A semi-supervised approach for cell phenotypic and functional estimation in tissue microenvironment

Wennan Chang[1, 2+], Changlin Wan[1, 2+], Xiaoyu Lu[1], Szu-wei Tu[1], Yu Zhang[3], Brooke Richardson[1], Yifan Sun[1], Yingnan Hou[4], Xinna Zhang[1], Yong Zang[5], Anru Zhang[6], Kun Huang[7], Milan Radovich[8], Yunlong Liu[1], Xiongbin Lu[1*], Sha Cao[5]*, Chi Zhang[1, 2*]

[1]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, [4]Department of Medicine, [5]Department of Biostatistics, [7]Department of Medicine, [8]Department of Surgery, Indiana University School of Medicine, Indianapolis, IN,46202, USA.

[2]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, 46202, USA

[3]Colleges of Computer Science and Technology, Jilin University, Changchun,130012, China,

[4]Department of Mathematics, Shandong University, Jinan, 250100, China,

[6]Department of Statistics, University of Wisconsin–Madison, Madison, WI 53706, USA

*To whom correspondence should be addressed. +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Xiongbin Lu, Email: xiolu@iu.edu; Sha Cao, Email: shacao@iu.edu.

[+]These authors are with equal contribution to this work.

Abstract

Traditional deconvolution methods infer the relative proportions of predefined cell types through either regression- or enrichment- based approaches. However, there are several challenges that remain unsolved in the current formulations, including (1) identifying the Immune/Stromal (I/S) cell types that truly exists in a tissue, (2) identifying the marker genes for each cell type that are specifically expressed by one or a few I/S cell types in a TME, (3) co-linearity among I/S proportions due to their co-infiltration. We have developed a novel semi-supervised deconvolution method namely ICTD (Inference of Cell Types and Deconvolution), addressing the three challenges via (i) developing a Bi-Cross Validation (BCV) based matrix rank test to assess the significance of the existence of cell types and signature genes, (ii) utilizing a constrained Non-negative Matrix Factorization (NMF) to handle co-linearity caused by co-infiltration of multiple cell types. ICTD is shown to have largely improved prediction accuracy of tissue compositions, and it is capable of identifying novel or sub-cell types with specific functional activity.

Introduction

Deconvolution of tissue transcriptomic profiles aims to estimate the quantity of cellular components, as well as the cell type-specific functions and their cross-talks in the tissue microenvironment (TME) [1-4]. These algorithms usually assume the observed expression matrix of selected gene markers as a product of a cell type-specific expression signature matrix S and a cell type proportion matrix P, and they differ mostly on assumptions on the matrix S [2-4]. To learn about S, independent training data is usually needed to select representative genes that could maximally differentiate various cell types, thereby quantifying their expression levels [2]. The recent emergence of single cell RNA-seq (scRNA-seq) allows researchers to uncover new and potentially unexpected biological traits relative to traditional profiling methods that assess cell populations in bulk tissue, and thus represent unprecedently valuable training data in this matter. Cell Population Mapping (CPM) and CIBERSORTx seek to explain more genes' expression with incorporating the possible variations trained from scRNA-seq data, which also calls for a higher demand of the accuracy of cell proportion prediction [5]. Notably, the knowledge transfer from training data including single cells and other pure bulk cells, to unknown bulk tissue samples regarding matrix S, should be carefully handled, as the gene expression distribution of the two domains could be highly variable. Current deconvolution methods tend to oversimplify the knowledge transfer process, and often only rough transformation was applied to align the two domains [6]. The composition of the bulk tissues is of great interest to researchers, especially when it comes to novel or rare cell subtypes that are often indicative of the complexity of the TME. However, current deconvolution methods usually assume a fixed number of cell types, which clearly is incapable of identifying novel sub cell types [1-3]. Moreover, certain cell types such as immune cells tend to co-infiltrate, indicating that the proportions of these cell populations are highly co-linear [7]. As a result, the estimation of proportions in plain linear regression formation would suffer from the multi-collinearity, which would erratically change in response to small changes in the input data [8, 9].

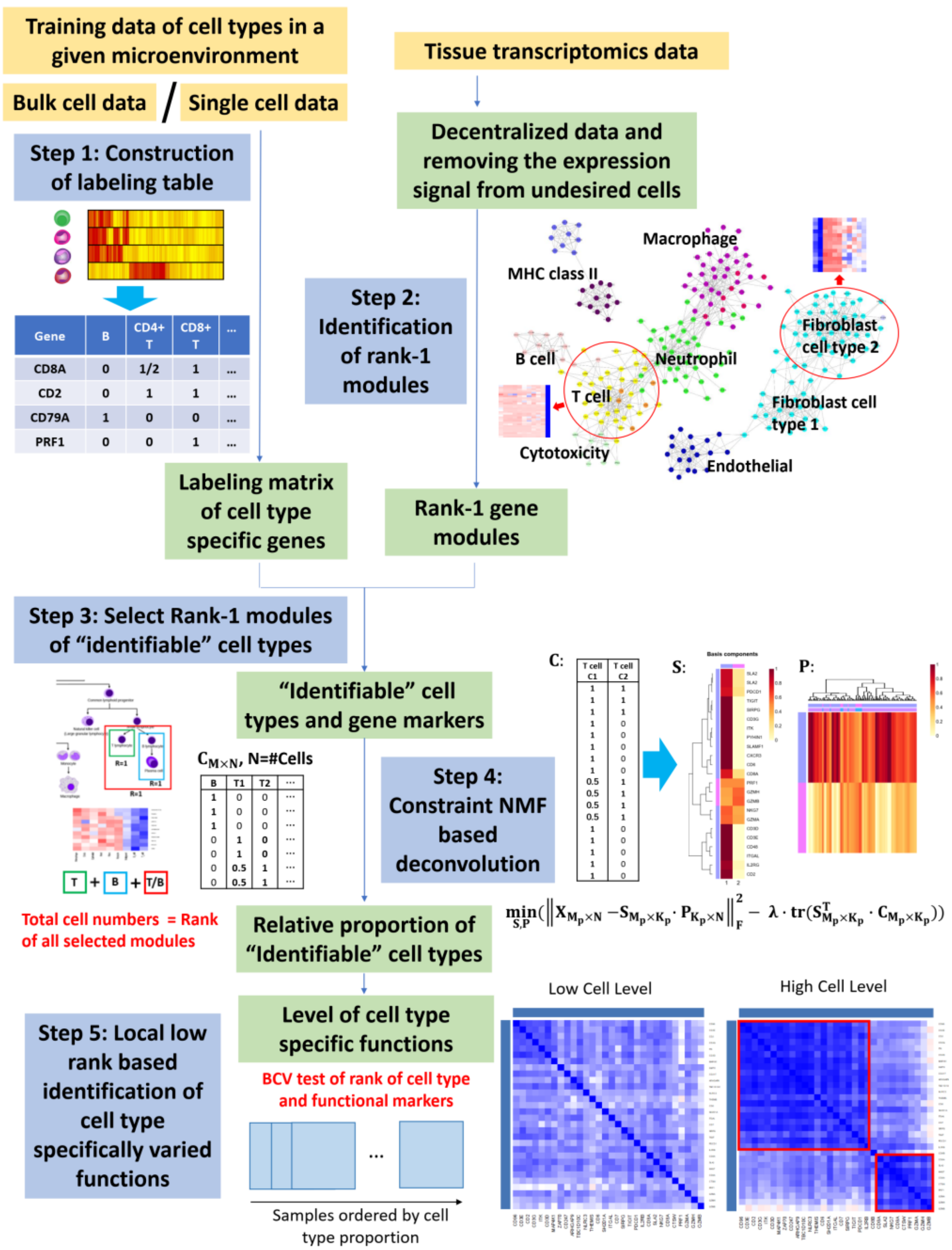
Hence, we summarize the key challenges of deconvolution methods as (i) detecting all existing (sub) cell types and their true marker genes in a specific context of TME; (ii) handling expression variations caused by different experimental platforms and batches from training to target data domain; (iii) dealing with the prevalent co-linearity in the cell type specific expression signatures as well as cell proportions; and (iv) defining the genes expression patterns representing varied cell type specific functions. More detailed discussions and comparisons of the formulations of existing methods are provided in the **Supplementary Note**.

Delineating all cell types and subtypes existing in a TME heavily relies on their specificities, as well as the resolution and scale of the collected data, in addition to the efficacy of the computational method. To put it in mathematical context, we conducted a comprehensive evaluation of the expression profiles of known cell type signature genes in both single cell and bulk tissue data sets of different disease microenvironment and experimental platforms. We derived the mathematical condition that a cell type is “identifiable” in a transcriptomic data only if (1) the cell type has uniquely expressed genes, the expression values of which over any subset of samples form a rank-1 matrix, or (2) there are genes expressed by the cell type and likely other cell types satisfying (1), but their expression values contributed by the cell type over any subset of samples form a rank-1 matrix; and a cell type-specific function is “identifiable” if there are marker genes of the function forming a local low rank submatrix in a subset of samples with significant presence of the cell type (**Online Methods and Supplementary Notes**). The stringent one or low rank condition grants the potential to detect novel cell subtypes from a given data, and yet there is no need to pre-specify the cell types.

We next developed a semi-supervised method, namely inference of cell types and deconvolution (ICTD), featured by: (1) a novel notion of “identifiable” cell types and marker genes; (2) a comprehensively extracted cell type signature genes used as information basis to annotate the identified cell types, which is TME-specific and data-driven; (3) a constrained non-negative matrix factorization (NMF) method to decrease the bias caused by knowledge transfer from training data, as well as to effectively handle the co-infiltrated cells; and (4) a local low rank screening approach to identify cell type specific functions.

Results

Our core algorithm ICTD consists of the following five steps (**Fig 1**): (1) Construct a labeling matrix from training data that represents cell type-specific gene expression for a given microenvironment. A labeling matrix $L_{\mathrm{M}\times\mathrm{K}}$ of M genes and K selected cell types is first trained from independent single or bulk cell transcriptomics data, which takes value in the set $\left\{1,\frac{1}{2},\frac{1}{3},\dots,\frac{1}{K-1}\right\}$, $L_{\mathrm{I,j}}=\frac{1}{l}$ if the expression level of gene i in cell type j is significantly lower than $l-1$ cell types and higher than the other $K-l$ cell types, and $L_{i,j}=0$ if gene i is not significantly expressed in cell type j (Supplementary Fig S1A). (2) Detect all rank-1 gene modules that are potentially markers of “identifiable” cell types. A potential cell type is recognized by ICTD as a gene module in which the genes are all highly expressed in one or several cell type according to the labeling matrix. ICTD particularly implemented a modified Bi-Cross Validation (BCV) for rank test, and a non-parametric hub and module detection method, to tease out all rank-1 modules [10]. In this step, ICTD can exclude undesired cell types, such as cancer or other disease cells, from further analysis, by non-negatively projecting the input data to the complementary space of the one spanned by the marker genes of undesired cell types. (3) Infer the number of “identifiable” cell types and cell type specific genes. The total number of “identifiable” cell types is computed as the total rank of the expression matrix of composed by all genes in the selected rank-1 modules in (2). Linear dependency among the selected modules is evaluated, and only those unique to one cell type are conserved. Here, each module is annotated by the genes’ significant enrichment to a cell type based on the labeling matrix. (4) Predict cell type proportions using constrained NMF. With the “identifiable” cell types and their uniquely expressed genes, a constraint matrix $\mathrm{C}_{\mathrm{M}\times\mathrm{K}}$ can be constructed for the NMF problem: $\mathrm{X}_{\mathrm{M}\times\mathrm{N}}=\mathrm{S}_{\mathrm{M}\times\mathrm{K}}\cdot\mathrm{P}_{\mathrm{K}\times\mathrm{N}}+\mathrm{E}$ [11]. Specifically, for cell type $k, k=1..K$ with M_{k} marker genes, $\mathrm{C}_{\sum_{k=1,..,K}\mathrm{M}_{\mathrm{k}}\times\mathrm{K}}[i,j]=1$, if gene i is marker of the cell type j, and 0 otherwise. The constraint matrix is then enforced upon the regular NMF formulation to guarantee similarity of the signature matrix with the constraint matrix, namely, we solve by $\min_{\mathrm{S,P}}\left(\|\mathrm{X}-\mathrm{S}\cdot\mathrm{P}\|_F^2+\lambda\cdot\mathrm{tr}\left(\mathrm{S}(1-\mathrm{C}^{\mathrm{T}})\right)\right)$. (5) Estimate cell type specific functions. For each cell type detected in (4), ICTD screens the matrix rank of the expression profile of the cell type and functional marker genes through the samples of different level of the cell types. Marker genes of a varied cell type specific function are identified if they form at least one distinct dimension comparing to the cell type markers only in the samples with high level of the cell.


Training data of cell types in a given microenvironment
Bulk cell data / Single cell data
Tissue transcriptomics data
Step 1: Construction of labeling table
Decentralized data and removing the expression signal from undesired cells
Step 2: Identification of rank-1 modules
Macrophage
MHC class II
B cell
Neutrophil
T cell
Fibroblast cell type 2
Fibroblast cell type 1
Cytotoxicity
Endothelial
Labeling matrix of cell type specific genes
Rank-1 gene modules
Step 3: Select Rank-1 modules of "identifiable" cell types
"Identifiable" cell types and gene markers
R=1
T + B + T/B
Total cell numbers = Rank of all selected modules
$C_{M\times N}$, N=#Cells
Step 4: Constraint NMF based deconvolution
C:
S:
P:
Basis components
$\min_{S,P}(\|X_{M_p\times N} - S_{M_p\times K_p}\cdot P_{K_p\times N}\|_F^2 - \lambda\cdot tr(S^T_{M_p\times K_p}\cdot C_{M_p\times K_p}))$
Relative proportion of "Identifiable" cell types
Level of cell type specific functions
BCV test of rank of cell type and functional markers
Step 5: Local low rank based identification of cell type specifically varied functions
Samples ordered by cell type proportion
Low Cell Level
High Cell Level

Gene	B	CD4+ T	CD8+ T	...
CD8A	0	1/2	1	...
CD2	0	1	1	...
CD79A	1	0	0	...
PRF1	0	0	1	...

B	T1	T2	...
1	0	0	...
1	0	0	...
1	0	0	...
0	1	0	...
0	1	0	...
0	0.5	1	...
0	0.5	1	...

Figure 1. Analysis pipeline of ICTD. Input data, analysis steps, and key results are yellow, blue and green colored, respectively. Rank 1 modules are identified by using decentralized data and the constrained NMF-based deconvolution is conducted on the input data of the "identifiable" cell markers. Based on the availability of training data and knowledge of the tissue microenvironment, the labeling matrix can be trained with a low resolution of cell types and further steps can identify potentially "identifiable" cell subtypes. The network in step 2 illustrate the co-expression association of nine cell type specific modules identified in a simulated bulk tissue data by using the human melanoma single cell RNA-seq data, and the two heatmaps show the expression profile of the identified fibroblast and T cell markers through different cell types (**Supplementary Note**). In step 4, the rank-1 modules of genes expressed by several cell types identified in step 3 are excluded from the NMF analysis and the relative proportions of the "cell types" defined by these modules are also computed and output for the downstream analysis. In step 5, marker genes of cell type specific functions are then identified by testing the matrix rank of the cell type and functional marker genes in the samples sliced by different level of the cell proportion. The two heatmaps in the step 5 illustrate the correlation between T cell marker genes and cytotoxicity markers in the samples of low T cell infiltration and high T cell infiltration level in TCGA COAD data, i.e. the cytotoxicity markers form a rank-1 module conditional to the T cell level specifically in the samples with high T cell infiltration level.

The core algorithms for each step are described in the **Online Methods**. Detailed algorithms, data used for method validation, and model comparisons with other methods, are provided in the **Supplementary Notes and Methods**. Below we present the application of ICTD on simulated bulk data using single cell RNA-seq data (**Fig. 2**) and real tissue data (**Fig. 3**), from which we demonstrate (1) the ability of ICTD to identify both known and novel (sub) cell types with a high analysis resolution, (2) the accuracy and goodness of fitting of ICTD in analyzing data of different tissue microenvironment and experimental platforms, (3) the robustness of ICTD in cases when cell type are with highly correlated proportions, (4) inference of cell type specific functions, and (5) the biological and clinical implications derived by correlating ICTD predicted cell and functional levels with other omics, imaging and clinical data.

We first benchmarked ICTD on predicting the existence of cell types and their relative proportions in scRNA-seq data simulated bulk tissue datasets and compared with three state-of-art deconvolution methods, namely CIBERSORT, TIMER, and EPIC (**Online Methods**). The bulk tissue datasets were simulated by RNA-seq data of single cells from five different TMEs, including five types of human solid cancer (namely, breast, colon, head and neck, lung, and melanoma), three types of human brain cancer (namely glioblastoma, oligodendroglioma and astrocytoma), two types of normal human brain microenvironment, three types of human immune microenvironment (monocyte and dendritic cell, lymphoid, and myeloid progenitor cells), and mouse melanoma microenvironment. On all five types of human solid cancer microenvironment, all available cell types in the input scRNA-Seq data were detected as “identifiable” cell types by ICTD. In addition, ICTD achieved significantly higher accuracy in predicting total B-, T-, mast, fibroblast, endothelial cells and macrophages comparing to other methods. On 92% of the 25 cells type in the simulated bulk cancer datasets, their relative proportions predicted by ICTD have at least 0.95 Pearson correlation with their true proportions, while the average correlation is 0.86, 0.63 and 0.52 for EPIC, TIMER and CIBERSORT, respectively (**Fig 2a**). On the five types of simulated human normal and brain cancer microenvironment, ICTD also successfully detected astrocyte, oligodendrocyte and progenitors, exhibitory and inhibitory neuron, microglial and Schwann cells as identifiable cell types, all with high accuracy in predicting relative proportions (**Fig 2b**). Similarly, ICTD also accurately identified sub cell types from the mixture of multiple classes of monocyte and dendritic cells, human lymphoid and myeloid progenitors, and the immune and stromal cells in mouse melanoma microenvironment and can accurately assess their relative proportions (**Fig 2b**).

A unique feature of ICTD is its capability to detect novel cell types along with their marker genes that could effectively provide biological annotation of the cell types. Our analysis on simulated cancer tissue data suggested that each of the rank-1 module corresponds to one cell or sub-cell type (Supplementary Fig X). On the simulated human solid cancer bulk tissue datasets, ICTD was able to identify subtypes of immune/stromal cells, such as CD4+ and CD8+ T cells, and subtypes of fibroblast and myeloid cells (Supplementary Table SX and Fig SX). As illustrated in **Fig 2c**, among the three rank-1 modules identified by ICTD, one clearly corresponds to the general fibroblast (with COL1A1 expression) type, and the other two correspond to two fibroblast subtypes (with COL8A1 or SERPINF1 expression) in the simulated human melanoma data. The cell type markers are further validated by the tSNE visualization, where the expression level of each marker set has a good correspondence a certain cell type or subtype. It is noteworthy that the number of identifiable cell types could vary through disease contexts and data sets (**Fig 2d**), indicating that it may not rationale to fix the cell types beforehand, as was used in other methods. We further evaluated the variation of cell type specific markers through different disease contexts and data set. As shown in **Fig 2e**, there is a strong disease context specificity of T cell markers, where only four common T cell markers were identified in all the five data sets, and 19 T cell markers were identified in four out of the five data sets. We observed on average 93.75%, 90.36% and 83.33% of the T cell markers utilized in CIBERSORT, TIMER and EPIC are specific to three or less cancer types and only 65.21%, 69.57% and 13.04% of the common cell type marker genes were included in their signature matrix. Similar patterns are also seen for B, dendritic, myeloid, endothelial and fibroblast cells (Supplementary Fig X). In contrast, ICTD considers the variations in both of identifiable cell types but also cell type markers in different TMES and datasets, resulting in a better prediction accuracy (**Fig 2e-f**).

An explanation score (ES) is defined for each marker gene to evaluate the goodness of fitting of the gene’s expression by the predicted proportions of the cell types expressing the gene (**Supplementary Methods**). Intuitively, high ES scores of the marker genes for one cell type suggest high prediction accuracy and specificity of the marker genes. We observed strong positive correlations between the ES scores and prediction accuracy using ICTD and EPIC, as these two methods rely on cell type uniquely expressed genes. Similarly, for CIBERSORT and Timer, positive associations were also observed (**Fig 2g**). Analysis of six major immune and stromal cell types in five simulated bulk data sets suggested that the prediction accuracy cannot exceed 0.8

when ES<0.8; while ES>0.9 is a necessary condition for a prediction accuracy to be above 0.9 (Fig 2g). It is noteworthy that the ES of all the general immune and stromal cell markers in the simulated data identified by ICTD were as high as above 0.95, due to stringent rank-1 threshold.

ICTD also demonstrated its superiority in handling co-linearity of cell proportions, caused by cells' co-infiltrations. Our preliminary analysis on TCGA data suggested correlation among the immune and stromal cells to be as high as 0.57 (**Supplementary Notes**). To evaluate the robustness of ICTD in estimating the cell types proportion with a co-linearity, we simulated batches of bulk tissue samples in each of which the cell proportions are intentionally set to have different levels of correlations to mimic co-infiltration (**Online Methods**). Not surprisingly, ICTD achieved good robustness and high prediction accuracy at different levels of co-infiltration. On the contrary other methods that are based on linear regression showed significantly decreased prediction accuracy when co-infiltration level was high. ICTD overcomes the co-linearity issue owing to its data-driven selection of cell type specific markers and constrained NMF formulation. The four methods' prediction accuracy of B and T cells across different co-infiltration levels in simulated human melanoma tissue data is shown in Fig 2h. In addition, significant associations among ES, prediction accuracy, and co-linearity levels were identified (Supplementary Fig X).

ICTD can identify varied function of a certain cell type by using a local low rank based formulation [12]. In human head and neck cancer data, we identified varied expression level of cytotoxic gene in the CD8+ T cells in different patient stratifications, suggesting varied T cell exhaustion levels (**Supplementary Note**). To evaluate the capability of ICTD in identifying varied T cell cytotoxicity level, we simulated bulk tissue data with different proportion and cytotoxicity level of T cells, by using the inter- and intra- sample variance derived from TCGA and single cell data (**Online Methods**). ICTD conducted a local low rank screening with a kernel function along samples ordered by predicted T cell proportions. Our analysis clearly identified the linear spaced spanned by the T cell and cytotoxicity marker genes clearly switches from rank-1 to rank-2 through the samples with low to high T cell levels, suggesting the identifiability of varied cytotoxic level in the samples of high T cell infiltrations (**Fig 2i-j**). On average, correlation level of 0.86 between the true cytotoxicity level per unit T cell and the prediction made by ICTD was observed (Supplementary Fig SX).

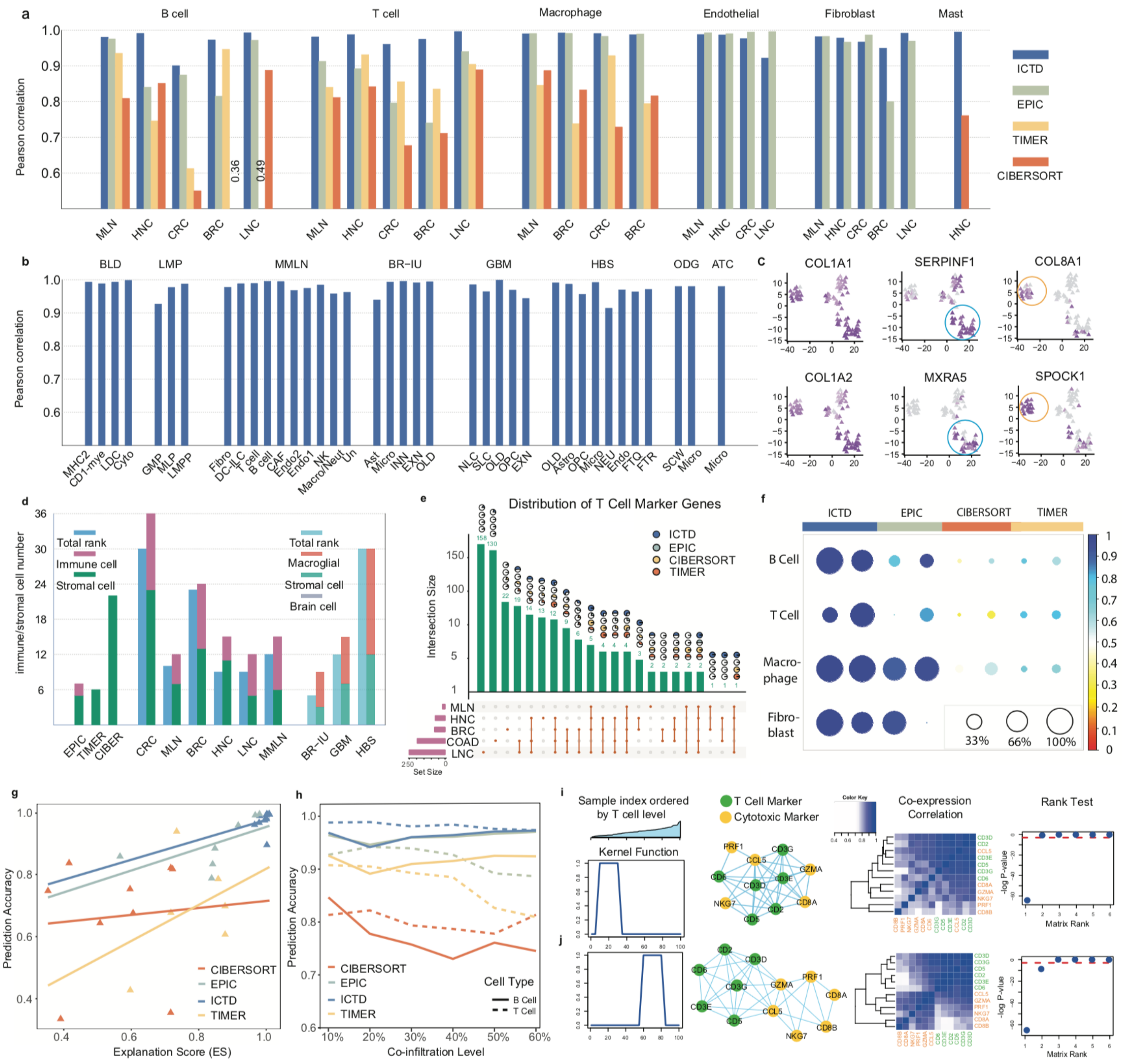


Figure 2. Validation of ICTD by using single cell simulated bulk tissue data. (a) Pearson correlation between known and predicted proportion of B, T, macrophage, endothelial, fibroblast and mast cells by ICTD, EPIC, TIMER and CIBERSORT, in the bulk tissue data simulated by the scRNA-seq data of Melanoma (MLN), Head and Neck Cancer (HNC), Colorectal Cancer (CRC), Breast Cancer (BRC), and Lung Cancer ((LNC). (b) Pearson correlation between known and predicted proportion of cell types and subtypes identified by ICTD in the bulk tissue data simulated by scRNA-seq data of myeloid and dendritic cell mixture (BLD), lymphoid and myeloid progenitor mixture (LMP), mouse melanoma (MMLN), normal brain cells nucleic sequencing generated in this study (BR-IU), glioblastoma (GBM), human normal brain (HBS), oligodendroglioma (ODG), and astrocytoma (ATC). Detailed cell type codes are given in Supplementary Table X. (c) t-SNE plot of fibroblast subtypes identified by ICTD in simulated human melanoma tissue data. (d) Consistency of the number of ICTD identified cell types and the matrix rank of the expression profile of the markers of identified cell types. (e) Distribution of the T cell marker genes identified in the five cancer data and their overlap with T cell signature genes used in CIBERSORT, TIMER and EPIC. Each bar and number represent the number of T cell markers in certain cancer types labeled in the dot plot below. The pie charts illustrate the proportion of the T cell markers in certain cancer types used as T cell markers in ICTD, CIBERSORT, TIMER and EPIC. It is noteworthy that the ICTD identifies

TME and data specific markers while the other methods assume constant markers. (f) Accuracy of cell type specific markers used by each method. The circle size represents the ratio of cell type signature genes used in each method as the true cell type specific markers, scaled as the uncolored circles in the bottom right, while the color represents the prediction accuracy. The left and right column of each method illustrate the results of MLN and HNC data. Statistics of the other three cancer types were provided in Supplementary Fig SX. (g) Dependency explanation score and prediction accuracy of the cell type proportions predicted by the four methods. X- and y-axis represents explanation score and prediction accuracy, respectively. (f) Prediction accuracy of the proportion of T and B cells made in the data with different level of T and B cell co-infiltration. X- and y-axis represents the Pearson correlation of co-infiltration and prediction accuracy, respectively. Detailed simulation method is given in **Online Methods**. (i-j) prediction of varied T cell cytotoxicity level in simulated HNC data. From left to right, the four plots illustrate the kernel function used for local low rank screening, co-expression correlation network between T cell and cytotoxic marker genes, heatmap of correlations between T cell and cytotoxic marker genes, and p values of the matrix rank of the gene expression profile of T cell and cytotoxic marker genes, in the samples of low T cell infiltration (i) and high T cell infiltration level.

We then applied ICTD on a collection of human cancer, normal, blood and inflammatory tissue (CNBI) data, including 28 cancer and 11 normal tissue types from TCGA, 17 colorectal cancer, 7 triple negative breast cancer, 7 blood tissue, and 11 human inflammatory disease data sets from GEO (Supplementary Table S1). Similar to the analysis of the simulated tissue data, we identified rank-1 markers of B, T, dendritic, general myeloid, macrophage, monocytes, neutrophil, fibroblast, endothelial and adipocyte cell and their sub cell types in each dataset (**Fig 3a** and Supplementary Fig SX). A strong association between the number of identified cell types and the total rank of the matrix of marker genes was observed (**Fig 3b**). It is noteworthy that the cell types most variable across different cancer types are the subtypes of Myeloid, Fibroblast, and Adipocytes, which seem to be higher in breast, colorectal, lung, pancreatic and stomach cancers, commonly known to have more stromal components. The complete set of cell types and their marker genes identified in each data set were summarized in Supplementary Table SX. In the TCGA datasets, 21 common "identifiable" cell and subtype types have been observed in more than 10 cancer types, including CD19/CD22 expressing regulatory-like and CD79A/CD79B expressing activated B cell; total, CD8+, and CD4+ T cell; Neurexin and Caytaxin expressing Neuron cell; myofibroblast-like cell; Collagen 1/3/5, Collagen 4/15/18, Collagen 6, and Non-collagen expressing Fibroblast; Endothelial cell; MHC class II antigen presenting cell; MHC class I, pro-inflammatory cytokine releasing, chemokine and cytokine releasing Myeloid cells; complement pathway activated Macrophage and Monocytes; granulocytes; and adipocytes (**Fig 3c**). It is noteworthy that markers of each commonly identifiable cell types have certain overlaps with the immune and stromal cell markers identified in normal tissue data, suggesting these marker genes were not affected by cancer cells (Supplementary Fig X). On average, the ICTD marker genes of each cell type have ES higher than 0.9, while the ES scores of the signature genes used by CIBERSORT, TIMER and EPIC are 0.26, 0.39 and 0.46, respectively. **Fig 3d** illustrate the ES of T and B cell (sub)type markers of the four methods. The level of tumor infiltrated lymphocytes (TIL) in 12 TCGA cancer types have been previously assessed by imaging data [13]. On average, the correlation between imaging predicted TIL and ICTD predicted T cell level is 0.4, comparing to 0.14, 0.2, and -0.11 with CIBERORT, TIMER, and EPIC predicted T cell level (**Fig 3e**). On 3552 pre-identified immune and stromal cells specifically expressed genes, we compared how well ICTD-predicted proportions explain the cell type specific expressions with the three methods. It turns out that ICTD-predicted cell proportions achieved on average 0.56 R^2 value in explaining the expression level of predefined immune and stromal cell specifically expressed genes, while the R^2 is 0.2, 0.24, and 0.18 for CIBERSORT, TIMER, and EPIC (Supplementary Table X). Application of ICTD on 7 human normal brain, 5 neuro-degenerative disease and 4 brain cancer data sets identified 23 common cell types in brain microenvironment, including eight subtypes of astrocytes, glial cell and oligodendrocytes, exhibitory and inhibitory neuron, MHC class I and II antigen presenting cells, innate immune responsive microglial and two other microglial subtypes, one endothelial, four ependymal, and two stromal-like cell types (Supplementary Fig XX).

A critical assumption of ICTD Is that the existent cell types and their marker genes are varied across tissue microenvironments and technology platforms, which was validated by our comprehensive data analysis. **Fig 3f** illustrate the ES of T cell expressing genes in different CNBI data sets, suggesting a significant variation of the T cell markers in the TME of different cancer, inflammatory disease and blood samples, as well as under different experimental platform [14, 15]. To further investigate how the data set specific makers vary by disease/tissue micro-environments or experimental platforms, we further computed the averaged Jaccard distance between the marker genes of same cell types identified in any two CNBI or single cell simulated bulk datasets (**Supplementary Methods**). As illustrated in **Fig 3h**, the cell type marker genes vary drastically between cancer, normal inflammatory and blood tissues. Three distinct clusters were observed (1) TCGA cancer and other cancer, (2) single cell simulated cancer, and (3) TCGA normal and other inflammatory disease, and blood tissue. Among the cancer data, TCGA cancer data set and other RNA-seq based data sets is well separated from scRNA-seq simulated data and the Affymetrix Microarray data sets, and the later one is further divided into two sub-clusters containing independent CRC and TNBC data sets. Similarly, the TCGA RNA-seq and microarray data of normal, inflammatory conditions, and blood tissue form three distinct sub-clusters. Among the microarray data of inflammatory conditions, the disease of digestive system and airway and skin tissues from two sub-clusters.

ICTD detected general T cell, fibroblast, and myeloid cells in all 28 analyzed TCGA cancer types, while the CD8+ T, non-collagen extracellular component expressing fibroblast, and oxidative stress producing myeloid cells were identified as distinct cell types in only 10, 12, and 15 cancer types, respectively. We found that the markers of these functional sub cell types are detected as cell type specific functions rather than a cell type in some cancer types by the local low rank screening function. For the 19 cancer types where CD8+ T cell is not identified as a cell type, CD8+ T cell markers are identified as varied T cell specific function in 15 cancer types, while in 4 cancer types, high concordance is observed between total T cell and CD8+T cell markers in all the samples, making the CD8+ T subtype not differentially from the general T cell. **Fig 3i** illustrated the marker genes of general T, CD8+ T, CD4+ T and T-reg cells form a distinct rank-4 submatrix in samples with high T cell infiltration, while the genes were less distinguishable in the complete TCGA COAD data (**Fig 3j**). This suggests the "locality" of finding identifiable cell types and functions, and hence it is necessary to implement a local low rank module detection approach. Similar locality was also observed for the marker genes of CD4+/CD8+ T, non-collagen expressing fibroblast and NADPH oxidase expressing myeloid cells in certain TCGA cancer types and other analyzed CRC and TNBC data sets (Supplementary Fig X). We also conducted comprehensive screening to identify unknown immune/stromal cell type specific functional genes (**Online Methods**). 84 major functional modules were identified as common cell type specific functions in TCGA data (**Supplementary Notes**).

The deconvoluted cell type functions in a collection of tissue samples makes it possible to computationally characterize cell-cell interactions at low cost. We observed co-infiltrations among immune and stromal cell types with Pearson correlation in the range of 0.2-0.7 in all the analyzed TCGA cancer data (Supplementary Table X). The inhibition or promotion of a function in cell type A on cell type B could now be seen by the correlations between the abundance level of B and the activity level of the function in A, conditional to the predicted proportion of A. Consequently, we found seven genes expressed by fibroblast cells with significant negative conditional correlation with T cell infiltration in at least 10 out of 15 cancer types with high level of stromal cells ($p<0.01$) (**Fig 3h**). The genes execute functions related to the modification and synthesis of collagen and extracellular polysaccharide, suggesting a possible role of the dysregulated extracellular matrix composition in affecting the T cell infiltration. Similarly, the interactions of functions in two cell types can be computed by the correlation of the levels of the two functions conditional to the proportion of the two cell types. We identified a low association among CD8 T cell markers such as CD8A/CD8B and cytotoxic genes conditional to the total T and NK cell level in 19 cancer types and a high correlation among general T, CD8+ T, and cytotoxic genes in 4 cancer types, suggesting possibly perturbed cytotoxicity of T cells, like T cell exhaustion, in the first 19 cancer types. We also observed a significant negative correlation ($p <0.01$) between the NADPH oxidase and T cell cytotoxicity levels conditional to the total myeloid and T cell in 11 out of the 25 TCGA cancer types. This is consistent with previous observation that NADPH oxidases produce reactive oxygen species (ROS) on the surface of myeloid-derived suppressor cells that suppress the cytotoxic function of T cells [16].

We further investigated the clinical implication of the cell types and cell type-specific functions inferred by ICTD, by associating the predicted cell proportions and varied functions with patient's overall survival in TCGA data, as well as five clinical trial data with immune checkpoint inhibitor treatment (**Supplementary Notes**). We identified significant associations of patients' overall survival with T cell infiltration and relative cytotoxicity levels in 12 and 7 TCGA cancer types, respectively. More interestingly, in colorectal and ovarian cancer, we observed that patients with moderate level of T cell infiltration have the best overall survival comparing to the patients with high and low T cell levels (**Fig 3k**). We define the T cell's relative cytotoxicity (RC) level as the predicted cytotoxic level divided by the predicted total T cell level in each sample and observed patients with higher RC have significantly better overall survival. This clearly suggests the existence of T cell exhaustion and its association with poor prognosis (**Fig 3k**). On the five clinical trial data, we noticed that patients with high T cell infiltration have better response to the treatment (**Fig 3l**), which is consistent with previously reported [ref]. Moreover, the level of T cell cytotoxicity was observed to vary significantly in four data of melanoma, lung adenocarcinoma and lung squamous carcinoma. We observed the patients with lower RC tend to have better clinical response (**Fig 3m**), possibly due to more PD-1/PD-L1-mediated immuno-suppression in these tumors. It is noteworthy that association between T cell infiltration and patients' clinical outcome, and the identifiability of varied cytotoxic function show a high consistency between TCGA and clinical trial data (Supplementary Table X).

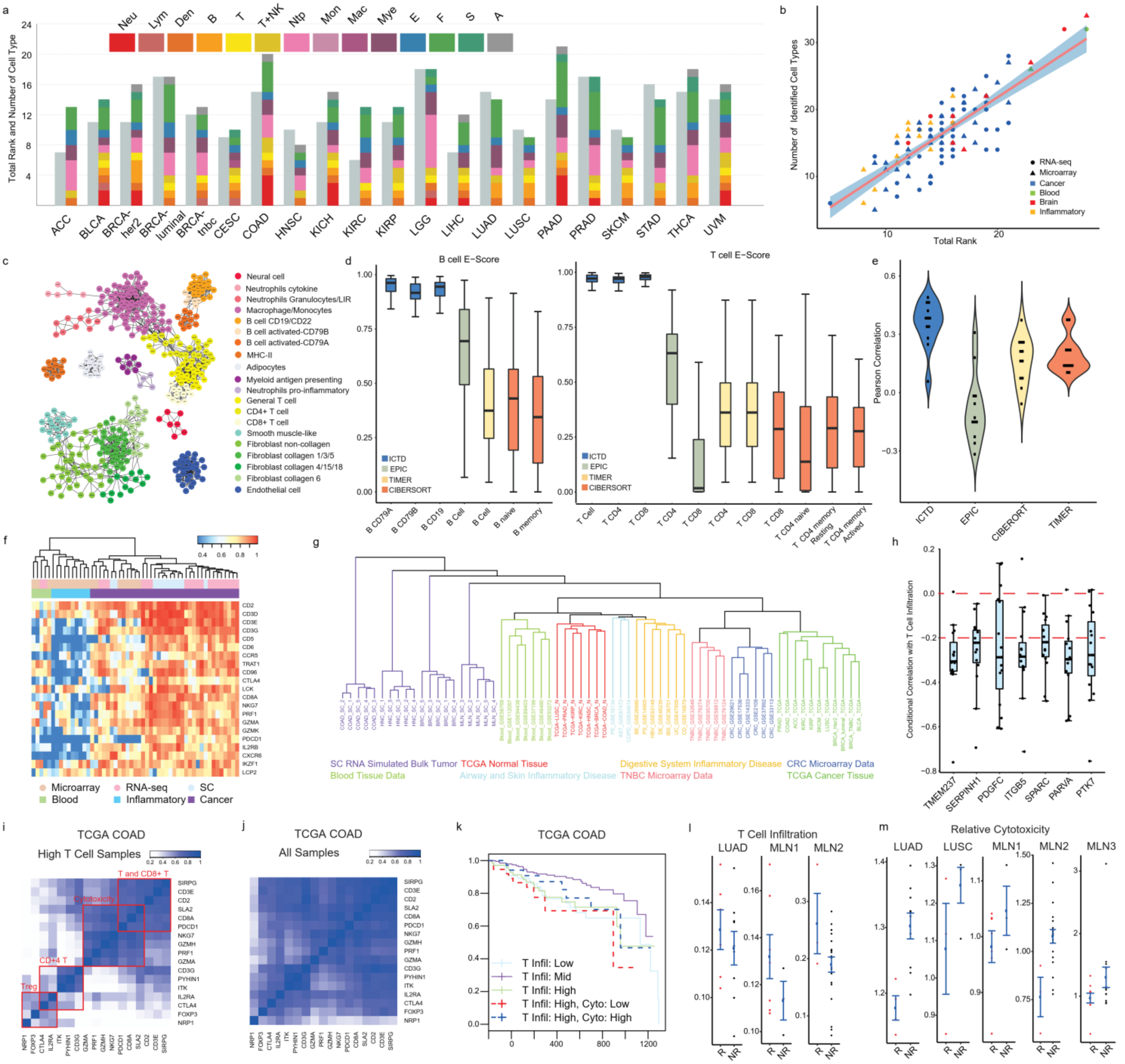


Figure 3. Application of ICTD one real bulk tissue transcriptomics data. (a) Consistency of the number of ICTD identified cell types and the matrix rank of the expression profile of the markers of identified cell types in selected TCGA data. (b) Significant correlation between the number of cell types and the matrix rank of the marker genes identified by ICTD through all the analyzed data. (c) Network of the marker genes of the commonly identified cell (sub) types in the TCGA data. An edge between two genes means the two genes are co-identified as markers of one cell type in more than 10 analyzed TCGA data. (d) E-score of the B and T cell marker genes identified by ICTD and used by EPIC, TIMER and CIBERSORT in TCGA data. Statistics of other cell types are given in Supplementary Fig SX. (e) Correlation between the imaging data derived tumor infiltrated lymphocyte level and T cell proportion predicted by the four methods in 11 TCGA cancer. (f) Consistency of T cell markers in identified CNBI data. In the heatmap, each row is the commonly identified T cell markers and each column is one data set. The heatmap represent the E-score of each gene in each data set. Statistics of other cell types are

given in Supplementary Fig SX. (g) Consistency of overall identified immune and stromal (sub) cell types and marker genes identified in the CNBI data. Detailed formula the distance between of the cell types identified in two disease types is given in **Online Methods**. (h) Correlation between selected fibroblast cell expressing genes and T cell infiltration level conditional to the fibroblast cell level in TCGA data. (i-j) Co-expression correlation between T cell, CD8+ T cell, cytotoxic function, CD4+ T cell and T-reg marker genes in the samples with high T cell infiltration (i) and all samples (j) in TCGA COAD data. (k) Survival curves of the TCGA COAD patients with low, medium and high T cell infiltration, and the high T cell infiltration patients with low and high cytotoxic functions predicted by ICTD. (l) Variation of T cell infiltration level in response (R) and non-response (NR) patients in three independent checkpoint inhibitor treated clinical data. (m) Variation of T cell relative cytotoxic level in response (R) and non-response (NR) patients in five independent checkpoint inhibitor treated clinical data. LUAD, LUSC and MLN* represents different sets of lung adenocarcinoma, lung squamous cell carcinoma and melanoma.

Discussion

Our semi-supervised deconvolution method ICTD brought up the notion of “identifiability” of a cell type and cell type specific functions, and consequently, suffers much less bias resulted from fixing the cell types and their marker genes. ICTD outperformed other methods not only in accurately predicting cell type proportions, but also in that it enables detection of novel cell (sub) types, and evaluation of cell type functional activities, which opens up the possibilities of characterizing cell-cell interactions in large-scale tissue transcriptomic profiles. It is noteworthy that the transcriptionally “identifiable” cell types differ from those defined by cell differentiation lineage: some cell types on the lineage may not be identifiable, while an “identifiable” cell type can be a certain cell or cell subtype, or the total of several cell types on the lineage that express same gene markers. Nonetheless, we believe deconvoluted cell types defined this way is more data-driven, less biased to the training data, and more suitable for downstream correlation analysis with other clinical and biological features. ICTD relies on good training data to construct the labeling matrix, and we noticed the labeling matrix trained from scRNA-seq data in the TME of a certain cancer type cover more tissue specific cell type markers than those trained from microarray data of primary cells collected from healthy donors (**Supplementary Notes**). It is worthy of mention that since ICTD is not fully supervised, we suggest at least 20 samples is needed for the method to work. While the method has increased type II error when the sample size is small, the identified rank-1 genes can still guide the flexible selection of signature genes integrated with a regression based approached, and this feature is available for our ICTD R package. Our analysis of a post-chemotherapy data set suggested ICTD could be sensitive to outlier samples in the rank-1 module identification step (**Supplementary Methods**), and in the ICTD R package, it is suggested to remove outlier samples before the analysis. Application of ICTD on TCGA pan-cancer data identified variations of T cell marker, cytotoxic marker and T cell exhaustion level, association between fibroblast expressing genes and T cell infiltration level, and association between ROS produced by myeloid cell and T cell cytotoxic level in different cancer types, suggesting the capacity of ICTD in providing a comprehensive evaluation of TME specific cell types, cell type specific function, and cell-cell interactions. Nevertheless, the sensitivity of detecting cell type varied function can be largely improved if the functional marker genes are predefined, and additionally, more novel cell type functions can be predicted if the module detection approach could be replaced by a local low rank module detection method. The semi-supervised formulation minimizes the bias in data platform and batches in knowledge transfer. A possible application of the ICTD framework is to transfer the cell type or functional gene modules derived from scRNA-seq of a small or moderate amount of sample, to guide the characterization of a certain TME in large scale bulk tissue data.

Methods

Single cell, bulk cell and tissue transcriptomics data sets used in this study

The semi-supervised formulation identifies the rank of a gene’s expression through considered cell types instead of its exact expression level, hence a much larger set of training data can be utilized to maximize the utilization of training data. Considering the availability of number of samples and data sets, we collected large scale bulk cell data of 11 cell types in human cancer and inflammatory tissue microenvironment, 8 cell types in human brain microenvironment generated by Affymetrix UA133 plus 2.0 Array, and 13 cell types in mouse cancer and inflammatory tissue microenvironment generated by Affymetrix Mouse Genome 430 2.0 Array, with cell types detailed below: *human stromal and immune cells:* fibroblast , adipocytes, endothelial cell, B cell, CD4+ T cell, CD8+ T cell, natural killer cell, dendritic cell, monocytes, macrophages, and neutrophil; *human central nervous* *system:* neuron, Schwann cell, astrocyte, ependymal cell, oligodendrocyte, and microglial cells, endothelial, and stromal-like cell; *mouse stromal and immune cells:* fibroblast, adipocytes, myofibroblast, endothelial cell, B cell , CD4+ T cell, CD8+ T cell, natural killer cell, dendritic cell, monocytes, macrophages, neutrophils, and mast cell. We believe these cell types in combination with disease cells and tissue specific cells can cover major cell populations in the tissue microenvironment of solid cancer, inflammatory disease, brain and hematopoietic system [17]. CCLE cell line, human and mouse tissue index, and other cancer cell line and tissue data were utilized as background expression profile for marker training.

The method was validated on single cell simulated bulk tissue data by using 13 public single cell RNA-seq data sets generated by either C1/SMART-seq2 or 10x Genomics pipelines, including cells collected from (1) the TME of human solid cancer melanoma, breast, colorectal, head and neck, and lung cancer, (2) human glioma, oligodendroglioma, and astrocytoma, (3) two human normal brain sets, (4) human myeloid cell lineage and lymphoid cell lineage and monocyte/dendritic cell populations, and (5) TME of mouse melanoma.

We applied ICTD on bulk tissue transcriptomic data of (1) 28 TCGA cancer types, (2) 11 TCGA normal tissue data, (3) 17 independent microarray data sets of colorectal cancer measured by different platforms; (4) metabric and 6 other triple negative breast cancer data sets; (5) 7 blood tissue RNA-seq and microarray data; (6) 11 human inflammatory disease data sets generated by Affymetrix UA133 plus 2.0 Array, and (7) 7 human normal brain, 5 neuro-degenerative disease and 4 brain cancer types. Detailed information of the bulk cell, scRNA-seq and bulk tissue data were provided in Supp Table 1S. Public data selection, downloading and processing procedures, and sample and sequencing information of the newly generated data are given in **Supplementary Note**.

Preliminary derivation of the mathematical conditions of "Identifiable" cell types and cell type specific functions

A preliminary evaluation of the gene expression scale and matrix rank of the cell type uniquely expressed genes and signature genes derived by other deconvolution methods was conducted. As detailed in **Supplementary Note**, we analyzed the following characteristics of the genes in the scRNA-seq and bulk tissue data of different disease context, experimental platforms and batches: (1) the consistency and specificity of cell type uniquely expressed genes were evaluated by their averaged expression level through different cell types of different data sets; (2) inter- and intra- sample variations of cell type signature genes were characterized by the "drop-out" rates and multimodality of each gene's expression profile in the scRNA-seq data of different samples, computed by using a left truncated mixture Gaussian distribution; (3) matrix rank and expression scale of cell type uniquely expressed genes in bulk tissue data were evaluated by using BCV based rank test and Kolmogorov Smirnov (KS) test, and (4) immune and stromal cell co-infiltrations in cancer and inflammatory tissues were further assessed by using the averaged co-expression correlations among a small number of known cell type uniquely expressed genes that truly form rank-1 structure in each data set.

Our evaluation suggested that the uniqueness condition of an NMF problem cannot hold for the gene with expression spanned by more than one cell type's proportion (**Supplementary Note**). Hence only the cell type with uniquely expressed genes are transcriptomic "identifiable", and the markers genes should also be stably expressed in the cell type so that its expression level can reflect the cell's population. Specifically, if gene i is uniquely and stably expressed in cell type k, its gene expression can be expressed as $\mathrm{X}_{i,\cdot} = \mathrm{S}_{i,k} \cdot \mathrm{P}_{k,\cdot} + e$, where $S_{i,k}$ is the unit expression of i in k, and $P_{k,\cdot}$ is the relative proportion of cell type k across all the samples. This shows that genes uniquely expressed by a cell type forms a (matrix) rank-1 submatrix spanned by the vector of relative proportion of the cell type. On the other hand, a significant rank-1 structure of the expression profile of multiple genes $\mathrm{X}_{i,\cdot} = \sum_k \mathrm{S}_{i,k} \cdot \mathrm{P}_{k,\cdot} + e, i = 1 \dots m$ suggests that these genes are highly possibly expressed by a dominating cell type in the current tissue microenvironment or the genes are with similar expression pattern in several cell types.

Noting cell type specific functional activities, such as T cell cytotoxicity, reactive oxygen species production by myeloid derived suppressor cells, and synthesis of certain extracellular components by fibroblast cells, are highly varied in through tissue and disease micro-environment of different patients, thus it is not feasible to use constant gene expressions to characterize their activities. Even a group of functional related genes. If there is no variation, the cell type specific functional genes should share the same rank-1 space with the cell type markers. If all the samples have a variation, the functional genes should be identified as the markers of a cell type defined by the function. Only if a subset of samples are with the functional variation, the low rank structure of the functional genes will be absorbed by the cell type markers and diminished on the co-expression network of all the samples. For such a case, the linear base of the varied function can be distinguished when the computation was limited to the samples with the functional variation, i.e. a local low rank identification method is needed.

Construction of labeling matrix

A labeling matrix $L_{M\times K}$ was first constructed to represent the genes that are overly expressed in a certain cell type, where *M*=number of genes and *K*=number of cell types, $L_{i,j}=\frac{1}{R}$ stands for the gene $G_i's$ expression in cell type C_j is among the *R*th highest among its expression in all the cells, and $L_{i,j}=0$ stands for G_i is not a significant signature of cell type C_j. The labeling matrix is to ensure the rank-1 modules identified by ICTD are true cell markers, and is used for inference and annotation of the cell type and functions of the identified markers.

To construct the labeling matrix, a non-parametric random walk based approach was developed to identify if a gene is more expressed in certain cell types comparing to others by using the collected bulk cell data. Cell types and analysis resolution of sub cell types are first selected for the training of a tissue context specific labeling matrix.

Exclusion of the expression of undesired cells

Before the rank-1 module identification, ICTD eliminates the low rank space spanned by the linear bases of the gene expressions from predefined undesired cell types. To maximally eliminate the impact of the gene expressions from undesired cells, ICTD first identifies gene co-expression modules from the expression matrix of predefined marker genes by using WGCNA and computes the first row base of each module by using SVD [ref]. The genes in the whole data that are positively correlated with the first row base(s) of one or several module(s) are further project to the complementary space of the row space spanned by the base(s). Denote the whole gene expression data as X, the data of pseudo-code of exclusion of the expression of undesired cells are given below:

```
$Modules_C \leftarrow WCGNA(X_C)$
$for\ i\ in\ Modules_C$
$U_i\Sigma_iV_i^T = SVD(X_i)$
    $RB_c[i,] \leftarrow V_i^T[,1]$
$for\ gene\ in\ X$
    $for\ k\ in\ 1{:}K$
        $if(\max(cor(RB_c, X_{genes})) > 0)$
            $i \leftarrow \underset{i}{\operatorname{argmax}}(cor(RB_c[i,], X_{genes}))$
            $X_{gene} \leftarrow X_{gene} - X_{gene}\frac{RB_c[i,]RB_c[i,]^t}{||RB_c[i,]||^2}$
$return(X)$
```

In this paper, we first identified 1089 genes consistently up-regulated in 11 cancer types of TCGA data and with significant expression in CCLE cell line data, as the genes commonly expressed by cancer cells (Supp Table XX). Differential gene expression analysis was conduct by using Mann-Whitney test with FDR<0.05 as the significant cutoff and significant expression in cancer cell line data is determined by log(FPKM)>2. In the analysis of one specific cancer type, gene co-expression modules of the cancer genes were first identified. The linear space spanned by the modules were further excluded by the complementary space projection. Our analysis on single cell simulated and real bulk tissue data validated that such an elimination procedure can largely remove the expression of the genes stably expressed in cancer cells and keep the low rank structure of the gene expressions from other cells (See **Supplementary Note**).

Identification of rank-1 modules

Noting the prevalently existed dependency among cell types, such as immune cell co-infiltration in cancer microenvironment, we utilized our recently developed non-parametric co-expression module identification method namely MRHCA [18, 19]. Comparing to other gene co-expression network analysis methods, MRHCA is especially sensitive to small and strongly co-expressed gene modules that can identify hub genes of small modules with a hierarchical structure in a large module. The method evaluates if a hub gene is significantly ranked among top correlated ones by a group of highly co-expressed genes, which exactly fits the relationship between the gene mostly purely expressed by one cell type and the other genes majorly expressed by the cell type. The method is also with a rigorously derived statistical significance assessment formula and identifies strongly co-expressed modules rather than making hard partitions. More details about the MRHCA based module identification and its rationale are given in **Supplementary Note.**

A Bi-Cross Validation (BCV) based rank test is further applied to the modules to find the ones forming rank-1 in the whole matrix, which are possibly the genes uniquely expressed by one cell type. The matrix rank of a module centered by a cell type uniquely expressed genes always increases with the module size, due to the genes less co-expressed with the hub may be expressed by other cell types. In the analysis of this paper, we selected the modules of with hub significance p<1e-3, average co-expression coefficient>0.8, rank=1 (p<1e-3) and with at least eight genes, as possible markers of identifiable cell types.

Determine the number and select Rank-1 modules of “identifiable” cell types

Rather than a predefined cell types and lineage, de novo identification of “identifiable” cell types and corresponding markers may detect cell types of resolution levels over a tree-like cell type classification system [refs], such as the identified rank-1 markers can be specifically expressed by general myeloid cells, myeloid cell subtype macrophages and neutrophils, or even all MHC class II antigen presenting cells including macrophages, dendritic cell and other antigen presenting immune cell types. In some circumstance, the proportion of the cell type with a lower resolution is a non-negative linear sum of the proportion of several cell types with higher resolutions, such as the myeloid cell proportion equals to the sum of macrophage and neutrophils when these two cell types dominate the myeloid cell populations in the tissue. This linear dependency may correspond to a linear dependency between the row base of marker genes of cell types of different resolutions. Such a linear dependency may cause more identifiable cell types than the rank of the linear space generated by the identified rank-1 markers.

After identifying all sets of rank-1 marker genes, ICTD further determines the number of identifiable cell types, eliminates redundant and insignificant cell type marker genes, annotates each set of marker genes with a most likely cell type by using the labeling matrix, and build a marker gene – cell type representing matrix for the downstream deconvolution analysis.

To fully annotate the rank-1 markers, ICTD expands the labeling matrix constructed for each specific tissue microenvironment by including master cell types of closely related cell types. The following master cell types were considered in the analysis of immune and blood systems: general T cells expressing both CD8+ T and CD4+ T markers, TNK cell expressing T and NK cell markers, lymphoid cells expressing B, T and NK cell markers, antigen presenting immune cells expressing MHC class II genes, myeloid cells expressing both macrophage and monocytes markers, myeloid cells expressing macrophage, monocytes and neutrophil markers, and general glial cells expressing astrocyte, ependymal cell, oligodendrocyte, and microglial cells were considered in the analysis of brain data. In the expanded labeling matrix, the master cells were scored by the mean of the scores of its downstream cell types. For a rank-1 marker set $G_i = \{g_1, \dots, g_{n_i}\}$ and labeling matrix $L_{M\times K}$, we first compute $S_i = \{s_{i,1}, \dots, s_{i,K}\}$, where $s_{i,k} = \sum_{j=1}^{n_i} L_{g_j,k}$ representing the enrichment level of G_i to the genes top expressed in cell type k. The significance level of $s_{i,k}$, $p_{s_{i,k}}$, is assessed by a permutation test, and G_i is annotated as cell type with the minimal $p_{s_{i,k}}$ if $\min\left(FDR(p_{s_{i,k}})\right) < Cutoff_{CES}$. In this study, $Cutoff_{CES}$ is selected as 0.01. The rank-1 markers annotated without a significant cell type annotation are excluded from further analysis. It is noteworthy that a larger $Cutoff_{CES}$ can be selected for identification of possible unknown cell types.

To determine the number of identifiable cell types covered by the rank-1 marker genes, ICTD first construct a tree structure to represent the linear dependency among identify the rank-1 marker sets. A rank-1 marker set is considered as a root node if its row base can be non-negatively fitted by the row bases of other nodes with $R^{\mathbf{2}} > Cutoff_{R^2}$. In this study, $Cutoff_{R^2} = 0.9$ is selected. The rank-1 marker sets fit each other with $R^{\mathbf{2}} > Cutoff_{R^2}$ are merged together. All the root rank-1 marker sets are considered as markers of "identifiable" cell types and excluded from the further analysis. ICTD further computes the rank of the expression matrix of all the non-root rank-1 maker genes. Denoting the number of non-root rank-1 maker sets and their total rank as P and $\hat{P}$. The total number of "identifiable" cell types among the non-root rank-1 marker sets is determined as $\hat{P}$.

A marker gene – cell type representation matrix is further computed for the downstream NMF analysis. For a marker set a rank-1 marker set $G_i = \{g_1, \ldots, g_{n_i}\}, i = 1 \ldots P$, denote its gene expression profile as X_{G_i} and its SVD as $X_{G_i} = U_i \Sigma_i V_i^t$, G_i's self-explanation score is defined as $\frac{\sum_{g \in G_i} cor(X_g, V_i[,1])^2}{|G_i|}$, i.e. the averaged R square of the genes' expression fitted by their first row base. Our single cell simulated study suggests the rank-1 marker set with a larger self-explanation score is more like the genes uniquely expressed by one cell type. The marker gene – cell type representation matrix C is constructed by the following algorithm:

```
$for\ i\ in\ 1 \ldots P$
    $Compute\ the\ SVD\ of\ X_{G_i}\ as\ U_i \Sigma_i V_i^t$
$Conduct\ a\ hierachical\ clustering\ of\ G_i\ in\ to\ \hat{P}\ clusters\ C_j, i = 1 \ldots \hat{P}, by\ using\ euclie an\ distance\ between\ V_i[,1]$
$for\ j\ in\ 1 \ldots \hat{P}$
    $Select\ \ rank\ 1\ marker\ set\ G_{k_j}\ by\ \underset{j_k}{\operatorname{argmax}}(\frac{\sum_{g \in G_{j_k}} cor(X_g, V_{j_k}[,1])^2}{|G_{j_k}|} | G_{j_k} \in C_j)$
$C_{\sum_{j=1 \ldots \hat{P}} n_{j_k} \times \hat{P}}[i,j] = \{ \begin{matrix} 0, if\ \ gene\ i \notin G_{k_j} \\ 1, if\ \ gene\ i \in G_{k_j} \end{matrix}$
$return(C_{\sum_{j=1 \ldots \hat{P}} n_{j_k} \times \hat{P}})$
```

Noting this step assigns marker genes of identifiable cell types that highly determines the prediction accuracy of the deconvolution analysis. ICTD also includes three other options in constructing marker genes and C matrix of identifiable cell types. The computational details and performance comparison of these methods were given **in Supplementary Note.**

Constrained Non-negative Matrix Factorization

With the NMF constraint matrix $\mathrm{CS}_{\mathrm{X \times K}}^{\mathrm{NMF}}$, each of the K cell type is assigned with at least one cell type uniquely expressed gene (see derivations in method), hence constraint NMF problem $\mathrm{X_{M \times N} = S_{M \times K} \cdot P_{K \times N}, S[I,k] \geq 0, P[k,j] \geq 0, S[I,k] =}$ $0\ if\ \mathrm{CS^{NMF}[I,k] = 0}$ is with a unique solution [20]. The rationale here is that the analysis only focus on cell types with uniquely expressed markers that form rank-1 structure, and the analysis is robust to I/S cell co-infiltration due to the uniqueness of solution. Specifically, for the pth disconnected subgraph with $\mathrm{M_p}$ genes, rank= $\mathrm{K_p}$, and constraint matrix $\mathrm{C_{M_p \times K_p}}$, the NMF of $\mathrm{X_{M_p \times N} = S_{M_p \times K_p} \cdot P_{K_p \times N}}$ is solved by $\min_{\mathrm{S,P}}(\left\| \mathrm{X_{M_p \times N} - S_{M_p \times K_p} \cdot P_{K_p \times N}} \right\|_F^2 + \lambda \cdot \mathrm{tr(S_{M_p \times K_p}^T \cdot (I - C_{M_p \times K_p}))})$, where $\mathrm{S_{M_p \times K_p}}$ and $\mathrm{P_{K_p \times N}}$ are the predicted signature and proportion of the $\mathrm{K_p}$ I/S cell types (Figure 1X). Variables with fitted S that are highly varied from C are further removed. It is noteworthy when $\lambda \rightarrow \ \infty$, $P_{i,j}$ is the first row base of the SVD of $Diag(C_{.,j}) \cdot X$, where $Diag(C_{.,j})$ is the diagonal matrix generated by $C_{.,j}$. In this study, λ was selected based the best prediction accuracy trained on single cell simulated bulk data

Conditional local low rank test of cell type varied function

Identifiable cell type specific function is defined by a group of genes show local rank-1 structure conditional to the estimated proportion of the cell type. A kernal function based local low rank structure screening method is developed for identification of such local rank-1 structures. Denote $P_k = \{p_1^k, p_2^k, \dots, p_n^k\}$ as predicted proportion of cell type k for the n samples and $P_{(k)} = \{p_{k(1)}, \dots, p_{k(n)}\}$ as sorted P_k by increasing order, $\mathrm{G}_{\mathrm{I}_\mathrm{k}}$ as the rank-1 marker genes of cell type k, and $\mathrm{G}_{\mathrm{F}_k}$ is a gene set containing possible marker genes of a varied function of k, then the level of functional activity and its associated marker genes can be identified by the following local low rank testing algorithm:

```
Algorithm: BCV screening of a local low rank structure
For $\mathrm{i} = 1 \dots \mathrm{n} - \mathrm{l} + 1$
    Do BCV test of $\mathrm{X_i} \triangleq \mathrm{X}\left[\left(\mathrm{G_{I_k}}, \mathrm{G_{F_k}}\right), \mathrm{k(i)} \dots \mathrm{k(i+l)}\right]$
        $\mathrm{p_{ij}}$ = FDR correted p value of the rank j of $\mathrm{X_i}$
If $\exists\ \mathrm{i}^*$ and $\mathrm{j} > 1$,
s.t. $\mathrm{p_{ij}} < 0.05$ for all $\mathrm{i} \geq \mathrm{i}^*$ and $\mathrm{p_{ij}} \geq 0.05$ for all $\mathrm{i} < \mathrm{i}^*$
    $\rightarrow \mathrm{G_{F_k}}$ contains marker genes of a varied function
Identify markers of top $\mathrm{j} - 1$ rank in $\mathrm{X}\left[\mathrm{G_{F_k}}, \mathrm{k(i^*)} \dots \mathrm{k(n)}\right]$
```

The idea of this algorithm is that the genes of a cell type specific function may form additional ranks in the samples with high proportion of the cells, which can be identified by the BCV test when only screening on those samples. The kernel function is to smooth the screening boundary to ensure enough inter-sample variation in cell proportions (see more details in **Supplementary Note**).

In this paper, $\mathrm{G}_{\mathrm{F}_k}$ is selected for each cell type k by the genes annotated as top expressed by cell type k in the labeling matrix and with more than 0.8 co-expression correlation with the cell type k's proportion. ICTD enables users to predefine $\mathrm{G}_{\mathrm{F}_k}$ and select proportion of cell type k for a specified analysis, such as prediction of T cell cytotoxicity. The results of this analysis is a list of gene markers forming a local low rank structure in the subset of samples with high proportion of each cell type, which possibly corresponds to a varied functional activity of the cell type. The activity level are then predicted by the aforementioned NMF approach specifically in the samples with high proportion of the cell type.

Single cell simulated Bulk Tissue data

Bulk tissue data were simulated by using the scRNA-seq data of human head and neck (10), breast (12), lung(10), and colon cancer (8), melanoma (8), oligodendroglioma (13), astrocytoma (12), glioblastoma (5), two normal brain tissues (5 and 8), dendritic and myeloid cell (7), mouse melanoma (10) and hematopoietic system (4). The number after each data set indicates the number of cell types used for simulation. Simulation of cancer tissue included malignant cell types to mimic the true tumor microenvironment. Cell types in each scRNA-seq data were labeled by the cell clusters provided in the original works or by using Seurat pipeline with default parameters. Detailed information of the scRNA-seq data and cell type annotation is given in **Supplementary Table X and Notes**. For each data set, we simulate bulk tissue data with or without considering the dependency among cell types. The simulation of bulk tissue is composed by three steps: 1) generating the proportion of each cell type which called true proportion in this paper from a Dirichlet distribution, 2) if need, adding co-infiltrating for the two selected cell types with a given co-infiltrating level, 3) drawing cells randomly from the cell pool with replacement and combining them into a pseudo bulk tissue with pre-defined patients number, as detailed in the following pseudo code:

```
Input:
Single cell gene expression matrix B, which is M (genes) by N (single cell);
Single cell type label vector L, which correspond to N single cell;
Co-infiltration flag CoF, which is a bool value; Corr is the co-infiltration level parameter if CoF is True; row1, row2 are the cell type location that indicate two selected cell types adding CoF dependency;
Patient number n.
Function of Bulk_Simu_coinfil:
        1. Find the cell type number k based on L.
        2. Generate a k by n matrix D, which each row form a Dirichlet distribution.
        3. Judge the flag variable CoF, go to 4 if ture, else go to 7.
        4. Find the cell type string based on row1, row2 and locate the two vector v1, v2 in the matrix D.
        5. Generate two vectors, u1, u2, which satisfy the distance of the correlation of two vectors and Corr is
           less than $\delta, where\ \delta = 0.05$.
        6. Replace v1, v2 by u1, u2.
        7. Check the true proportion matrix satisfy required distribution and co-infiltration (if Cof is true).
        8. For i = 1, … , n:do
              i)    Sample single cells randomly from the cell pool with replacement for each cell type.
              ii)   Combine the gene expression value of selected single cells and normalize based on the
                    number of the selected cells.
Output:
The pseudo bulk tissue expression value matrix and true proportion D.
```

To generate the Dirichlet distribution matrix, R package “DirichletReg” (version 3.5.3) was used.

In order to evaluate the robustness of the deconvolution method while co-infiltrating existed, we simulated cancer tissue data of the five cancer types with setting the co-infiltration between the four pairs of cells that are commonly co-infiltrated in cancer tissue, namely, B/T cell, T/NK cell, Fibroblast/Endothelial cell, and B/Dendritic cell. (Supp fig xx). For a robust method evaluation, five replicates were generated in the simulation of each data set and co-infiltration parameter. The performance of ICTD including prediction accuracy of cell type and cell type specific function, as well as the prediction made by other methods, were assessed by the mean of achieved on the five data sets.

Explanation Score to evaluate the performance of our deconvolution method

We assessed the methods’ performance by the correlation between predicted and known proportion of each cell type in simulated data, which is inapplicable in the real tissue data. Thus, we developed an explanation score (ES) of the goodness that each marker gene’s expression is fitted by the predicted cell proportions, with the following formula

$$EScore(x) = 1 - \sum_{j=1}^{N} (x_j^* - \hat{x}_j)^2 \,/\, \sum_{j=1}^{N} (x_j^*)^2$$

$$\widehat{x_j} = \sum_{k=1}^{k_x} \beta_k^x p_j^k \, , \beta_k^x \geq 0$$

, where x_j^* is the observed expression of marker gene x in sample j, $\hat{x}_j$ is the x's expression level in j predicted by a non-linear regression model of the predicted proportion $p_j^k, k = 1 \ldots k_x$ of k_x cell types that express x, and β_k^x are parameters. Intuitively, with correctly selected marker genes, the marker gene's expression can be well explained by the predicted proportions of the cell types that express the gene. Hence, a high ES score is a necessary but not sufficient condition for correctly selected marker genes and predicted cell proportion.

References

1. Hackl, Hubert., et al., *Computational genomics tools for dissecting tumour–immune cell interactions*. Nat Reviews Genetics, 2016. **17**(8): p. 441.

2. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles.* Nat Methods, 2015. **12**(5): p. 453-7.

3. Li, B., et al., *Comprehensive analyses of tumor immunity: implications for cancer immunotherapy.* Genome Biol, 2016. **17**(1): p. 174.

4. Racle, J., et al., *Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data.* Elife, 2017. **6**.

5. Frishberg, Amit., et al., *Cell composition analysis of bulk genomics using single-cell data.* Nat methods, 2019. **16**(4): p. 327.

6. Newman, Aaron M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry*. Nat biotechnology, 2019.**37**(7): p. 773-782.

7. Varn, F.S., et al., *Systematic Pan-Cancer Analysis Reveals Immune Cell Interactions in the Tumor Microenvironment*. Cancer Res, 2017. **77**(6): p. 1271-1282.

8. Li, B., J.S. Liu, and X.S. Liu, *Revisit linear regression-based deconvolution methods for tumor gene expression data*. Genome Biol, 2017. **18**(1): p. 127.

9. Dormann, Carsten F., et al., *Collinearity: a review of methods to deal with it and a simulation study evaluating their performance*. Ecography, 2013. **36**(1): p.27-46.

10. Art B. Owen, P.O.P., *Bi-cross-validation of the SVD and the nonnegative matrix factorization*. Annals of Applied Statistics 2009. **3**(2): p. 564-594.

11. Gaujoux, R. and C. Seoighe*, CellMix: a comprehensive toolbox for gene expression deconvolution*. Bioinformatics, 2013. **29**(17): p. 2211-2.

12. Joonseok Lee., et al., *LLORMA: Local Low-Rank Matrix Approximation*. Journal of Machine Learning Research, 2016. **17**: p. 1–24.

13. Saltz, J., et al., *Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images*. Cell Rep, 2018. **23**(1): p. 181-193 e7.

14. Venteicher, A.S., et al., *Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq*. Science, 2017. **355**(6332).

15. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq*. Science, 2016. **352**(6282): p. 189-96.

16. Gabrilovich, Dmitry I., et al., *Myeloid-derived suppressor cells as regulators of the immune system*. Nat reviews immunology, 2009. **9**(3) : p. 162.

17. Johnson, W.E., et al., *Adjusting batch effects in microarray expression data using empirical Bayes methods*. Biostatistics, 2007. **8**(1): p. 118-27.

18. Yu Zhang, S.C., Jing Zhao, Burair Alsaihati, Qin Ma, Chi Zhang, *MRHCA: a nonparametric statistics based method for hub and co-expression module identification in large gene co-expression network*. Quant. Biol., 2018. **6**(1): p. 40-55.

19. Zhang, C., et al., *Elucidation of drivers of high-level production of lactates throughout a cancer development.* J Mol Cell Biol, 2015. **7**(3): p. 267-79.

20. Kejun Huang, et al., *Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition*. IEEE Transactions on Signal Processing 2014. **62**(1).