub (<https://github.com/hanjunwei-lab/iATMEcell>).

References

1. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. *Nat Med* 2013;**19**:1423–37.
2. Kaymak I, Williams KS, Cantor JR, et al. Immunometabolic interplay in the tumor microenvironment. *Cancer Cell* 2021;**39**:28–37.
3. Gajewski TF, Schreiber H, Fu YX. Innate and adaptive immune cells in the tumor microenvironment. *Nat Immunol* 2013;**14**:1014–22.
4. Nishikawa H, Sakaguchi S. Regulatory T cells in cancer immunotherapy. *Curr Opin Immunol* 2014;**27**:1–7.
5. De Palma M, Lewis CE. Macrophage regulation of tumor responses to anticancer therapies. *Cancer Cell* 2013;**23**:277–86.
6. Becht E, Giraldo NA, Germain C, et al. Immune contexture, immunoscore, and malignant cell molecular subgroups for prognostic and theranostic classifications of cancers. *Adv Immunol* 2016;**130**:95–190.
7. Brune B, Weigert A, Dehne N. Macrophage polarization in the tumor microenvironment. *Redox Biol* 2015;**5**:419.
8. Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* 2015;**12**:453–7.
9. Miao YR, Zhang Q, Lei Q, et al. ImmuCellAI: a unique method for comprehensive T-cell subsets abundance prediction and its application in cancer immunotherapy. *Adv Sci (Weinh)* 2020;**7**:1902880.
10. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;**18**:220.
11. Armingol E, Officer A, Harismendy O, et al. Deciphering cell-cell interactions and communication from gene expression. *Nat Rev Genet* 2020;**22**:71–88.
12. Efremova M, Vento-Tormo M, Teichmann SA, et al. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. *Nat Protoc* 2020;**15**:1484–506.
13. Wang Y, Wang R, Zhang S, et al. iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *bioRxiv* 2019, 507871. <https://doi.org/10.1101/507871>.
14. Bindea G, Mlecnik B, Tosolini M, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 2013;**39**:782–95.
15. Charoentong P, Finotello F, Angelova M, et al. Pan-cancer Immunogenomic analyses reveal genotype-ImmunoPhenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017;**18**:248–62.
16. Danaher P, Warren S, Dennis L, et al. Gene expression markers of tumor infiltrating leukocytes. *J Immunother Cancer* 2017;**5**:18.
17. Davoli T, Uno H, Wooten EC, et al. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* 2017;355.
18. He Y, Jiang Z, Chen C, et al. Classification of triple-negative breast cancers based on Immunogenomic profiling. *J Exp Clin Cancer Res* 2018;**37**:327.
19. Rooney MS, Shukla SA, Wu CJ, et al. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 2015;**160**:48–61.

20. Tirosch I, Izar B, Prakadan SM, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 2016;**352**:189–96.
21. Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol* 2016;**17**:218.
22. Racle J, de Jonge K, Baumgaertner P, et al. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *Elife* 2017;**6**:e26476.
23. Jiang P, Gu S, Pan D, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med* 2018;**24**:1550–8.
24. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. *Gene Ontol Consortium, Nat Genet* 2000;**25**:25–9.
25. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.
26. Mariathasan S, Turley SJ, Nickles D, et al. TGFbeta attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* 2018;**554**:544–8.
27. Rose TL, Weir WH, Mayhew GM, et al. Fibroblast growth factor receptor 3 alterations and response to immune checkpoint inhibition in metastatic urothelial cancer: a real world experience. *Br J Cancer* 2021;**125**:1251–60.
28. Sheng Y, Jiang Y, Yang Y, et al. CNA2Subpathway: identification of dysregulated subpathway driven by copy number alterations in cancer. *Brief Bioinform* 2021;**22**(5):bbaa413.
29. Kohler S, Bauer S, Horn D, et al. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;**57**(1):289–300.
31. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med* 1997;**16**:385–95.
32. Prager I, Watzl C. Mechanisms of natural killer cell-mediated cellular cytotoxicity. *J Leukoc Biol* 2019;**105**:1319–29.
33. Riggan L, Shah S, O'Sullivan TE. Arrested development: suppression of NK cell function in the tumor microenvironment. *Clin Transl Immunol* 2021;**10**:e1238.
34. Sun Y, Sedgwick AJ, Khan MA, et al. A transcriptional signature of IL-2 expanded natural killer cells predicts more favorable prognosis in bladder cancer. *Front Immunol* 2021;**12**:724107.
35. Hartana CA, Ahlen Bergman E, Zirakzadeh AA, et al. Urothelial bladder cancer may suppress perforin expression in CD8+ T cells by an ICAM-1/TGFbeta2 mediated pathway. *PLoS One* 2018;**13**:e0200079.
36. Pan Y, Chiu YH, Chiu SC, et al. Gamma/Delta T-cells enhance carboplatin-induced cytotoxicity towards advanced bladder cancer cells. *Anticancer Res* 2020;**40**:5221–7.
37. Kachrilas S, Dellis A, Papatsoris A, et al. PI3K/AKT pathway genetic alterations and dysregulation of expression in bladder cancer. *J BUON* 2019;**24**:329–37.
38. Liu F, Yang X, Geng M, et al. Targeting ERK, an Achilles' heel of the MAPK pathway, in cancer therapy. *Acta Pharm Sin B* 2018;**8**:552–62.
39. Pishesha N, Harmand TJ, Ploegh HL. A guide to antigen processing and presentation. *Nat Rev Immunol* 2022;**22**:751–64.
40. Latchman Y, Wood CR, Chernova T, et al. PD-L2 is a second ligand for PD-1 and inhibits T cell activation. *Nat Immunol* 2001;**2**:261–8.
41. Diefenbach A, Raulet DH. The innate immune response to tumors and its role in the induction of T-cell immunity. *Immunol Rev* 2002;**188**:9–21.
42. Gerada C, Ryan KM. Autophagy, the innate immune response and cancer. *Mol Oncol* 2020;**14**:1913–29.
43. Silva-Santos B, Mensurado S, Coffelt SB. Gammadelta T cells: pleiotropic immune effectors with therapeutic potential in cancer. *Nat Rev Cancer* 2019;**19**:392–404.