

ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy

Ya-Ru Miao, Qiong Zhang, Qian Lei, Mei Luo, Gui-Yan Xie, Hongxiang Wang, and An-Yuan Guo*

The distribution and abundance of immune cells, particularly T-cell subsets, play pivotal roles in cancer immunology and therapy. T cells have many subsets with specific function and current methods are limited in estimating them, thus, a method for predicting comprehensive T-cell subsets is urgently needed in cancer immunology research. Here, Immune Cell Abundance Identifier (ImmuCellAI), a gene set signature-based method, is introduced for precisely estimating the abundance of 24 immune cell types including 18 T-cell subsets, from gene expression data. Performance evaluation on both the sequencing data with flow cytometry results and public expression data indicate that ImmuCellAI can estimate the abundance of immune cells with superior accuracy to other methods especially on many T-cell subsets. Application of ImmuCellAI to immunotherapy datasets reveals that the abundance of dendritic cells, cytotoxic T, and gamma delta T cells is significantly higher both in comparisons of on-treatment versus pre-treatment and responders versus non-responders. Meanwhile, an ImmuCellAI result-based model is built for predicting the immunotherapy response with high accuracy (area under curve 0.80–0.91). These results demonstrate the powerful and unique function of ImmuCellAI in tumor immune infiltration estimation and immunotherapy response prediction.

1. Introduction


The immune system, comprising various proteins, immune cells, and tissues, is complex and important for host defense.^[1] Immune cells, including innate immune cells [e.g., macrophages, neutrophils, natural killer (NK) cells, and dendritic cells (DC)] and adaptive immune cells (e.g., B and T cells), are important components of the immune system. Dysfunctions of immune cells such as abnormal distributions of abundance and type as well as abnormal development and functions are always associated with diseases, including cancers.^[2,3] Thus, investigating immune cell distribution in individuals could provide important insights into immune status, disease progression, and prognosis, as well as therapy (particularly in cancer immunotherapy).^[4]

Tumor-infiltrating immune cells are considered to be primary immune signatures and are strongly associated with the clinical outcomes of immunotherapies.^[5]

T cells play pivotal roles in cancer initiation, progression, and therapy (particularly immunotherapy)^[6] and are composed of two major groups: CD4⁺ T cells and CD8⁺ T cells with each including numerous functional subpopulations (or subsets). The CD4⁺ T-cell subsets, such as T helper cells (e.g., Th1, Th2, Th17, and Tfh) and regulatory T cells (e.g., nTreg, iTreg, and Tr1), primarily display helper and/or regulatory activities on other immune cells.^[7] The CD8⁺ T-cell subsets, cytotoxic T cells (Tc) and mucosal-associated invariant T cells (MAIT), function in killing target cells. Importantly, the abundance of T-cell subsets, particularly that of tumor-infiltrating T cells, could influence clinical curative effects and prognosis.^[8] In addition, strategies used for adjusting the proportion of T-cell subsets have demonstrated profound efficacy in cancer immunotherapies. For example, increasing the ratio between effector T cell and Treg cell subsets could enhance the antitumor effects of anti-CTLA-4 therapy against melanoma.^[9] Thus, investigating the landscape of immune cells, particularly T cells, can help us better understand the interplay between the immune system and diseases and provide important clues for improving the efficacy of immunotherapy in precision medicine.^[10]

Y.-R. Miao, Dr. Q. Zhang, Dr. Q. Lei, M. Luo, G.-Y. Xie, Prof. A.-Y. Guo
Center for Artificial Intelligence Biology
Hubei Bioinformatics and Molecular Imaging Key Laboratory
Department of Bioinformatics and Systems Biology
Key Laboratory of Molecular Biophysics of the Ministry of Education
College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan 430074, China
E-mail: guoay@hust.edu.cn

Prof. H. Wang
Department of Hematology
Wuhan Central Hospital
Tongji Medical College
Huazhong University of Science and Technology
Wuhan 430074, China

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/advs.201902880>.

© 2020 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.201902880

High-throughput technologies, including microarrays and RNA sequencing (RNA-Seq), produce large-scale transcriptome data and provide opportunities for estimating the abundance of immune cells using gene expression profiles. Several methods, including xCell,^[11] CIBERSORT,^[12] EPIC,^[13] TIMER,^[14] MCP-counter,^[15] and DeconRNASeq^[16] have been developed for enumerating immune cells from bulk transcriptome data of tumor samples, whereas a rare method has been designed for estimating the abundance of numerous T-cell subsets, such as iTreg, Tc, and exhausted T cells (Tex). As such, there is an urgent need to develop a method focusing on abundance prediction of T-cell subsets and other important immune cells in immuno-oncology and immunotherapy studies.

In this study, we developed Immune Cell Abundance Identifier (ImmuCellAI), a method to robustly and precisely estimate the abundance of 24 immune cell types (including 18 T-cell subsets) from transcriptome data. ImmuCellAI was applicable to both microarray and RNA-Seq expression profiles from various resources (e.g., tumor, adjacent or normal tissue, and peripheral blood). Furthermore, we applied ImmuCellAI to cancer immunotherapy and The Cancer Genome Atlas (TCGA) pan-cancer data to explore the influence of immune cells on the efficacy of immunotherapy and clinical progression of patients with cancer.

2. Results

2.1. Algorithmic Overview of the ImmuCellAI Method

ImmuCellAI was designed to estimate the abundance of 18 T-cell subsets [CD4⁺, CD8⁺, CD4⁺ naïve, CD8⁺ naïve, central memory T (Tcm), effector memory T (Tem), Tr1, iTreg, nTreg, Th1, Th2, Th17, Tfh, Tc, MAIT, Tex, gamma delta T ($\gamma\delta$ T), and natural killer T (NKT) cells] and six other important immune cells (B cells, macrophages, monocytes, neutrophils, DC, and NK cells) (Figure 1a). A brief illustration of the core algorithm of ImmuCellAI is represented in Figure 1b, and its detailed algorithm is described in the Experimental Section. Briefly, we curated a specific gene set from publications as gene signature (Table S1, Supporting Information) and obtained the reference expression profile from the Gene Expression Omnibus (GEO) database for each cell type (Table S2, Supporting Information). Then, we calculated the total expression deviation of the gene signature in the input expression profile in comparison with the reference expression profiles of the 24 immune cell types. We assigned the deviation to corresponding immune cell type based on the enrichment score of its gene signature, which was calculated using the single sample gene set enrichment analysis (ssGSEA) algorithm.^[17] To correct the bias due to shared genes in the gene signatures of different immune cell types, a compensation matrix was introduced and least square regression was implemented to measure the weight of shared genes on these immune cells and to re-estimate their abundance (Figure 1b). ImmuCellAI was suitable for application to both RNA-Seq and microarray expression data from blood or tissue samples. To better utilize ImmuCellAI, we designed a user-friendly web server, which is freely available at <http://bioinfo.life.hust.edu.cn/web/ImmuCellAI/>, for estimating the abundance of 24 immune cell types from gene expression profiles.

2.2. Performance of ImmuCellAI in RNA-Seq and Microarray Datasets

To evaluate the performance of ImmuCellAI, we applied it to multiple RNA-Seq and microarray expression datasets, performed benchmark tests, and compared the results with other five methods (xCell,^[11] CIBERSORT,^[12] EPIC,^[13] MCP-counter,^[15] and TIMER^[14]). Pearson correlation between the abundance estimated by flow cytometry and in silico method was used to assess the performance of each method in estimating the abundance of individual immune cell type, whereas the correlation deviation for all cell types was calculated to systematically evaluate the overall prediction power of each method (details are discussed in the Experimental Section).

First, we enumerated the amount of immune cell types available in each of the six analytical methods, among which ImmuCellAI proved capable of predicting more T cell subsets than other methods (Figure 2a). Then, we used six RNA-Seq datasets as benchmark resources for evaluating the performance of ImmuCellAI (Figure 2b,c) on RNA-Seq data. Three of them were simulated and integrated from single-cell sequencing data of liver cancer (GSE98638),^[18] lung cancer (GSE99254),^[19] and melanoma (GSE72056),^[20] their immune cell proportions were calculated from single cell barcode information (Tables S5–S7, Supporting Information). One dataset was taken from the lymph nodes of four patients with melanoma included in the EPIC^[13] project and their flow cytometry result was also obtained. Furthermore, because of the limited number of T-cell subsets in currently available data, to evaluate the performance of ImmuCellAI in estimating the abundance of unique T-cell subsets, we generated two datasets using flow cytometry analysis for all 24 immune cell types (Table S6, Supporting Information) and sequenced their RNA (BIG Data Center ids: CRA001839 and CRA001840). One of these datasets contained five samples from healthy donors and the other contained seven samples from patients with acute myelocytic leukemia. Based on the results, the abundance of most immune cells estimated by ImmuCellAI showed a higher positive correlation with the counting results of flow cytometry than that estimated by the other methods, particularly for T-cell subsets. These results suggested that ImmuCellAI is able to robustly and accurately estimate the abundance of 24 immune cell types in RNA-Seq datasets (Figure 2b,c; Figures S1a,b and S2a, Supporting Information).

Meanwhile, we used two microarray datasets (GSE65135^[12] and GSE65133^[12] from GEO), which were disaggregated lymph node biopsy samples from patients with follicular lymphoma and peripheral blood samples from individuals vaccinated for influenza with immune cell fractions determined by flow cytometry. The abundance of each cell type measured by ImmuCellAI showed overall high positive correlations with the flow cytometry results in both datasets (Figure 2d; Figure S2b, Supporting Information). In addition, ImmuCellAI showed the least correlation deviation in both RNA-Seq and microarray datasets (Figure 2e). The performance evaluation results indicated that ImmuCellAI has the best performance in both microarray and RNA-Seq data with stable and high precision in terms of estimation of abundance of the 24 immune cell types.

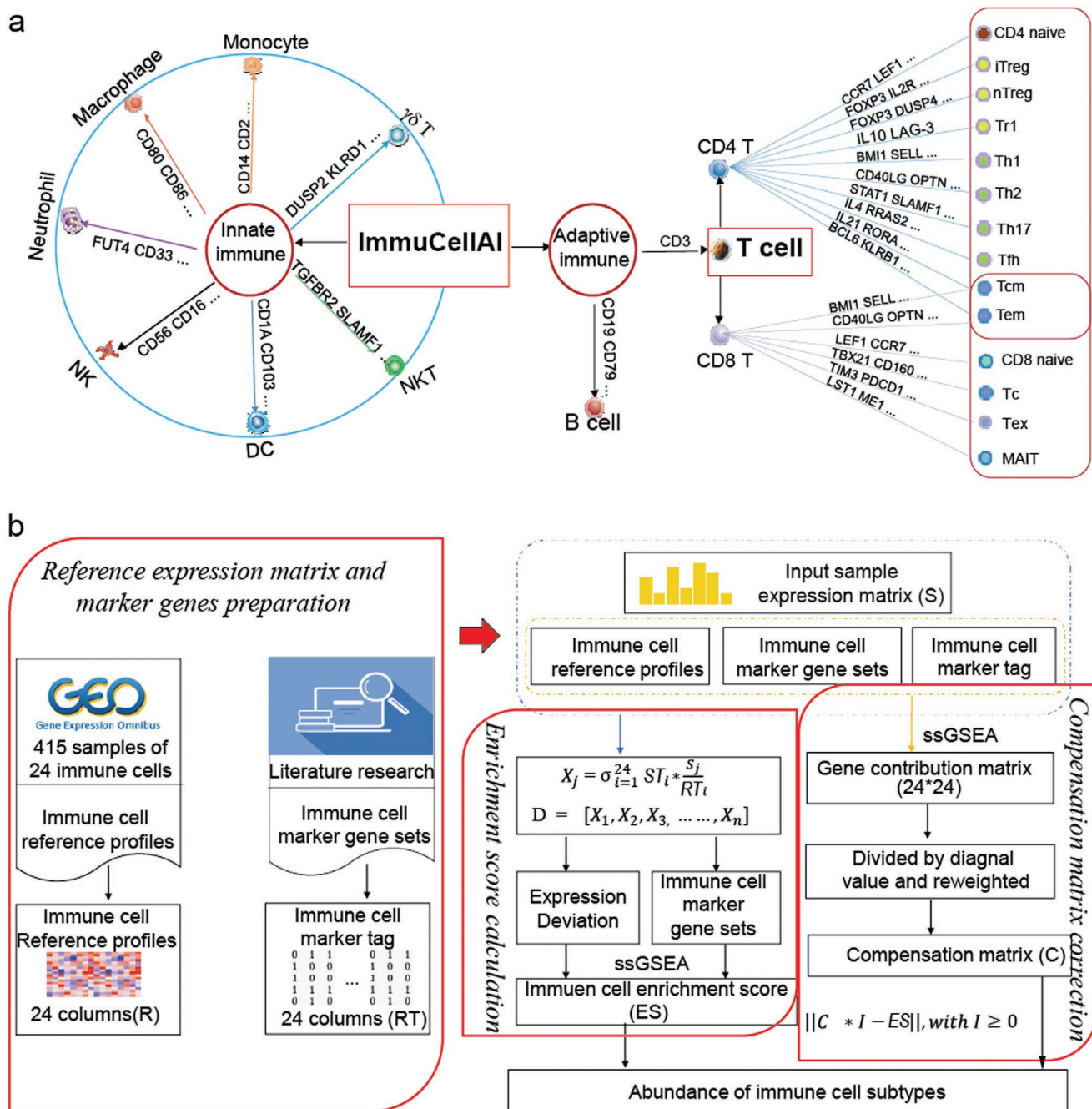


Figure 1. Immune cell types estimated by ImmuCellAI and the workflow of ImmuCellAI. a) Immune cell subsets enumerated by ImmuCellAI. Genes on the line to cell types are the examples of their marker genes. b) The pipeline of the ImmuCellAI algorithm. The three red boxes are the three main steps of ImmuCellAI algorithm. The reference expression profiles of the immune cells were obtained from GEO, and marker genes per immune cell type were obtained from the literature and analytical methods. For each queried sample, the enrichment score of total expression deviation of the signal gene sets was calculated and assigned to each immune cell type by the ssGSEA algorithm. The compensation matrix and least square regression were implemented to correct the bias caused by the shared marker genes among different immune cell types.

2.3. Case Study of ImmuCellAI Application for Cancer Immunotherapy Response Prediction

To investigate the impact of immune cell abundance on cancer immunotherapy, we applied ImmuCellAI to a dataset, GSE91061,^[21] which comprised 58 melanoma samples from

a clinical trial on anti-PD1 therapy. We analyzed the results using two comparisons: responders versus non-responders and on-treatment versus pre-treatment. The abundance of three immune cells including Tc, $\gamma\delta$ T, and DC cells significantly increased with anti-PD1 treatment (Figure 3a; Mann–Whitney U-test, $p < 0.05$). Besides, Tc, $\gamma\delta$ T, and DC cells also significantly

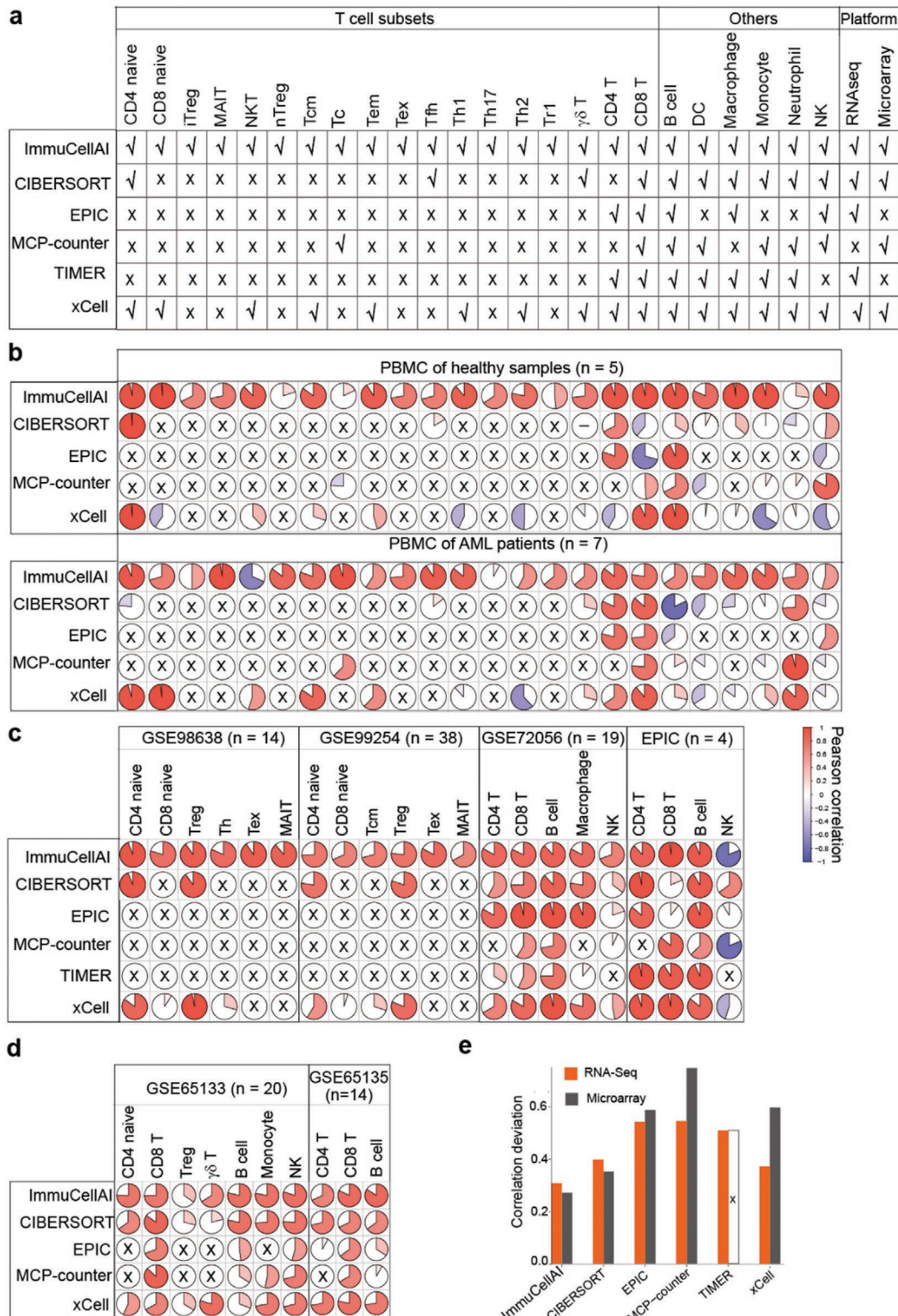


Figure 2. Performance comparison of ImmuCellAI and other methods. a) Immune cell types can be estimated and platforms applicable in ImmuCellAI and other five methods. b) Prediction accuracy of ImmuCellAI and other methods for our sequenced blood samples from healthy donors and AML patients. The rows correspond to methods and the columns indicate the Pearson coefficient for the corresponding cell in the pie graph. Cell types not available in the corresponding methods are marked with a black “x.” The “–” in the circle denotes the correlation analysis result was “NA.” c,d) Performance of ImmuCellAI and other methods on public RNA-Seq datasets (c) and microarray datasets (d). e) Correlation deviation of each method, which took sample size and overall accuracy into consideration to measure the global performance of each tool. “x” means that TIMER was not suitable for estimating the cell fraction of the two microarray datasets (PBMC: GSE65133 and FL: GSE65136).

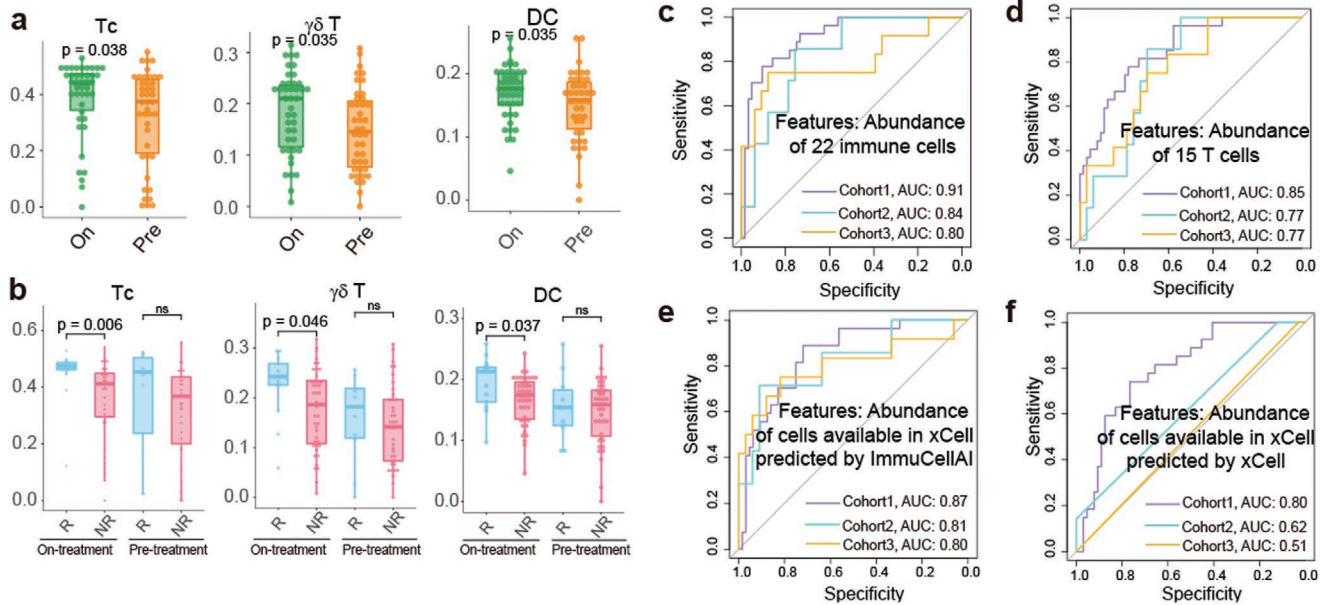


Figure 3. Case study of the application of ImmuCellAI on immunotherapy. a,b) The significant abundance differences of three types of immune cells in before (Pre) and during (On) anti-PD1 treatment (a), and responders (R) and non-responders (NR) at on-treatment (anti-PD1) time point (b). c,d) The receiver operating characteristic (ROC) curve of the immunotherapy response prediction model using ImmuCellAI estimated abundance of 22 immune cells (c) or 15 T-cell subsets (d) as features in the test and validation cohorts. e,f) Performance of prediction model using the abundance of immune cells available in both xCell and ImmuCellAI estimated by ImmuCellAI (e) or xCell (f) as features in the test and validation cohorts. “Cohort 1” contains 53 samples (eight responders and 45 non-responders, random sampling from GSE91061, GSE78220 and GSE115821). “Cohort 2” contains 41 samples (seven responders and 33 non-responders, SRP011540). “Cohort 3” contains 45 samples (12 responders and 33 non-responders, ERP107734).

infiltrated more in responders compared with non-responders at the on-treatment time point (Figure 3b; Mann–Whitney *U*-test, $p < 0.05$). The results suggested that ImmuCellAI can provide important insights on the dynamic immune cell infiltration during immunotherapy and offer valuable indicator for immunotherapy response during the treatment.

Next, we built a prediction model for immune checkpoint blockade (ICB) therapy response based on the immune cell abundance estimated by ImmuCellAI. Five anti-PD1 or anti-CTLA4 therapy datasets (GSE91061,^[21] GSE78220,^[22] and GSE115821,^[23] ERP107734,^[24] and SRP011540^[25]) with a total of 176 patients were involved in the process. The former three datasets from GEO were used for training and testing in a support vector machine model based on the abundance of immune cells, and the last two cohorts from dbGAP were used for further validation (details in methods). By taking the predicted abundance of 22 immune cells (B cell, CD4⁺ naïve, CD8⁺ naïve, Tcm, Tc, DC, $\gamma\delta$ T, Tem, Tex, iTreg, macrophage, MAIT, monocyte, neutrophil, NK, NKT, nTreg, Tfh, Th1, Th17, Th2, and Tr1) as features, the model achieved relatively high accuracy in predicting immunotherapy response in the test data (area under curve (AUC) 0.91) and other two validation cohorts (AUC 0.84 and 0.80, Figure 3c). And T-cell subsets made great contribution in the model by using only the abundance of T-cell subsets as features achieved AUC of 0.77–0.85 (Figure 3d). Furthermore, using the abundance of immune cell subsets available in both ImmuCellAI and xCell as features, the model with abundance predicted by ImmuCellAI performed better than xCell in all three cohorts (Figure 3e,f). Overall, the abundance of immune cells measured by ImmuCellAI was highly predictive

of immune therapy sensitivity (Figure 3c–f), suggesting that ImmuCellAI can serve as an ideal method for immunotherapy studies. We implemented the model for immune therapy response prediction as a functional module on the ImmuCellAI server.

2.4. Case Study of ImmuCellAI Application to TCGA Pan-Cancer Data for Predicting the Infiltration of Immune Cells and Patient Survival

Increasing evidence has demonstrated that immune cells are critical in cancer progression, and the infiltration of different T-cell subsets could dramatically influence the treatment strategy and prognosis.^[26] In this study, to demonstrate the application of ImmuCellAI to cancer research, we analyzed 17 cancer types in TCGA with gene expression data of both the tumor and adjacent tissues to survey the infiltration difference of immune cells. Partial correlation analysis was implemented to reduce false correlation, which may be caused by other factors, such as age and gender (Figure S3, Supporting Information). The results indicated that the abundance of many immune cell types was significantly different (false discovery rate (FDR) < 0.1) between samples from tumor and adjacent tissue in most cancers, particularly for Tc, NK, NKT, Th2, iTreg, nTreg, and DC (Figure 4a). The iTreg, nTreg, Tr1, and monocyte cells were markedly enriched in the nidus of most cancer types, which is consistent with their immunosuppressive properties (Figure 4a). In contrast, several antitumor cells, such as $\gamma\delta$ T, MAIT, NK, NKT, and Th2 cells, showed higher infiltration

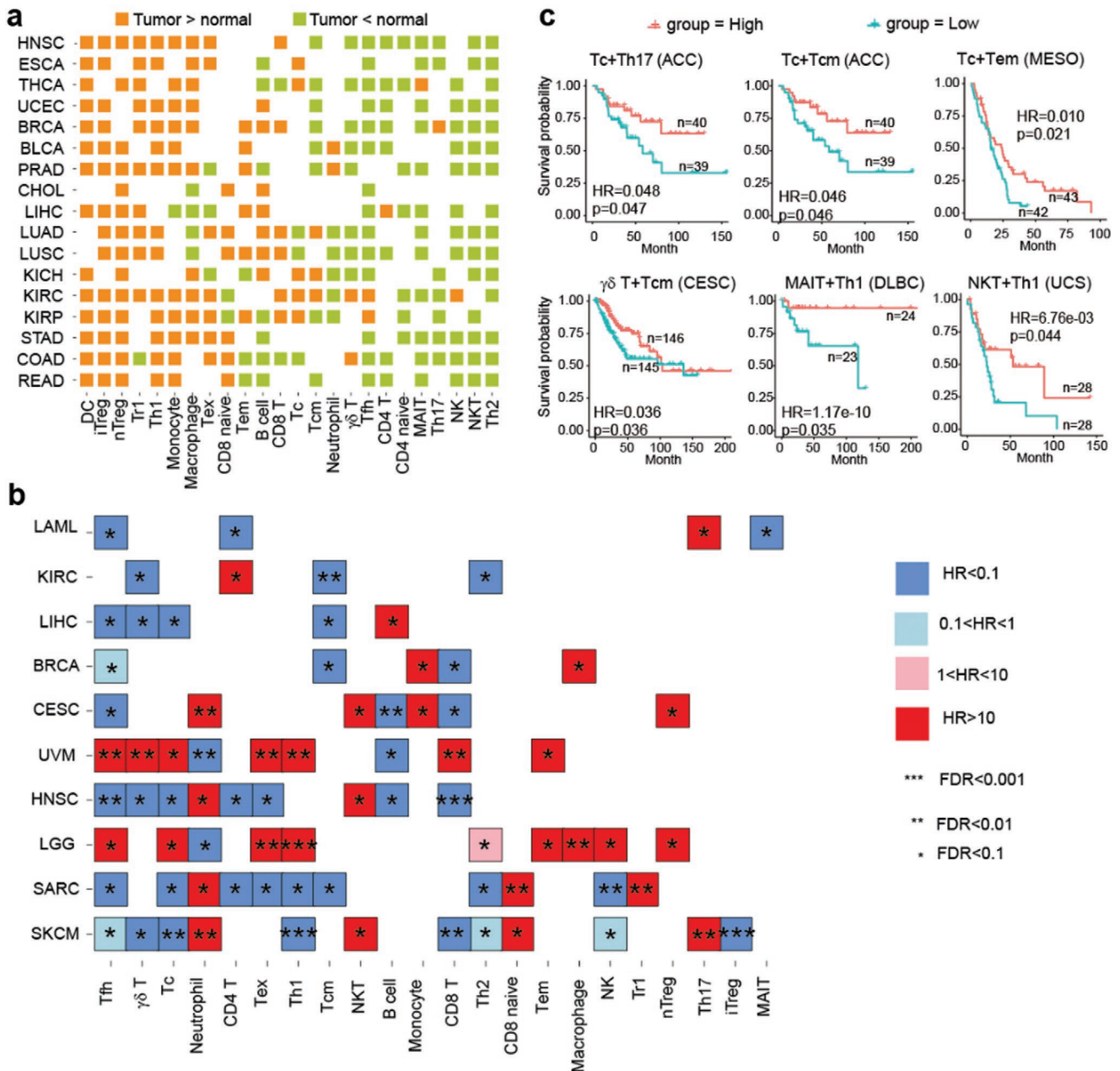


Figure 4. Analysis of the infiltration of immune cells in TCGA data by ImmuCellAI. a) A landscape of the comparison of the infiltration of immune cells between the tumor and adjacent tissues. The orange blocks indicate that cells infiltrated more in the nidus tissue and green blocks indicate the opposite. Statistical significance was evaluated using Wilcoxon's rank sum test with an FDR of 0.10. b) Association of tumor-infiltrating immune cells with patient survival. For each cancer type, multivariate Cox regression was performed, with covariates including the abundance of immune cell, patient age at diagnosis, gender, and clinical stage. c) Kaplan–Meier curves of cancers by the combination of multiple immune cell types. Statistical significance and hazard ratios were calculated using multivariate Cox regression.

in adjacent tissues of most cancers, indicating that the tumor microenvironment may prohibit their access to the nidus.

Furthermore, we investigated the effects of the immune cell infiltration on patient survival by controlling other factors (i.e., age, gender, and stage) using cox regression analysis. In a nutshell, the infiltration of most immune cell types significantly affected the overall survival of patients in different cancers (Figure 4b). The infiltration of most immune cells had opposite effects on survival in brain lower grade glioma and uveal

melanoma compared with that in other cancers (Figure 4b). Furthermore, skin cutaneous melanoma (SKCM) had the most immune cell types (12/24) significantly associated with patient survival (FDR < 0.1; Figure 4b; Figure S4, Supporting Information). The infiltration of T-cell subsets (e.g., $\gamma\delta$ T cells, Th1, Th2, and iTreg) had positive effects on long-term survival in patients with SKCM, whereas patients with high infiltration of CD8⁺ naïve and neutrophils were associated with worse outcomes (Figure 4b; Figure S4, Supporting Information). Although the

infiltration of a single immune cell type in some cancer types was not related with patient survival, the infiltration of a combination of multiple immune cell types was significantly associated with survival (multivariate Cox regression, $p < 0.05$), such as Tc + Th17 in adrenocortical carcinoma (ACC), Tc + Tcm in ACC, Tc + Tem in mesothelioma, $\gamma\delta$ T + Tcm in cervical and endocervical cancers, MAIT + Th1 in diffuse large B-cell lymphoma, and NKT + Th1 in uterine carcinosarcoma (Figure 4c). In addition, the infiltration of immune cells was correlated with microsatellite instability (MSI) after partial correlation analysis was performed to reduce false correlation caused by other features, such as age and gender (Figure S5a,b, Supporting Information). In colon adenocarcinoma (COAD) and stomach adenocarcinoma (STAD), patients with high-MSI cancer (MSI-H) showed a significantly higher infiltration of antitumor and tumor helper cells, such as Tc, $\gamma\delta$ T, NK, and DC (Figure S6a, Supporting Information), but a significantly lower infiltration of tumor suppressor cells (Tr1 and neutrophils) and CD8 naïve cells (Figure S6b, Supporting Information). These results may partially explain the better outcomes of patients with MSI-H colorectal cancer undergoing immunotherapy.^[27] Furthermore, we observed that some immune cell types showed stage-related profiles in cancers. For example, the infiltration of Tex, Th1, iTreg, and CD8⁺ T cells gradually increased with the development of kidney renal clear cell carcinoma (Figure S7, Supporting Information).

3. Discussions

Increasing evidence suggests that immune cells play critical roles in carcinogenesis and progression, and a proper proportion of T-cell subsets could contribute to long-term clinical benefits of anticancer treatments.^[28] Investigating the abundance of immune cell types could help to have a more comprehensive understanding of the immune status of patients and could thus benefit disease therapy.^[29] In this study, we developed ImmuCellAI, a highly accurate method of estimating the abundance of immune cells, particularly T-cell subsets, from transcriptome data. The case study application results on immunotherapy and pan-cancer data suggest that ImmuCellAI is a very useful tool in cancer immunology.

ImmuCellAI focuses on immune cell prediction from expression profiles of array or RNA-seq data and is able to enumerate the abundance of 18 T-cell subsets (Figure 1a). Although there are several methods available for immune cell abundance estimation, most of them have intrinsic limitations. The 4 immune cell focused tools (CIBERSORT, EPIC, TIMER, and MCP-counter) can only predict two to four subsets of T cells and only CIBERSORT can be used on both arrays and RNA-Seq data (Figure 1a). xCell was designed for immune and stroma cells including 10 T-cell subsets, while it cannot retrieve the abundance of some important T-cell subsets (iTreg, Tc, and exhausted T cells etc., which were proved playing vital roles in the immune system and immunotherapy^[30–32]) and does not show robust accuracy in some cases (e.g., Figure 2b). Compared with those alive tools, the unique function of ImmuCellAI is that it can accurately estimate the abundance of comprehensive T-cell subsets, which is particularly important in

cancer therapy. For those immune cells that could be identified by other methods, our comparison results showed that ImmuCellAI had the highest consistency with flow cytometry results for most cells (Figure 2b–e). Because there are very limited T-cell subsets with both flow cytometry data and RNA-Seq data, we produced two datasets using flow cytometry analysis for all 24 immune cell types and sequenced their RNA. The results confirmed that ImmuCellAI has the best performance in terms of accurately identifying these T-cell subsets, which is an advantage of this method.

Tumor-infiltrating T cells could serve as a prognostic factor and predictor of therapeutic efficacy.^[33] In our result, the abundance of Tc, $\gamma\delta$ T, and DC cells were significantly increased in both comparisons of on-treatment versus pre-treatment and responders versus non-responders (Figure 3a,b). DC cells are antigen-presenting cells (APC) that are essential for the activation of immune responses, which have the potential to turn immunologically “cold” tumors into “hot” tumors.^[34] Tc cells play key roles in the tumor cell killing process,^[35] and $\gamma\delta$ T cells also function in the antigen recognition and tumor killing process.^[36] To extend these results, taking the abundance of 22 immune cells of pre-treatment samples into account, an immunotherapy model was proposed and reached high accuracy (AUC 0.80–0.91) (Figure 3c). The model using abundance of 22 immune cells predicted by ImmuCellAI was implemented for immunotherapy response prediction on the ImmuCellAI server. This is the second user-friendly web server for immunotherapy response prediction, except for the TIDE.^[37] Comparing with the reported accuracy of TIDE, our ImmuCellAI has a little bit higher accuracy. In addition, to date, most studies have focused on the infiltration of CD8⁺ T cells as a predictive biomarker for response to ICB therapy.^[38] Our study indicated that integrating many immune cells, particularly different T-cell subsets, could serve as a biomarker for better therapy response prediction (Figure 3c). Thus, the systematic evaluation of immune cell abundance could be an effective approach for predicting immunotherapy response and improving the effects of cancer immunotherapy.^[39] Some of the results of the infiltration of immune cells in TCGA cancer data were consistent with those of previous reports, for example, the infiltration of Tc cells is more often observed in kidney cancer but less so in colorectal cancer.^[40] These results indicate the powerful and unique function of ImmuCellAI on cancer immunology and immunotherapy research.

Although ImmuCellAI had the best performance in comparison with other methods, it still has several limitations that need to be addressed. First, ImmuCellAI could only estimate the relative abundance of immune cells based on the deviation of gene signatures. It could not provide the absolute amount of each immune cell type. Moreover, ImmuCellAI did not consider the spatiotemporal localization of immune cells and the abundance of cancer cells. In addition, the sample size used in the immunotherapy case study was relatively small, and the performance of our model needs to be tested in larger cohorts. Other immune cell subsets, besides the T cells used in our method, also need to be tested in future studies.

In summary, this study presented an accurate and reliable tool ImmuCellAI to dissect T-cell properties and explore the infiltration of immune cells in cancer. The best advantage of ImmuCellAI is its ability to accurately estimate the abundance of 18

T-cell subsets, which is its unique function. Besides, ImmuCellAI can be applied to predict the patient response of ICB therapy. The results of ImmuCellAI provided valuable prognostic predictors and comprehensive resources to elucidate cancer-immune interactions, which could facilitate applications of cancer immunotherapy and precision medicine.

4. Experimental Section

The Main Algorithm of ImmuCellAI: The main algorithm of ImmuCellAI, presented in Figure 1b, includes three main steps: 1) reference expression matrix (RT) and marker gene preparation, 2) enrichment score calculation, and 3) compensation matrix correction.

Reference Expression Matrix and Marker Gene Preparation: The datasets of the expression profiles of 24 immune cell types (Figure 1a) were downloaded from the GEO database. In total, 415 datasets from 26 studies were manually curated to build RT of the immune cell types (Table S2, Supporting Information). Gene expression data were obtained from CEL files according to the frozen robust multiarray analysis protocol with batch effect correction.^[41] Each line of the matrix denotes the expression of a gene in the 24 immune cell types. The median value was used if there were multiple samples of a cell type.

Furthermore, a gene signature per cell type was developed by integrating the marker genes obtained from the literature and other analytical methods, such as CIBERSORT and xCell; thus, a total of 2547 genes were collected (denoted as G_a , Table S1, Supporting Information). Next, a robust marker gene set per immune cell type was selected using in silico simulated data by taking advantage of the TCGA data, which was based on the work of Li et al.^[14] For each cancer type in the TCGA data, the expression of G_a in the samples (\log_2 transferred) and immune cell reference profiles were used to simulate the immune cell infiltrated tumor samples with known fractions. To control the mixing ratios of immune cell components for maintaining the correlative structure of real data, the gene-gene covariance matrix Σ_a was first calculated for all genes in G_a using tumor expression data. Then, 24 numbers (f_1 – f_{24}) were randomly sampled from Uniform (0,1) and μ_a (length n) was calculated, which was the average of gene expression in the reference profiles of the 24 immune cell types weighted by f_1 – f_{24} . Next, a vector of length n was sampled from the multivariate normal distribution with mean μ_a and covariance Σ_a . For each cancer type, the same number of samples was simulated as its sample size in the TCGA data.

Then, for all collected marker genes in G_a , the average correlation between gene expression in simulated samples with cell fractions was calculated using Pearson correlation for all cancers, and genes with an average correlation of $r \geq 0.6$ were selected (denoted as G_1). Next, for each marker gene per immune cell, the standard correlation deviation among the cell with other cells was calculated, and genes with standard deviation larger than 1.5 were selected (denoted as G_2). The deviation between $CD4^+$ T and $CD4^+$ T-cell subsets (such as $CD4^+$ naïve and Th1) as well as that between $CD8^+$ T and $CD8^+$ T-cell subsets was not calculated. Finally, a robust marker gene set per immune cell type was obtained by intersecting G_1 with G_2 (denoted as G_f), which included 344 marker genes of the 24 immune cell types (Table S1, Supporting Information). In addition, a sparse matrix (ST) was constructed for these marker genes in which “1” means that the gene is a marker gene in the corresponding cell type.

Enrichment Score Calculation: For a user-uploaded expression dataset, ImmuCellAI first calculates the expression deviation of all marker genes compared with RT. Here two different approaches were implemented to deal with the microarray and RNA-Seq datasets.

$$D = [X_1, X_2, X_3, \dots, X_n], n = 344 \quad (1)$$

$$X = \sum_{i=1}^{24} ST_i * \frac{S}{RT_i} \quad (2)$$

$$S = \begin{cases} m, \text{Microarray} \\ \log_2(m+1), \text{RNA-Seq} \end{cases} \quad (3)$$

where vector D denotes the relative deviation of marker genes and ST_i is the vector in ST for marker genes of cell type i . RT_i is the reference marker gene expression in cell type i , whereas S indicates the gene expression in the user-provided dataset.

The ssGSEA algorithm in the GSEA package^[42] was used to estimate the abundance of immune cell types. The ssGSEA enrichment score for deviation vector D of the gene signature of each immune cell type (named ES) was used to indicate the relative abundance of immune cell types in the user-provided dataset. A higher enrichment score indicates a higher abundance of the immune cell type in the mixture sample than that of other cell types.

Compensation Matrix Correction: Some immune cell types may share a part of common marker genes, which will cause bias in the estimation of abundance of these immune cell types. Thus, ImmuCellAI used a compensation matrix and least square regression method based on the work of Aran et al.^[11] to fix this issue. After the estimation of abundance of the detected immune cell types in a dataset, the weights of common marker genes for these immune cell types were reassigned with the following steps: 1) A $N * N$ contribution matrix was produced by calculating the mutual contributions of marker genes in RT using ssGSEA. 2) Each column of the contribution matrix was divided by a diagonal value and weighted by the proportion of non-diagonal elements, and a compensation matrix was obtained (named C). 3) To reduce redundancy and overestimation of compensation between detected immune cell types, ImmuCellAI discarded the compensational calibration between the parental immune cell type and its subsets (e.g., $CD4^+$ T and Th1 cells), and limited the total compensation level at 0.5 for the non-diagonal cell types. 4) The least square method was used to calibrate the enrichment score based on the compensation matrix C .

$$\|C * I - ES\|, \text{with } I \geq 0 \quad (4)$$

where the parameter ES is the ssGSEA enrichment score of detected immune cell types and C is the compensation matrix. Finally, after calibration, we deemed the abundance of 24 immune cell types (named I) to be high confidence.

Benchmark Dataset Preparation: Our Datasets for the 24 Immune Cell Types: Heparinized blood samples from seven patients with leukemia and five healthy adult volunteers (Table S3, Supporting Information) were collected from Wuhan Central Hospital, China. Fresh blood samples were treated with Pharm Lyse (BD Biosciences, San Jose, CA, USA) to remove erythrocytes. Cells from each sample were used in parallel experiments of flow cytometry and RNA extraction. This study was approved by the ethics committee of Tongji Medical College, Huazhong University of Science and Technology, and followed the Declaration of Helsinki principles. The informed consent was obtained from all volunteers. The proportions of the 24 immune cell types used in the study were examined by flow cytometry using the combined markers listed in Table S3, Supporting Information, and antibodies listed in Table S4, Supporting Information. All antibodies were purchased from BD Biosciences, except those used against TCR-V β 2 and TCR-V α 7.2 (Miltenyi Biotec, Bergisch Gladbach, Germany).

The total RNA extracted from the cells of all 12 samples was used for RNA sequencing (RNA-Seq) (PE150) via the Illumina HiSeqTM4000 platform by Haplox (Jiangxi, China). RNA-Seq reads were mapped to Ensembl v81 (GRCH38) and processed using the HISAT2-StringTie-ballgown pipeline. Fragments per kilobase per million mapped reads were used to calculate gene expression levels.

Other Public Datasets: The microarray datasets and corresponding flow cytometry results were obtained from GEO (accession nos. GSE65135 and GSE65133), which included 14 disaggregated lymph node biopsy samples from patients with follicular lymphoma and 20 peripheral blood samples from individuals vaccinated for influenza, respectively.

Besides, RNA-Seq expression profile of samples from melanoma patients and their corresponding flow cytometry of four immune cell types (B, CD4⁺ T, CD8⁺ T, and NK) result were collected from EPIC publication.^[13] Because of the scarcity of bulk RNA-Seq datasets containing both gene expression profiles and flow cytometry counts for different immune cell types, particularly for T-cell subsets, we simulated two bulk RNA-Seq datasets by integrating the expression profiles of seven cell types (CD4⁺ naïve, CD8⁺ naïve, MAIT, Tcm, Tex, Treg, Th) from single-cell RNA sequencing data. The transcripts per million (TPM) normalized expression of liver and lung cancers from two Nature papers were collected from GEO (GSE98638^[18] and GSE99254^[19]). Based on the work of Max et al.,^[43] single-cell expression was normalized as follows:

$$exp = \log_2 (exp + 1) \quad (5)$$

for each single-cell dataset, the TPM values were transformed to

To ensure cross-sample comparability, the expression of all single-cell samples from the same dataset were normalized to the average expression of 3686 housekeeping genes^[44] as follows:

$$exp'_i = exp_i * \frac{\overline{HK}}{HK_i} \quad (6)$$

where exp_i represents the gene expression profile of sample i , HK_i denotes the average gene expression of all housekeeping genes in sample i , and \overline{HK} is the average expression of all housekeeping genes in all samples. Besides, a single-cell sequencing dataset from 19 patients with melanoma was collected from GEO (accession GSE72056), which is the normalized expression matrix as described above by Tirosh et al.^[20] and contains the single-cell RNA-Seq of B cells, T cells, macrophages, NK cells, and three other nonimmune cell types. Because CD8⁺ and CD4⁺ T cells can be easily distinguished by CD4, CD8A, and CD8B expression, we divided T cells into CD8⁺ T cells, CD4⁺ T cells, and others. Then, the bulk expression of each sample was identified by aggregating normalized expression from all cell barcodes for each patient sample. The cell ratio per cell type in a sample was calculated by the cell number of a specific cell type divided by the total number of cells (Tables S5–S7, Supporting Information).

Performance Assessment of ImmuCellAI: The performance of ImmuCellAI was evaluated using both microarray and RNA-Seq datasets and compared with that of five other methods (CIBERSORT, EPIC, MCP-counter, TIMER, and xCell). For a given immune cell type, the accuracy and sensitivity of each method were measured using the Pearson correlation between the results of in silico method and flow cytometry counting in samples (named r_i). In addition, we introduced the correlation deviation to measure the global performance of each method, which took the sample size and overall accuracy into consideration.

$$\text{Correlation deviation} = \sqrt{\frac{1}{n} \sum_{i=1}^n (1 - r_i)^2} \quad (7)$$

where n is the amount of immune cell types detected in samples and r_i is the Pearson correlation of immune cell type i .

Case Study of Immune Therapy and Prediction Model Building: Five immune checkpoint therapy datasets, including those from anti-PD1- or anti-CTLA4-treated patients with melanoma or gastric cancer, were collected from the GEO database (GSE91061,^[21] GSE78220,^[22] and GSE115821^[23]) and dbGAP (ERP107734^[24] and SRP011540^[25]). The abundance of infiltrating immune cells was calculated by ImmuCellAI and used to build the response prediction model.

The immunotherapy response prediction model was built using support vector machine with the radial basis function kernel. The training features were the abundance of immune cell types. The sequential backward feature selection algorithm was used to minimize the feature number and improve the performance. At first, three GEO datasets composed of 91 pre-treatment samples (response: complete response and partial response, $n = 27$, non-response: stable disease and progressive disease, $n = 64$) were used to train and test the model. The undersampling

method was used to fit the unbalanced sample size between responders and non-responders with 38 samples in the training and validation cohort (19 responders and 19 non-responders, fivefold cross validation) and 53 samples in the test cohort (eight responders and 45 non-responders). Then, the other two cohorts from dbGAP (ERP107734: 12 responders and 33 non-responders; SRP011540: seven responders and 33 non-responders) were used to further validate the model. The area under curve (AUC) was used to measure the model performance.

The gene expression profiles of TCGA samples and the clinical information were downloaded from Broad GDAC Firehose (<https://gdac.broadinstitute.org/>).

Statistical Analysis: Basic statistical analyses, such as Wilcoxon rank sum test and Pearson correlation, were performed using R language. The correlations between clinical indicators and the abundance of immune cell types were evaluated using partial correlation analysis in the R package “ppcor.” Multivariate Cox regression, log-rank test, and Kaplan–Meier in R package “survival” were used to assess the relationships between the abundance of immune cell types and survival time. The p values for each test were calibrated using FDR, and the FDR threshold was 0.1 in case studies. All results supported the current study and were deposited into the ImmuCellAI website (<http://bioinfo.life.hust.edu.cn/web/ImmuCellAI/>).

Data and Materials Availability: The sequence data sets reported in this paper have been deposited in the National Genomics Data Center with Accession Nos. CRA001839 and CRA001840.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Y.-M., Q.Z., and Q.L. contributed equally to this work. The authors express their gratitude to all study participants. The authors acknowledge funding from the National Natural Science Foundation of China (Grant Nos. 31822030, 31801113, and 31771458), National Key Research and Development Program of China (Grant No. 2017YFA0700403), China Postdoctoral Science Foundation (Grant Nos. 2018M632830, 2017M622455, and 2019T120664)

Conflict of Interest

The authors declare no conflict of interest.

Authors Contributions

Y.R.M. and Q.Z. performed formal analysis, conceptualized, conceived method, and wrote the original draft and edited the manuscript; Q.L. collected samples and performed the experiments; M.L. and G.Y.X. performed formal analysis; H.X.W. provided sample and experiment assistance; A.Y.G. conceptualized, wrote-reviewed and edited the manuscript, and funded and supervised the study

Keywords

cancer, immune cells, immunotherapy, T-cell subsets

Received: October 15, 2019

Revised: January 14, 2020

Published online: February 11, 2020

- [1] D. D. Chaplin, *J. Allergy Clin. Immunol.* **2010**, *125*, S3.
- [2] P. Brodin, M. M. Davis, *Nat. Rev. Immunol.* **2017**, *17*, 21.
- [3] G. P. Dunn, C. M. Koebel, R. D. Schreiber, *Nat. Rev. Immunol.* **2006**, *6*, 836.
- [4] T. F. Gajewski, H. Schreiber, Y.-X. Fu, *Nat. Immunol.* **2013**, *14*, 1014.
- [5] P. H. Pandya, M. E. Murray, K. E. Pollok, J. L. Renbarger, *J. Immunol. Res.* **2016**, *2016*, 1.
- [6] J. N. Kochenderfer, M. E. Dudley, S. H. Kassim, R. P. T. Somerville, R. O. Carpenter, M. Stetler-Stevenson, J. C. Yang, G. Q. Phan, M. S. Hughes, R. M. Sherry, M. Raffeld, S. Feldman, L. Lu, Y. F. Li, L. T. Ngo, A. Goy, T. Feldman, D. E. Spaner, M. L. Wang, C. C. Chen, S. M. Kranick, A. Nath, D.-A. N. Nathan, K. E. Morton, M. A. Toomey, S. A. Rosenberg, *J. Clin. Oncol.* **2015**, *33*, 540.
- [7] R. V. Luckheeram, R. Zhou, A. D. Verma, B. Xia, *Clin. Dev. Immunol.* **2012**, *2012*, 1.
- [8] A. O. Kamphorst, R. N. Pillai, S. Yang, T. H. Nasti, R. S. Akondy, A. Wieland, G. L. Sica, K. Yu, L. Koenig, N. T. Patel, M. Behera, H. Wu, M. McCausland, Z. Chen, C. Zhang, F. R. Khuri, T. K. Owonikoko, R. Ahmed, S. S. Ramalingam, *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 4993.
- [9] T. R. Simpson, F. Li, W. Montalvo-Ortiz, M. A. Sepulveda, K. Bergerhoff, F. Arce, C. Roddie, J. Y. Henry, H. Yagita, J. D. Wolchok, K. S. Peggs, J. V. Ravetch, J. P. Allison, S. A. Quezada, *J. Exp. Med.* **2013**, *210*, 1695.
- [10] D. Mittal, M. M. Gubin, R. D. Schreiber, M. J. Smyth, *Curr. Opin. Immunol.* **2014**, *27*, 16.
- [11] D. Aran, Z. Hu, A. J. Butte, *Genome Biol.* **2017**, *18*, 220.
- [12] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, *Nat. Methods* **2015**, *12*, 453.
- [13] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, D. Gfeller, *eLife* **2017**, *6*, e26476.
- [14] T. Li, J. Fan, B. Wang, N. Traugh, Q. Chen, J. S. Liu, B. Li, X. S. Liu, *Cancer Res.* **2017**, *77*, e108.
- [15] E. Becht, N. A. Giraldo, L. Lacroix, B. Buttard, N. Elarouci, F. Petitprez, J. Selves, P. Laurent-Puig, C. Sautès-Fridman, W. H. Fridman, A. de Reyniès, *Genome Biol.* **2016**, *17*, 218.
- [16] T. Gong, J. D. Szustakowski, *Bioinformatics* **2013**, *29*, 1083.
- [17] D. A. Barbie, P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl, S. Fröhling, E. M. Chan, M. L. Sos, K. Michel, C. Mermel, S. J. Silver, B. A. Weir, J. H. Reiling, Q. Sheng, P. B. Gupta, R. C. Wadlow, H. Le, S. Hoersch, B. S. Wittner, S. Ramaswamy, D. M. Livingston, D. M. Sabatini, M. Meyerson, R. K. Thomas, E. S. Lander et al., *Nature* **2009**, *462*, 108.
- [18] C. Zheng, L. Zheng, J.-K. Yoo, H. Guo, Y. Zhang, X. Guo, B. Kang, R. Hu, J. Y. Huang, Q. Zhang, Z. Liu, M. Dong, X. Hu, W. Ouyang, J. Peng, Z. Zhang, *Cell* **2017**, *169*, 1342.
- [19] X. Guo, Y. Zhang, L. Zheng, C. Zheng, J. Song, Q. Zhang, B. Kang, Z. Liu, L. Jin, R. Xing, R. Gao, L. Zhang, M. Dong, X. Hu, X. Ren, D. Kirchhoff, H. G. Roeder, T. Yan, Z. Zhang, *Nat. Med.* **2018**, *24*, 978.
- [20] I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. K. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A.-C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty et al., *Science* **2016**, *352*, 189.
- [21] N. Riaz, J. J. Havel, V. Makarov, A. Desrichard, W. J. Urba, J. S. Sims, F. S. Hodi, S. Martín-Algarra, R. Mandal, W. H. Sharfman, S. Bhatia, W.-J. Hwu, T. F. Gajewski, C. L. Slingluff, D. Chowell, S. M. Kendall, H. Chang, R. Shah, F. Kuo, L. G. T. Morris, J.-W. Sidhom, J. P. Schneck, C. E. Horak, N. Weinhold, T. A. Chan, *Cell* **2017**, *171*, 934.
- [22] W. Hugo, J. M. Zaretsky, L. Sun, C. Song, B. H. Moreno, S. Hu-Lieskovan, B. Berent-Maoz, J. Pang, B. Chmielowski, G. Cherry, E. Seja, S. Lomeli, X. Kong, M. C. Kelley, J. A. Sosman, D. B. Johnson, A. Ribas, R. S. Lo, *Cell* **2017**, *168*, 542.
- [23] N. Auslander, G. Zhang, J. S. Lee, D. T. Frederick, B. Miao, T. Moll, T. Tian, Z. Wei, S. Madan, R. J. Sullivan, G. Boland, K. Flaherty, M. Herlyn, E. Ruppin, *Nat. Med.* **2018**, *24*, 1545.
- [24] S. T. Kim, R. Cristescu, A. J. Bass, K.-M. Kim, J. I. Odegaard, K. Kim, X. Q. Liu, X. Sher, H. Jung, M. Lee, S. Lee, S. H. Park, J. O. Park, Y. S. Park, H. Y. Lim, H. Lee, M. Choi, A. Talasz, P. S. Kang, J. Cheng, A. Loboda, J. Lee, W. K. Kang, *Nat. Med.* **2018**, *24*, 1449.
- [25] D. Miao, C. A. Margolis, N. I. Vokes, D. Liu, A. Taylor-Weiner, S. M. Wankowicz, D. Adeegbe, D. Keliher, B. Schilling, A. Tracy, M. Manos, N. G. Chau, G. J. Hanna, P. Polak, S. J. Rodig, S. Signoretti, L. M. Sholl, J. A. Engelman, G. Getz, P. A. Jänne, R. I. Haddad, T. K. Choueiri, D. A. Barbie, R. Haq, M. M. Awad, D. Schadendorf, F. S. Hodi, J. Bellmunt, K.-K. Wong, P. Hammerman et al., *Nat. Genet.* **2018**, *50*, 1271.
- [26] F. Fend, L. Quintanilla-Martínez, *Haematologica* **2014**, *99*, 599.
- [27] A. Kalyan, S. Kircher, H. Shah, M. Mulcahy, A. Benson, *J. Gastrointest. Oncol.* **2018**, *9*, 160.
- [28] M. B. Schaaf, A. D. Garg, P. Agostinis, *Cell Death Dis.* **2018**, *9*, 115.
- [29] S. Delhalle, S. F. N. Bode, R. Balling, M. Ollert, F. Q. He, *npj Syst. Biol. Appl.* **2018**, *4*, 9.
- [30] D. O. Adeegbe, H. Nishikawa, *Front. Immunol.* **2013**, *4*, 190.
- [31] D. S. Thommen, T. N. Schumacher, *Cancer Cell* **2018**, *33*, 547.
- [32] A.-P. Ganesan, J. Clarke, O. Wood, E. M. Garrido-Martin, S. J. Chee, T. Mellows, D. Samaniego-Castruita, D. Singh, G. Seumois, A. Alzetani, E. Woo, P. S. Friedmann, E. V. King, G. J. Thomas, T. Sanchez-Elsner, P. Vijayanand, C. H. Ottensmeier, *Nat. Immunol.* **2017**, *18*, 940.
- [33] D. M. Pardoll, *Nat. Rev. Cancer* **2012**, *12*, 252.
- [34] M. W. L. Teng, S. F. Ngiow, A. Ribas, M. J. Smyth, *Cancer Res.* **2015**, *75*, 2139.
- [35] L. Martínez-Lostao, A. Anel, J. Pardo, *Clin. Cancer Res.* **2015**, *21*, 5047.
- [36] M. Lawand, J. Déchanet-Merville, M.-C. Dieu-Nosjean, *Front. Immunol.* **2017**, *8*, <https://doi.org/10.3389/fimmu.2017.00761>.
- [37] P. Jiang, S. Gu, D. Pan, J. Fu, A. Sahu, X. Hu, Z. Li, N. Traugh, X. Bu, B. Li, J. Liu, G. J. Freeman, M. A. Brown, K. W. Wucherpfennig, X. S. Liu, *Nat. Med.* **2018**, *24*, 1550.
- [38] L. W. Pfannenstiel, C. M. Diaz-Montero, Y. F. Tian, J. Scharpf, J. S. Ko, B. R. Gastman, *Cancer Immunol. Res.* **2019**, *7*, 510.
- [39] O. J. Finn, *Ann. Oncol.* **2012**, *23*, viii6.
- [40] M. S. Rooney, S. A. Shukla, C. J. Wu, G. Getz, N. Hacohen, *Cell* **2015**, *160*, 48.
- [41] M. N. McCall, B. M. Bolstad, R. A. Irizarry, *Biostatistics* **2010**, *11*, 242.
- [42] S. Hänzelmann, R. Castelo, J. Guinney, *BMC Bioinformatics* **2013**, *14*, 7.
- [43] M. Schelker, S. Feau, J. Du, N. Ranu, E. Klipp, G. MacBeath, B. Schoeberl, A. Raue, *Nat. Commun.* **2017**, *8*, 2032.
- [44] E. Eisenberg, E. Y. Levanon, *Trends Genet.* **2013**, *29*, 569.