ICTD: Inference of cell types and deconvolution -- a next generation deconvolution method for accurate assess cell population and activities in tumor microenvironment.

Wennan Chang[1, 2], Changlin Wan[1, 2], Yu Zhang[3], Kaman So[1], Brooke Richardson[1], Yifan Sun[1], Xinna Zhang[1], Kun Huang[1,4], Anru Zhang[6], Xiongbin Lu[1*], Sha Cao[1]*, 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, Indiana University, School of Medicine, Indianapolis, IN,46202, USA.

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

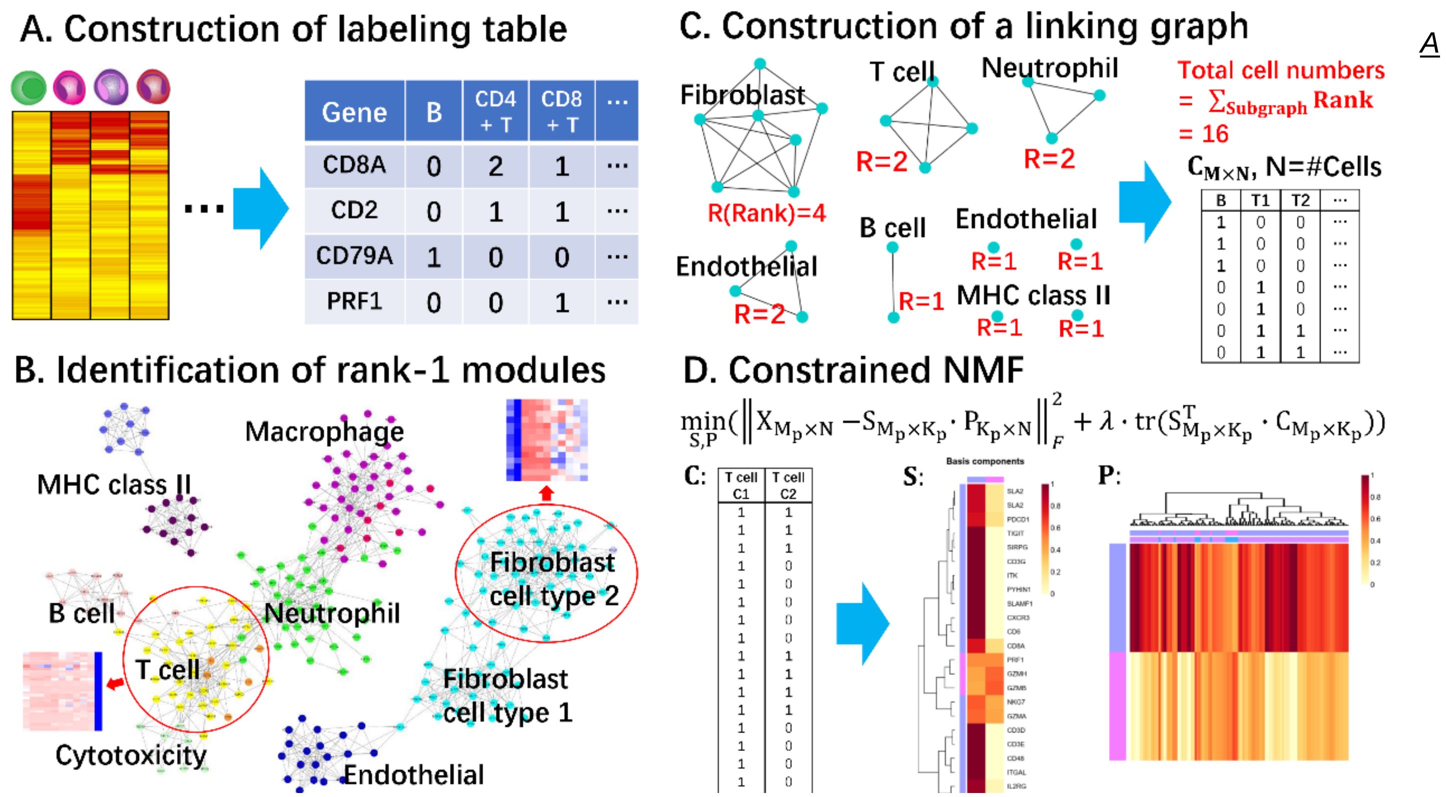
[3]Colleges of Computer Science and Technology, Jilin University, Changchun,130012, China,

[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@sdstate.edu.

Tumor tissue consists of various cell types, and the study of such complexity using transcriptomic data has been mathematically formulated as a linear equation: $X_{M\times N} = S_{M\times K} \cdot P_{K\times N} + E$, where X is the observed expression profiles of M genes and N tissue samples with K cell types, S and P are the gene signature and proportion matrices of the K cell types, including immune/stromal (I/S) cells, and E stands for error. Existing methods such as CIBERSORT, TIMER and EPIC [1-3] assume a constant signature matrix S, which are obtained from independent expression data sets with pure I/S cell, and the prediction of P is accomplished through regression methods. However, these supervised methods allow for minimum dynamics of S, and thus result in biased estimations of P, due to ignorance of the following two facts: (1) the gene expression of I/S cells in cancer microenvironment can be largely varied from what is obtained from the training data, and (2) the intrinsic batch effect between different studies cause the application of signature matrix from training data on a new dataset to be inappropriate, even with quantile normalization as in CIBERSORT [1]. In fact, recent studies revealed that these regression-based methods can only achieve very limited R^2 on TCGA and other cancer data sets [1, 3, 4]. In light of this, we develop a new deconvolution approach, where we allow for a high level of dynamics in signature matrix S, except for retaining its structural composition learnt from large amount of training data sets as well as the tissue expression data itself. The unbiasedly estimating proportion of I/S cells could then be reliably correlated with features derived from imaging, genomic and clinical data of these patients.

Key challenges for an unbiased deconvolution approach include: (i) detecting the I/S cell types and their true marker genes specific to the current cancer tissue micro-environment, (ii) eliminating variations of gene expression caused by different experimental platforms and batches, and (iii) dealing with the prevalent co-infiltrated I/S cells [5]. Among these challenges, (i) and (ii) require the signature matrix S to be highly dynamic to the current cancer microenvironment, as well as experimental setting variations; and (iii) complicates the problem as co-infiltration of I/S cells lead to high correlation among their proportion, which will jeopardize the uniqueness of the solution S and P, and the uniqueness of the solution could not be guaranteed unless cell type uniquely expressed genes exist among these co-infiltrating I/S cells [6].



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

$C_{\mathbf{M}\times\mathbf{N}}$, N=#Cells

B	T1	T2	···
1	0	0	···
1	0	0	···
1	0	0	···
0	1	0	···
0	1	0	···
0	1	1	···
0	1	1	···

$$\min_{S,P}\left(\left\|X_{M_p\times N}-S_{M_p\times K_p}\cdot P_{K_p\times N}\right\|_F^2+\lambda\cdot \mathrm{tr}\left(S^T_{M_p\times K_p}\cdot C_{M_p\times K_p}\right)\right)$$

C:

T cell C1	T cell C2
1	1
1	1
1	1
1	0
1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1
1	0
1	0
1	0
1	0

Figure 1. ICTD Analysis Pipeline

new computational method to unbiasedly predict I/S cells in cancer transcriptomics data. We develop a semi-supervised deconvolution method to handle the three challenges, with the following four steps: (1) Construct an ordering matrix to represent cell type specific gene expressions. We have collected large amount of gene expression data for fibroblast, adipocytes, endothelial, neuron, B cell, CD4+ T cell, CD8+ T cell, natural killer cell, dendritic cell, monocytes, macrophages, and neutrophils, all measured by Affymetrix U133 plus 2.0 array, totaling 387, 26, 606, 243, 404, 443, 130, 141, 410, 477, 277, 257 samples, respectively. We utilized COMBAT [7] to remove the batch effect across datasets. Using these as training data, we constructed an ordering matrix by comparing a gene expression's expression among different cell type. The ordering matrix $L_{m\times k}$ is takes value in the set $\{1,2,3,\dots,K-1\}$, $L_{\mathrm{i,j}}=l$ 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 i is not overly expressed in cell type j (Figure XA). (2) Predict cell type uniquely expressed genes in a cancer transcriptomic dataset via identification of rank-1 submatrices. If gene i is uniquely 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 N 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. To find such rank-1 submatrices in the whole matrix, we particularly realize that the level of co-expressions among a certain cell type's uniquely expressed genes are generally higher than their co-expression correlations with other genes, thus, we turn to find those strong co-expression modules using our in-house non-parametric network analysis method MRHCA [8, 9], wherein the rank of the module could be determined by Bi-Cross Validation (BCV) rank test [10](Figure XB). Expression of the genes coming from cancer cells will be eliminated by projecting $\mathrm{X}_{\mathrm{M}\times\mathrm{N}}$ to the complementary space of the row space spanned by the cancer genes' modules, under which the rank structure of the I/S gene expressions will not change. Further, cell type of each identified rank-1 module will be annotated by the labeling matrix. (3) Infer the number of cell types and their expressed genes via a graph partition method. Genes in rank-1 modules identified in (2) can be either genes uniquely expressed in one cell type or genes with similar expression patterns in multiple cell types. To determine the number of cell types, we sequentially examine the combinations of 2, 3, and 4 rank-1 modules and re-calculate the rank of the newly combined modules. The searched modules will be linked to each other if the rank of their merged expression matrix is smaller than the number of unconnected parts among them. Such a process will link the rank-1 modules of genes expressed in multiple cell types with the modules of uniquely expressed genes of the cell types and eliminate the redundant rank of such modules. Then the total number of cell types covered by the modules is determined by the sum of the rank of each disconnected subgraph in the linking graph (Figure XC). I/S cell types correspond to each disconnected subgraph will be annotated by the labeling matrix. (4) Predict cell type proportions using constrained NMF. We recently developed an algorithm to identify cell type unique modules in each disconnected subgraph, by which 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 the pth disconnected subgraph with M_{p} genes and rank= K_{p}, we first identify the top K_{p} cell type unique modules in the subgraph, and construct $\mathrm{C}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{K}_{\mathrm{p}}}$ by $\mathrm{C}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{K}_{\mathrm{p}}}[i,j] = 1/0$, if gene i is in or not in jth module. Then the NMF of $\mathrm{X}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{N}} = \mathrm{S}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{K}_{\mathrm{p}}} \cdot \mathrm{P}_{\mathrm{K}_{\mathrm{p}}\times\mathrm{N}}$ will be solve by $\min_{\mathrm{S,P}}\left(\left\|\mathrm{X}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{N}} - \mathrm{S}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{K}_{\mathrm{p}}} \cdot \mathrm{P}_{\mathrm{K}_{\mathrm{p}}\times\mathrm{N}}\right\|_F^2 + \lambda \cdot \mathrm{tr}(\mathrm{S}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{K}_{\mathrm{p}}}^{\mathrm{T}} \cdot \mathrm{C}_{\mathrm{M}_{\mathrm{p}}\times\mathrm{K}_{\mathrm{p}}})\right)$.

The advantages of this deconvolution method includes: (1) cell types and their uniquely expressed genes, instead of pre-defined, are specifically identified to the cancer microenvironment hidden in the dataset; (2) I/S proportions could be estimated free of batch effect by a NMF approach; (3) I/S co-infiltrations could be properly handled by the existence of cell type uniquely expressed genes that place the co-infiltrated cell types in disconnected subgraphs; and (4) commonly expressed genes in multiple cell types could be properly incorporated in a constrained NMF without harming the uniqueness of the solution. It is noteworthy the cell types are defined here as a combination of certain rank-1 markers, which may correspond to more than one commonly defined cell types by hematopoiesis lineage, namely, general T cell, total T and B cell, MHC class II antigen presenting cells, etc. This novel definition is entirely data-driven, without compromising the capacity of performing correlation analysis with features derived from genomics and imaging data, and most importantly, novel I/S sub cell types specific to a cancer micro-environment could be conveniently identified.

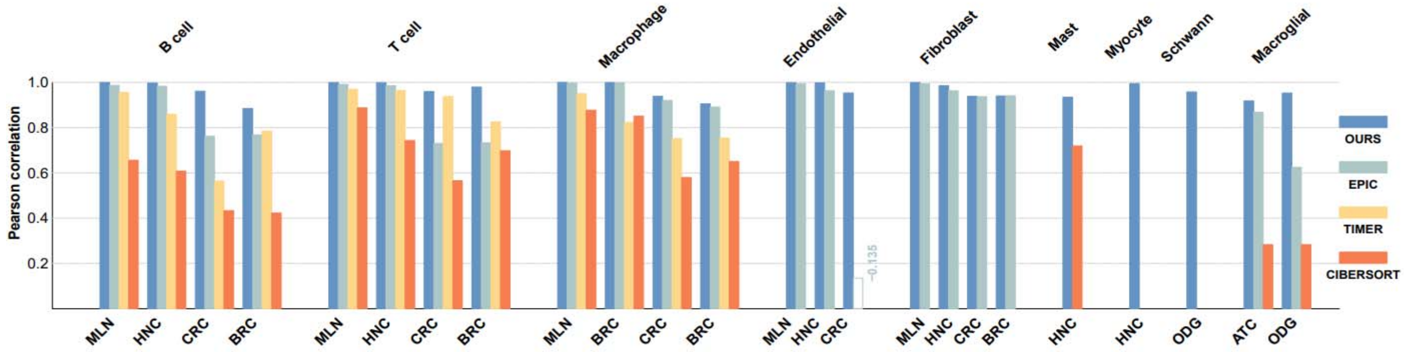


Figure 2. Comparison of our method with CIBERCORT, TIMER, and EPIC on scRNA-seq simulated bulk cancer data.

We have validated the method on six scRNA-seq simulated bulk tumor data sets of melanoma (MLN), head and neck cancer (HNC), colorectal cancer (CRC), breast cancer (BRC), glioma (ODG) and IDH-mutant astrocytoma (ATC) [12-14]. On average, our method achieves 0.92 correlation between the predicted and true cell proportions, and 0.958 R^2 in fitting the I/S uniquely expressed genes. Overall, ICTD outperformed the prediction by EPIC, TIMER and CIBEROSRT. Specifically, EPIC achieved a good performance in prediction of macrophage and fibroblast cells, but the predicted B cell and T cell proportion has only a ~0.85 correlation with the true proportion, especially in the colorectal and breast cancer set. Timer has a good performance in prediction of T cell proportional but have a relative less prediction accuracy for B and Macrophage cells. Meanwhile, CIBERSORT has achieved less than 0.6 correlation between the predicted and true cell proportions, and only 0.45 R^2 in the fitting of marker genes (Figure 2). In addition, the method can effectively identify novel sub cell types within the pre-identified cell classes, such as sub fibroblast and myeloid cell types, as well as the cell types not covered by the training data, namely Mast and Myocyte cell in the HNC, Schwann cells in the ODG and Microglial cells in the ATC and ODG data.

We applied our deconvolution method to two TNBC RNA-Seq data sets, one in TCGA and another in IUSM, and six microarray data sets retrieved from public domain, totaling 851 samples. Our analysis identified 29, 12, 16, 36, 8, 32, 9, and 12 uniquely expressed marker genes corresponding to T-, B-, macrophage, neutrophil, dendritic, fibroblast, adipocyte, and endothelial cells respectively that are consistent in all data sets (consistent markers are shown in Table 1). On average, our method achieves 0.87 by feeding these consistent markers to our NMF algorithm, comparing to 0.32 by CIBERSORT which uses SVM regression analysis based on a set of pre-defined genes. In addition, TCGA cohort has morphological data derived lymphocyte infiltration levels for all samples with pathological images, and our predicted T cell infiltration level has a Spearman correlation of 0.53 with the imaging-derived lymphocyte infiltration level in TNBC data, while the correlation is 0.3 for CIBERSORT predicted TILs [15]; We expect the analysis on the large amount of TNBC samples will generate robust gene markers and infiltration level of TILs.

T cell: CD2, CD3D, CD3G, CD52, GZMA, …
B cell: CD79A, CD79B, MZB1, FCRL2, IGL, …
Macrophage: TFEC, LAIR1, CD84, CD163, …
Neutrophil: CD48, CD53, IL10RA, LCP2, SLA,…
Fibroblast: COL1A2, COL3A1, COL5A1, FAP,…
Adipocyte: ADIPOQ, ADH1B, ACACB, FHL1,…

Table 1. Selected cell type uniquely expressed genes in the microenvironment of TNBC

ICTD is also applied to TCGA colorectal and breast cancer to compare with EPIC, TIMER and CIBERSORT on real cancer data. Specifically, we developed a goodness of fitting score for each gene by the of a non-negative regression of the gene expression by the predicted cell proportions. The advantage of this score is that the bias of the gene expression signature of different methods are eliminated in the analysis, and a pure evaluation of the predicted immune cell proportion enables the comparison between our semi-supervised method with the three supervised methods. The goodness of fitting clearly suggested the proportions predicted by ICTD significantly better explains the expression of each gene (with an averaged >0.6) than the other three methods.

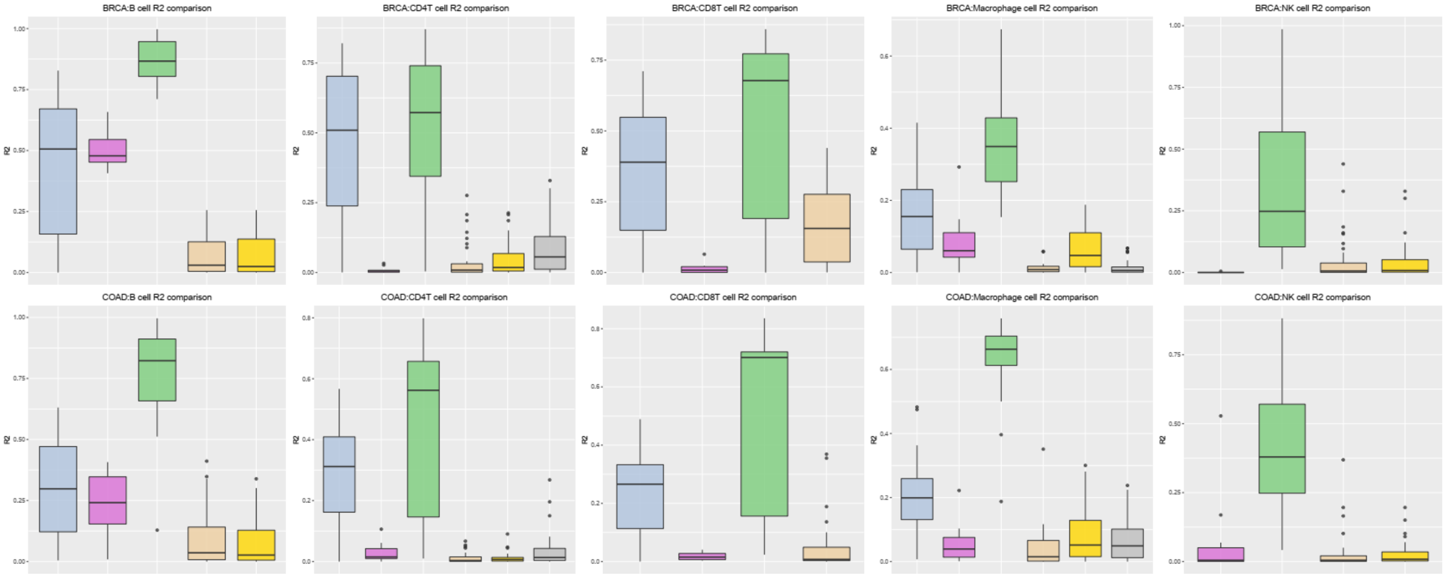


Figure 3. Comparison of the goodness of fitting of ICTD (green), EPIC (blue), TIMER (pink), and CIBERSORT (other colors) in the marker genes of B, CD4+T, CD8+ T, Myeloid, and Fibroblast markers.

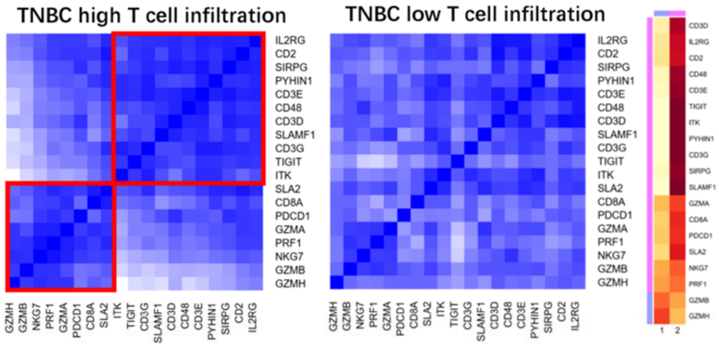


Figure 4. T cell (upper right) and cytotoxicity (bottom left) markers form two rank-1 structure in high T cell infiltration samples and one rank-1 structure in low T cell infiltration samples. The left and middle figures show the co-expression correlation between the genes and the figure on the right shows the predicted gene expression signature of cytotoxicity and T cell infiltration.

Quantitative prediction of TIL cytotoxicity and other varied immuno-activities. Our deconvolution method not only investigates biological I/S cell types, but also cells' functional activities, such as T cell cytotoxicity, T cell exhaustion or production of oxidative stress by myeloid derived suppressor cells, these immune functions are highly varied in specific cancer micro-environments, thus it is not feasible to use constant gene expressions to characterize their activities. Note that genes involved in such functions also form a 1- or low rank submatrix and our recent analysis on TNBC samples indicates that in samples with high T cell infiltration, the T cell cytotoxicity and infiltration marker genes form two distinct 1-rank modules, while in samples with low T cell infiltration, they only form one 1-rank module (Figure 3). Clearly, varied cytotoxicity in samples with high T cell infiltration is of more interest, hence we propose to identify functional activity levels by detecting the 1- or low rank structure in samples with high infiltration levels, i.e. a local low rank structure [16]. Denote as predicted proportion of cell type k for the n samples and as sorted by increasing order, as the rank-1 marker genes of cell type k, and is a gene set containing 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 algorithm using BCV tests. The expected results of this analysis include gene markers that form a local low rank structure specifically in the samples with high infiltration of an I/S cell type, which corresponds to a varied immune functional activity, and the activity level can be predicted by the aforementioned NMF approach.

Algorithm: BCV screening of a local low rank structure

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

References

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

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

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

4. 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.

5. 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.

6. Kejun Huang, N.D.S., Ananthram Swami, *Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition.* IEEE Transactions on Signal Processing 2014. **62**(1).

7. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

8. 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.

9. 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.

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

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

14. Puram, S.V., et al., *Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer.* Cell, 2017. **171**(7): p. 1611-1624 e24.

15. 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.

16. Joonseok Lee, S.K., Guy Lebanon, Yoram Singer, Samy Bengio, *LLORMA: Local Low-Rank Matrix Approximation.* Journal of Machine Learning Research, 2016. **17**: p. 1–24.