Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jinjing He Date

Estimating and Comparing immune cell infiltration in cancer

By

Jinjing He

Master of Science in Public Health

Biostatistics and Bioinformatics

_____

Hao Wu, PhD

(Thesis Advisor)

_____

Zhaohui (Steve) Qin, PhD

(Reader)

Estimating and Comparing immune cell infiltration in cancer

By

Jinjing He

B.A.

China Pharmaceutical University

2018

Thesis Committee Chair: Hao Wu, PhD

Reader: Zhaohui (Steve) Qin, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University in partial fulfillment of the requirements for the degree of Master of Public Health

In Biostatistics and Bioinformatics Department

2020

Abstract

Estimating and Comparing immune cell infiltration in cancer

By Jinjing He

Background: During last several years, cancer immunology and immunotherapy have become a promising field in cancer research. It is known that the tumor-infiltrating immune cells are related to tumor progression. In particular, the proportions of immune cells in tumor samples are often indicators for cancer stages and predictive for survival rate, thus these quantities are of great interest. To estimate the immune cell proportions, a practical approach is to perform signal deconvolution from high-throughput omics data. In this study, we aim to compare the immune cell proportions estimation from different omics data, using different method. To be specific, we apply two deconvolution methods on gene expression and DNA methylation data from individuals with and without breast invasive carcinoma (BRCA).

Methods: We used CIBERSORT and TOAST methods to estimate the proportions of B cells, CD4 T cell, CD8 T cell, Natural killer (NK) cell, Monocytes and Granulocytes in gene expression and DNA methylation. CIBERSORT is based on regression framework for reference-based deconvolution, and TOAST is based on matrix factorization for reference-free deconvolution. The Pearson correlation was applied to assess the relationship of estimates from gene expression and DNA methylation with the same method, or between the estimates with two methods.

Conclusion: Results from CIBERSORT showed that there is no significant difference in estimated proportions from tumor and normal samples. The estimations from gene expression and DNA methylation data are very low. With TOAST, the estimated proportions in tumor samples are significantly different from those in normal samples. Moreover, the estimated immune cells proportions from gene expression and DNA methylation show some correlation. Comparing CIBERSORT and TOAST, the correlations between proportion estimates from both gene expression and DNA methylation data are not very high.

KEYWORDS: Cancer, immunology, immune cells, gene expression, DNA methylation

Introduction

Immune system has the ability to protect the body from pathogen, allergen or toxin including the tumor cells that produced by body itself.[4] And many researchers have found that immune system is able to suppress the progression of tumors, thus many immunotherapeutic approaches are developed to activate immune effector or neutralize inhibitory.[12] For example, one research has reported that using siRNA to express the immunosuppressive enzyme indoleamine 2,3-dioxygenase (IDO) silence in dendritic cells can have a positive impact on cancer.[13] However, even if the immune system is important for us to defend the tumor cells in our body, we still cannot figure out the exact mechanism of immune interaction between tumor and immune cells. Hence, there is a great need for us to understand the relationship between immune cells and tumors.

Among all types of immune cells, the tumor-infiltrating immune cells can be a link between immune response and tumor development.[5] The tumor-infiltrating immune cell is a type of immune cells that some immune cells infiltrate into tissues in tumor and contact between tumor cells directly.[8] They are also related with the growth of tumor cells.[8] Some studies have reported the relationship between prognosis and infiltrating immune cells. [6] For example, T-cell correlates well with colorectal cancer.[10] Other immune cell types including, B cells, natural killer (NK) cells have also been shown to have the association with the prognosis of patients in tumor.[1] However, less is known about how tumor-infiltrating immune cells all together will predict the cancer prognosis. The mathematical model is needed to have a better estimation to the association between immune cells and cancer prognosis. For now, some mathematical methods have a good

result. One research has developed a method called Microenvironment Cell Population (MCP)-counter that allows to quantify the absolute abundance of eight immune and two stromal cell populations in heterogeneous tissues from transcriptomic data.[2] They used MCO-counter to estimates abundance score for CD3+ T cells, CD8+ T cells, B cells, cytotoxic lymphocytes, NK cells, as well as endothelial cells and fibroblasts in each sample, and used scores to compare the abundance of the cell types across samples with a cohort.[2] To validate this method, mRNA mixtures and immunohistochemical cell quantifications on paraffin-embedded tissue sections were tested.[2] Another method is experimental and visual. [3] They applied tissue microarrays (TMAs) and quantitative real-time PCR (qPCR) to evaluate the immune reaction in the location of center (CT) and invasive margin (IM) of the tumor, and to test its changes as a function for tumor stage in colorectal carcinoma.[3] Those methods suggest that the interaction of immune cells and tumor prognosis can be estimated. And more and more new methods should be developed to have a better estimation of that relationship.

In our study, we try to apply the mathematical model to figure out the relationship between the proportion of tumor-infiltrating immune cells and cancer prognosis. We aim at using the model to find out whether the proportions of tumor-infiltrating immune cells would be different between patients with breast invasive carcinoma (BRCA) and healthy people. In addition, we aim to calculate the correlation between gene expression and DNA methylation to find out their relationship.

Method

CIBERSORT and TOAST methods were applied in this study to produce the outcomes. We used CIBERSORT[9], a method with deconvolution that could characterize cell compositions from the genomic expression profiles, to produce "signature matrix" for gene expression and DNA methylation profiles with their corresponding reference cell-specific datasets. The "signature matrix" was the matrix of estimated coefficients generated with CIBERSORT method. The TOAST[7] method was a reference-free deconvolution method and was based on the iterative algorithm that could be iterated with features selected in the formal iteration to produce a mixture proportion from mixed samples in high-throughput datasets. After obtaining the "signature matrix" and mixture proportion, we applied Pearson correlation method to detect the correlation between the gene expression and DNA methylation. In addition, a comparison of two deconvolution methods was also conducted with Pearson correlation method.

Data Preparations

For gene expression, RNA sequencing (RNA_Seq) and corresponding reference cell-specific datasets for Breast invasive carcinoma (BRCA) were obtained from The Cancer Genome Atlas (TCGA). The cell types we chose as the reference were B cells, CD4 T cell, CD8 T cell, Natural killer (NK) cell, Monocytes, and Granulocytes. The RNA_Seq dataset had patients as columns and gene symbols as rows. The reference dataset regarded cell types as columns and gene symbols as rows. The scale of RNA_Seq dataset was raw count which was simply the number of reads overlapping a given feature.

For DNA methylation, DNA methylation (DNA_mthyl) and corresponding reference cell-specific datasets for BRCA were acquired from Illumina 450K array. We also

selected 6 cell types that were the same as those in gene expression reference cell-specific dataset. The DNA_mthyl dataset contained patients as columns and probe names as rows. And the corresponding reference dataset had the cell types as columns and probe names as rows.

CIBERSORT

The CIBERSORT deconvolution algorithm was reference-based, and had been developed to solve the following linear equation for X:

$$Y = X'\beta + b$$

Y: the vector containing input (RNA_Seq or DNA_mthyl dataset).

X: the vector containing each cell fraction (reference cell-specific dataset).

β: the "signature matrix" containing estimated coefficients for each cell type in each sample.

The equation illustrates that CIBERORT aimed at finding the "signature matrix" — the matrix of cell-specific expression signatures. The process of producing a signature matrix needed a preprocessing step that removed some features that are irrelevant before the application of the CIBERSORT method. We filtered the marker genes before processing that. We used the *limma* package in R to find the genes that have differential expression levels.

CIBERSORT method was developed with the machine learning method. After preprocessing, nu-support vector regression (ν-SVR)[11] was implied to minimize a loss function and penalty function. The ν -SVR used the same principles as the support vector

machine (SVM) for classification. To find the Y, it used criteria with the minimal norm value of $||w||^2$:

$$J(\beta) = \min_{\mathrm{w}} \left|\left|w\right|\right|^2$$

subject to all residuals having a value less than ϵ;

$$\forall n: |y_n - (X'\beta + b)| \leq \epsilon$$

The linear ϵ-insensitive loss function ignored errors that were within ϵ distance of the observed value by treating them as equal to zero. The loss was measured based on the distance between observed value y and the ε boundary. The formula was as follows:

$$L_\epsilon = \begin{cases} 0\,, & if\ |y_n - (X'\beta + b)| \leq \epsilon \\ |y_n - (X'\beta + b)| - \epsilon, & otherwise \end{cases}$$

With the constraint, it defined a hyperplane that could capture the points as many as possible in that area. In addition, it reduced overfitting by only penalizing data points outside a certain error radius using a linear "epsilon-insensitive" loss function. In CIBERSORT, ν represented the ϵ, and the set of ν was (0.25, 0.5, 0.75). The data points within the distance ν were ignored and outside of the distance ν were evaluated. The current implementation of CIBERSORT was to use "svm" function in the e1071 Package in R.

TOAST

The TOAST[7] deconvolution was a reference free deconvolution method to solve the linear equation mentioned above. It assumed that we had the K mixing proportion experimentally or computationally obtained. The mixing proportion for each sample i was $\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})$, And the constraint of this was $\sum_K \beta_{ik} = 1$. With the initial

mixing proportion, we used cross-cell type differential analysis to select the features. And the features were then applied for the next iteration to produce the new mixing proportion. After setting the specific times of iterations, we would get the best mixing proportions. As for the selecting features, we first assumed that the p-th feature for observed data was $Y_p = [Y_{p1}, Y_{p2}, \dots, Y_{pN}]^T$.With the obtained mixing proportion, we could model the observed data with linear equation:

$$E(Y_p) = \beta\theta$$

$$\beta = \begin{bmatrix} \beta_{11} & \cdots & \beta_{NK} \\ \vdots & \ddots & \vdots \\ \beta_{N1} & \cdots & \beta_{NK} \end{bmatrix}, \theta = \begin{bmatrix} \mu_{p1} \\ \mu_{p2} \\ \vdots \\ \mu_{pk} \end{bmatrix}$$

The μ_{pk} was the mean level of the p-th feature in the k-th cell type. Then we conducted cross-cell type differential analysis with the following hypothesis:

$$H_0: \mu_{pk} - \frac{1}{K-1}\sum_{i \neq k} \mu_{pi} = 0, k = 1, \dots, K$$

The features with the significant results were the features we would like to select.

Result

Results from CIBERSORT

152 samples were enrolled in this study among which were 76 tumor samples and 76 normal samples. CIBERSORT, the reference-based regression method, was applied to produce the “signature matrix” in tumor and normal samples. The value in the “signature matrix” ranged from 0 and 1 (Figure 1 and 2). The estimated coefficient of Monocytes

had the highest value that was about 0.6, while the values in CD8 T cells, CD56 NK cells and Granulocytes were close to 0 in gene expression (Figure 1). For DNA methylation, the estimated coefficient values of CD8 T cells and Monocytes were close to 0 (Figure 1). And the ranges of estimated coefficient values in CD4 T cells and CD56 NK cells were from 0.2 to 0.6, but the median value of CD 56 NK cells was higher than that of CD4 T cells. The median results of estimated coefficients in 6 cell types were similar in tumor and normal samples (Figure 1 and Figure 2).

The Pearson correlation method was used to calculate the correlation in the proportion estimates between gene expression and DNA methylation. Since we would like to figure out the impact of correlation towards the tumor, we focused on the relationship in the tumor sample. The ranges of correlation were from 0.159 to -0.167, which showed that there was no significant correlation between the proportion estimates of gene expression and methylation (Table 1 and S1, Figure S1). Moreover, based on the estimates from CIBERSORT method, the cell proportions were no significant difference for people with and without cancer in gene expression and DNA methylation (Figure 1 and 2, Table S5 and S6).

Table 1 The correlation matrix between gene expression and DNA methylation in tumor samples with CIBERSORT

	mCD4T	mCD8T	mBCell	mMono	mGran	mCD56NK
rBCell	-0.142	0.018	-0.011	-0.098	0.101	0.038
rCD8T	0.083	0.048	-0.029	-0.012	-0.015	-0.081
rCD4T	-0.095	-0.101	-0.019	0.018	0.068	0.122
rCD56NK	-0.138	0.118	0.113	-0.051	0.124	-0.122
rMono	0.159	-0.013	0.031	0.053	-0.125	-0.047
rGran	-0.038	0.262	-0.167	0.414	0.045	-0.167

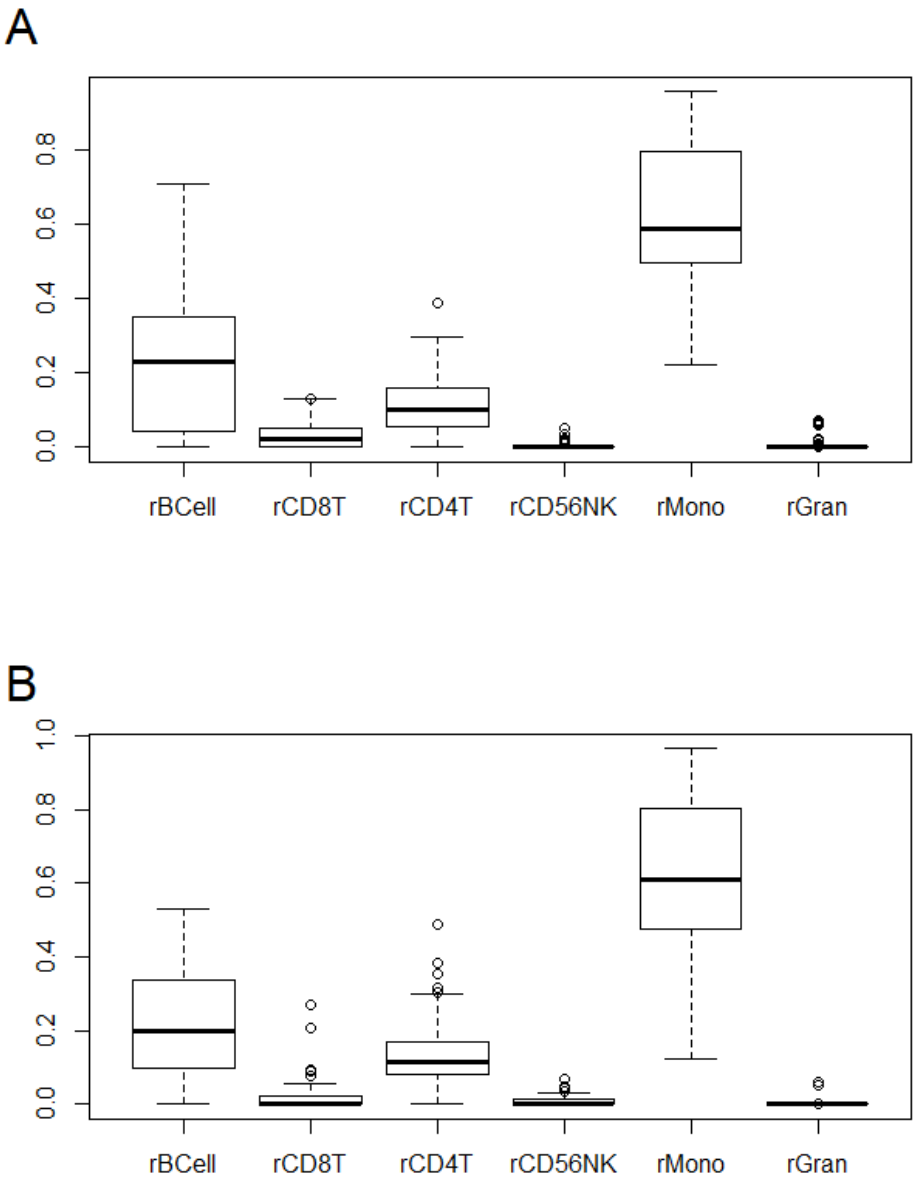

Figure 1. The summary of proportion estimation from gene expression with CIBERSORT. A shows the result from tumor samples. B shows the results from normal samples.

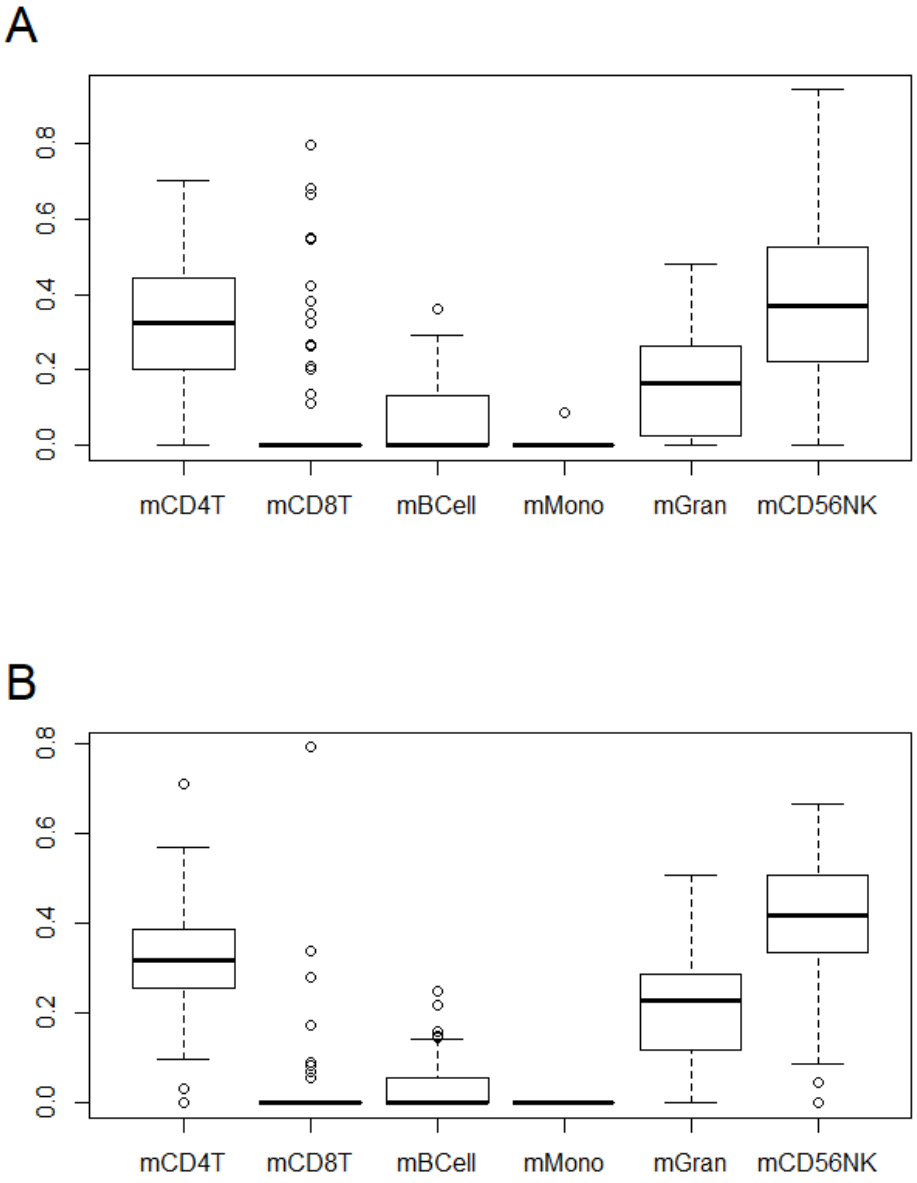

Figure 2. The summary of proportion estimation from DNA methylation with CIBERSORT. A shows the result in tumor samples. B shows the results in normal samples.

Results from TOAST

The sample was the same as that used in the CIBERSORT method. Here, we used TOAST, a reference-free deconvolution method, to produce the estimated proportions, and used Pearson correlation method to compare the relationship between gene expression and DNA methylation. We determined K=6 in this study. The 6 proportion estimates ranged from 0 to 1 (Figure 3 and 4). Six proportion estimations of tumor samples in gene expression were different from those of normal samples. The median of the highest value was 0.5 in tumor sample while 0.4 in the normal sample (Figure 3). The median of the first estimated proportion (r1) was close to 0.2 in the normal sample while 0 in tumor sample (Figure 3). But both of the third estimated proportions (r3) in two groups were around 0 (Figure 3). In DNA methylation, the median of 6 estimated proportions in tumor samples ranged from 0 to 0.4 while from 0 to 0.6 in normal samples (Figure 4). The variation of 6 median values in normal samples was larger than those in tumor samples.

The correlation values ranged from -0.4 to 0.6 (Table 2). There were some positive correlations in tumor samples between r1 and m1, r2 and m2, r3 and m3,respectively (Table 2 and S2, Figure S2). The “r” represents that the estimated proportions were produced in gene expression, and the “m” represents that the estimated proportions were produced in DNA methylation. Most of the estimated proportion in gene expression was correlated with estimated proportions in DNA methylation. And the proportion could be significantly different in tumor samples compared to those in normal samples in gene expression and DNA methylation (Figure 3 and 4, Table S7 and S8).

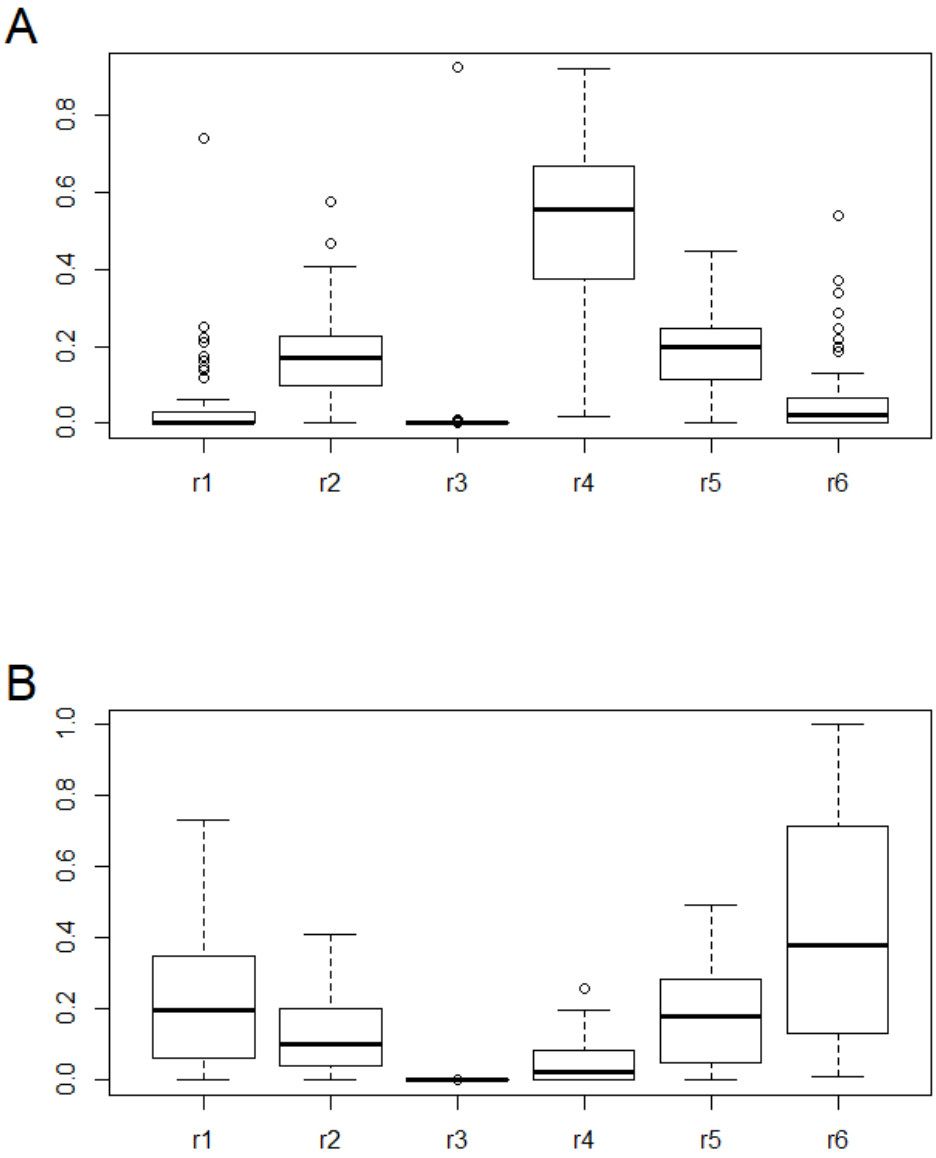

Figure 3. The summary of proportion estimation in gene expression with TOAST. A shows the result in tumor samples. B shows the results in normal samples

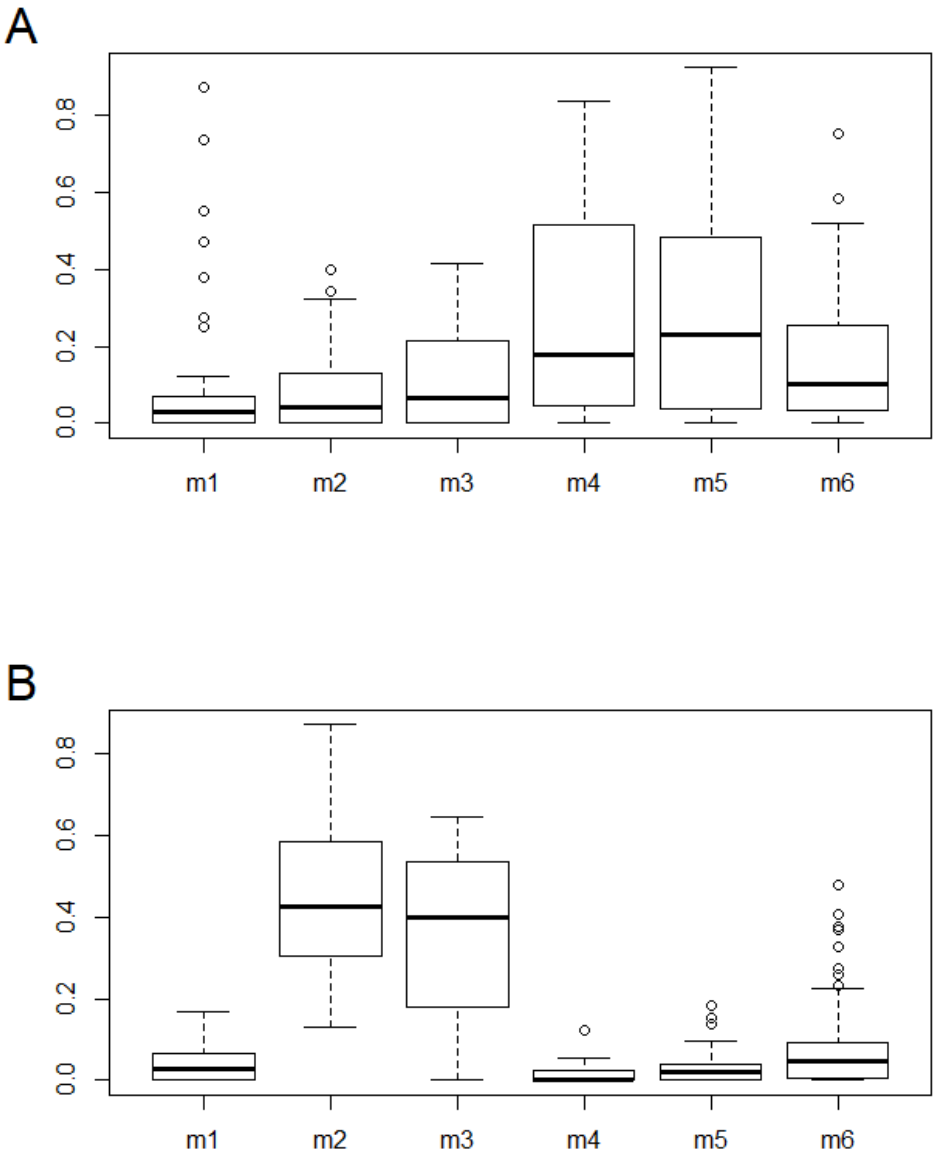

Figure 4. The summary of proportion estimation in DNA methylation with TOAST. A shows the result in tumor samples. B shows the results in normal samples

Table 2 The correlation matrix between gene expression and DNA methylation in tumor samples with TOAST

	m1	m2	m3	m4	m5	m6
r1	0.591	0.079	0.040	-0.263	-0.249	0.168
r2	-0.352	0.555	0.160	-0.027	-0.074	-0.003
r3	-0.056	-0.096	0.229	-0.049	0.044	-0.044
r4	0.129	-0.405	-0.378	0.157	0.124	-0.019
r5	-0.226	-0.063	0.178	0.030	0.068	-0.043
r6	-0.169	0.307	0.138	-0.011	-0.043	-0.035

Correlation between CIBERSORT and TOAST

We compared the estimated coefficients and proportions with CIBERSORT and TOAST methods in gene expression and DNA methylation. We firstly detected the correlation of estimations between two methods in gene expression. Most of the correlation values were located in (-0.2, 0.2), thus it showed the poor correlation between them (Table 3 and S3). The correlation results about estimations for DNA methylation with two methods also demonstrated that their estimations were not associated well (Table 4 and S4).

Table 3 Correlation results in gene expression between CIBERSORT and TOAST method

	rBCell	rCD8T	rCD4T	rCD56NK	rMono	rGran
r1	0.018	-0.113	0.165	0.007	-0.065	0.068
r2	0.171	0.002	0.021	0.008	-0.142	-0.282
r3	-0.092	-0.105	0.435	-0.051	-0.061	-0.037
r4	-0.317	0.126	-0.520	0.157	0.436	0.293
r5	0.362	0.077	0.275	-0.180	-0.427	-0.199

r6	0.173	-0.122	0.145	-0.103	-0.177	-0.115

Table 4 Correlation results in DNA methylation between CIBERSORT and TOAST method

	mCD4T	mCD8T	mBCell	mMono	mGran	mCD56NK
m1	-0.161	0.380	0.270	0.225	-0.082	-0.283
m2	-0.009	-0.155	-0.008	-0.076	0.166	0.043
m3	-0.049	-0.119	-0.080	-0.105	0.011	0.172
m4	-0.050	-0.104	0.087	-0.004	-0.379	0.323
m5	0.187	0.046	-0.150	-0.126	0.247	-0.260
m6	-0.011	-0.078	-0.123	0.137	0.185	0.007

Discussion

In this study, we tried reference-based (CIBERSORT) and reference-free (TOAST) regression methods to detect whether the estimated proportions of immune cell types would be changed when people were with the BRCA tumor compared to those without the BRCA tumor. In addition, the correlation between the proportion estimations of gene expression and DNA methylation were also produced to find out whether there were some relationships between them. With CIBERSORT method, the result showed that there was no significant difference in estimated coefficients of tumors and normal samples. And there was also no significant correlation between gene expression and DNA methylation. With TOAST, the estimated proportions in tumor samples were different from those in normal samples. And there were some proportions that had the correlation between gene expression and DNA methylation. Compared to two methods,

the correlations between the two methods for both gene expression and DNA methylation were not associated well.

During the study, there were some problems that should be addressed. Before processing the data analysis, gene markers would be selected firstly. In this study, we tried first selecting 1000 genes with the most variance using findRefinx function in the TOAST package. But the results showed that gene expression and DNA methylation did not correlated well with the TOAST method. After we removed the findRefinx function and put all the genes into the model in TOAST packages, the correlation appeared to be better. It appeared to me that we needed to select gene markers carefully and maintained more characteristics in datasets before processing the data.

Reference

1 Etienne Becht, Nicolas A Giraldo, Claire Germain, Aurelien de Reynies, Pierre Laurent-Puig, Jessica Zucman-Rossi, Marie-Caroline Dieu-Nosjean, Catherine Sautes-Fridman, and Wolf H Fridman, 'Immune Contexture, Immunoscore, and Malignant Cell Molecular Subgroups for Prognostic and Theranostic Classifications of Cancers', in *Advances in Immunology* (Elsevier, 2016), pp. 95-190.

2 Etienne Becht, Nicolas A Giraldo, Laetitia Lacroix, Bénédicte Buttard, Nabila Elarouci, Florent Petitprez, Janick Selves, Pierre Laurent-Puig, Catherine Sautès-Fridman, and Wolf H Fridman, 'Estimating the Population Abundance of Tissue-Infiltrating Immune and Stromal Cell Populations Using Gene Expression', *Genome biology,* 17 (2016), 218.

3 Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C Obenauf, Helen Angell, Tessa Fredriksen, Lucie Lafontaine, and Anne Berger, 'Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer', *Immunity,* 39 (2013), 782-95.

4 David D Chaplin, 'Overview of the Immune Response', *Journal of Allergy and Clinical Immunology,* 125 (2010), S3-S23.

5 Gulsun Erdag, Jochen T Schaefer, Mark E Smolkin, Donna H Deacon, Sofia M Shea, Lynn T Dengel, James W Patterson, and Craig L Slingluff, 'Immunotype and Immunohistologic Characteristics of Tumor-Infiltrating Immune Cells Are Associated with Clinical Outcome in Metastatic Melanoma', *Cancer research,* 72 (2012), 1070-80.

6 Francesca Finotello, and Zlatko Trajanoski, 'Quantifying Tumor-Infiltrating Immune Cells from Transcriptomics Data', *Cancer Immunology, Immunotherapy,* 67 (2018), 1031-40.

7 Ziyi Li, Zhijin Wu, Peng Jin, and Hao Wu, 'Dissecting Differential Signals in High-Throughput Data from Complex Tissues', *Bioinformatics,* 35 (2019), 3898-905.

8 Yan-gao Man, Alexander Stojadinovic, Jeffrey Mason, Itzhak Avital, Anton Bilchik, Bjoern Bruecher, Mladjan Protic, Aviram Nissan, Mina Izadjoo, and Xichen Zhang, 'Tumor-Infiltrating Immune Cells Promoting Tumor Invasion and Metastasis: Existing Theories', *Journal of Cancer,* 4 (2013), 84.

9 Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh, 'Robust Enumeration of Cell Subsets from Tissue Expression Profiles', *Nature methods,* 12 (2015), 453-57.

10 Franck Pagès, Anne Berger, Matthieu Camus, Fatima Sanchez-Cabo, Anne Costes, Robert Molidor, Bernhard Mlecnik, Amos Kirilovsky, Malin Nilsson, and Diane Damotte, 'Effector Memory T Cells, Early Metastasis, and Survival in Colorectal Cancer', *New England journal of medicine,* 353 (2005), 2654-66.

11 Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett, 'New Support Vector Algorithms', *Neural computation,* 12 (2000), 1207-45.

12 Padmanee Sharma, Klaus Wagner, Jedd D Wolchok, and James P Allison, 'Novel Cancer Immunotherapy Agents with Survival Benefit: Recent Successes and Next Steps', *Nature Reviews Cancer,* 11 (2011), 805-12.

13 Xiufen Zheng, James Koropatnick, Di Chen, Thomas Velenosi, Hong Ling, Xusheng Zhang, Nan Jiang, Benjamin Navarro, Thomas E Ichim, and Bradley Urquhart, 'Silencing Ido in Dendritic Cells: A Novel Approach to Enhance Cancer Immunotherapy in a Murine Breast Cancer Model', *International journal of cancer,* 132 (2013), 967-77.

Appendix

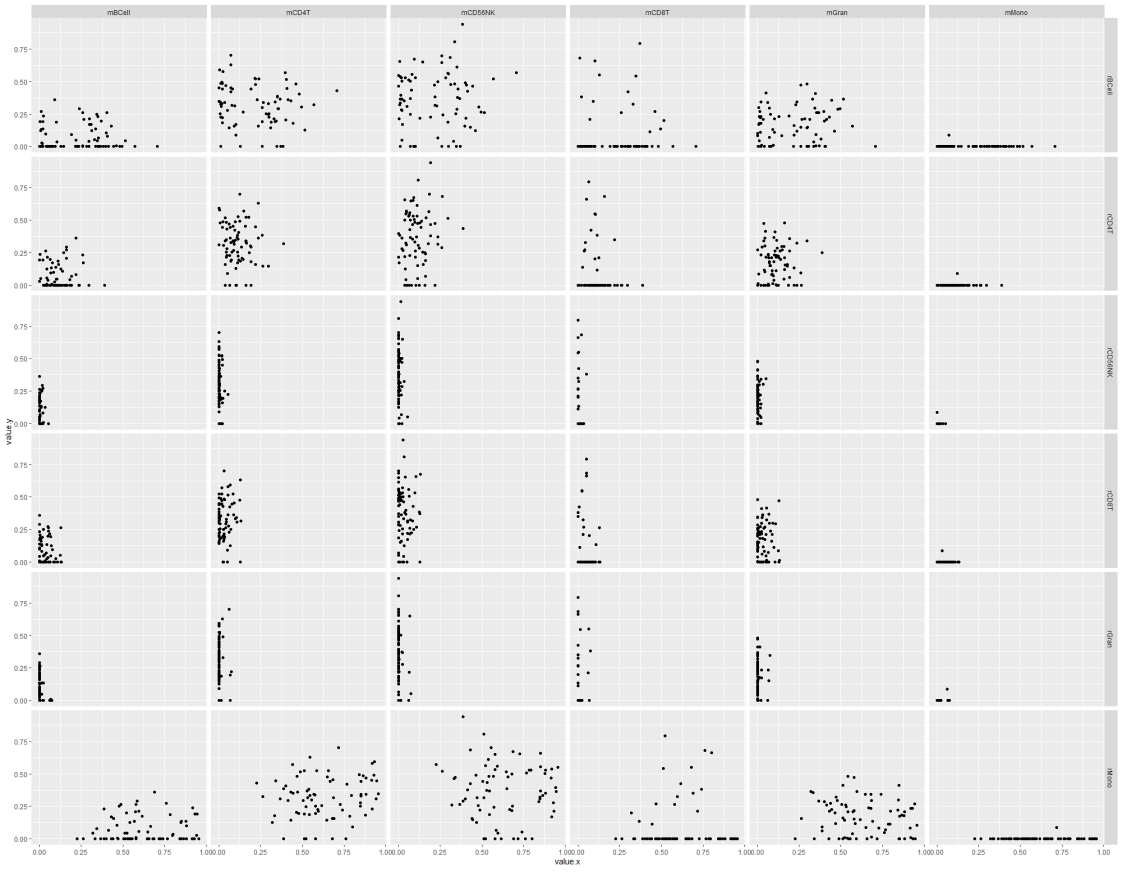

Figure S1. The scatter plot of correlation between gene expression and DNA methylation in tumor sample with CIBERSORT method.

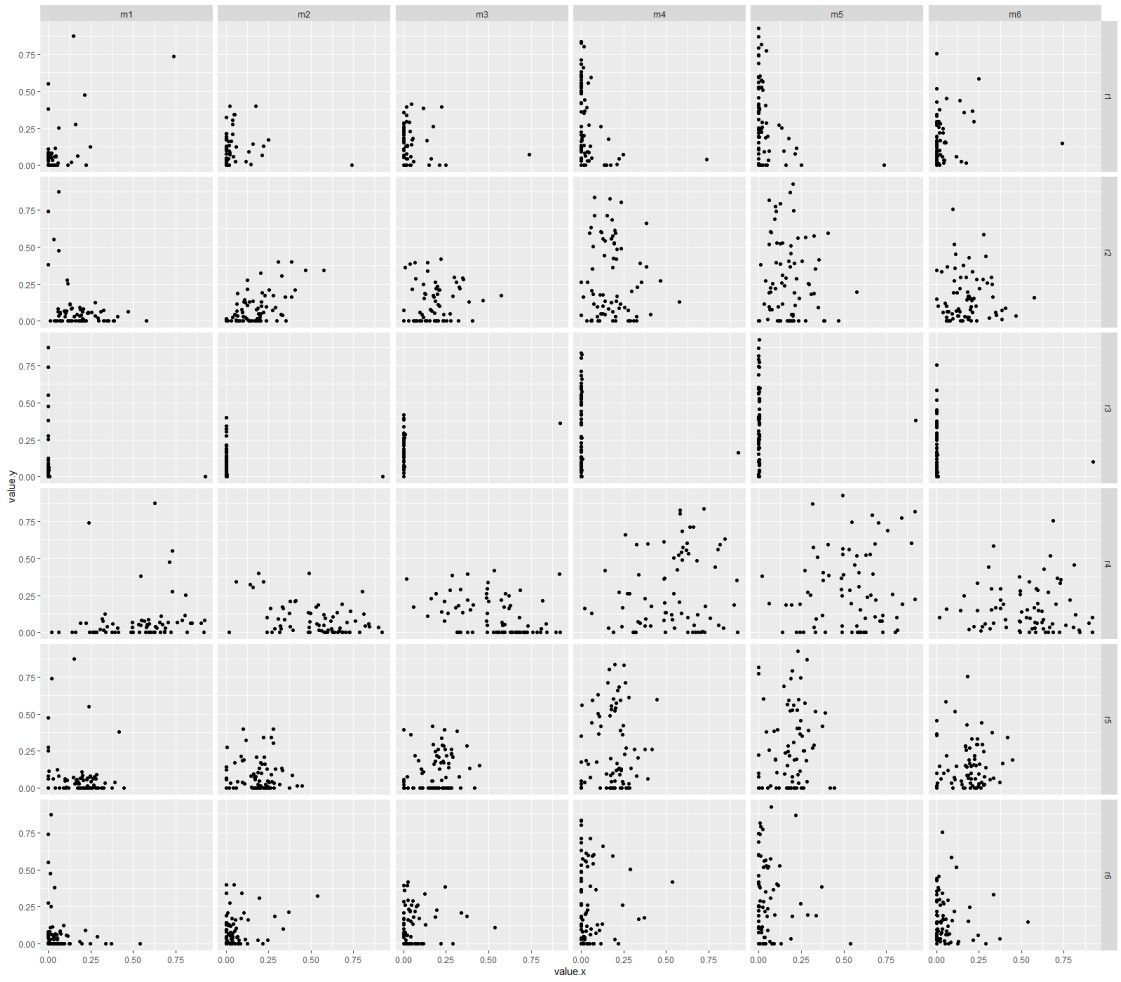

Figure S2. The scatter plot of correlation between gene expression and DNA methylation in tumor sample with TOAST method.

Table S1 P value for correlation matrix between gene expression and DNA methylation in tumor samples with CIBERSORT

	mCD4T	mCD8T	mBCell	mMono	mGran	mCD56NK
rBCell	0.222	0.875	0.923	0.400	0.386	0.746
rCD8T	0.474	0.678	0.805	0.920	0.898	0.488
rCD4T	0.416	0.384	0.874	0.876	0.560	0.292
rCD56NK	0.233	0.310	0.331	0.662	0.287	0.295
rMono	0.169	0.912	0.788	0.647	0.283	0.684
rGran	0.744	0.022	0.150	0.000	0.700	0.150

Table S2 P value for correlation matrix between gene expression and DNA methylation in tumor samples with TOAST

	m1	m2	m3	m4	m5	m6
r1	0.000	0.499	0.735	0.022	0.030	0.146
r2	0.002	0.000	0.167	0.816	0.523	0.981
r3	0.631	0.410	0.046	0.676	0.708	0.706
r4	0.267	0.000	0.001	0.177	0.286	0.869
r5	0.050	0.592	0.124	0.798	0.560	0.714
r6	0.144	0.007	0.235	0.924	0.713	0.767

Table S3 P values for correlation results in gene expression between CIBERSORT and TOAST method

	rBCell	rCD8T	rCD4T	rCD56NK	rMono	rGran
r1	0.874	0.330	0.154	0.951	0.575	0.559
r2	0.140	0.986	0.855	0.948	0.221	0.014
r3	0.430	0.367	0.000	0.659	0.601	0.752
r4	0.005	0.277	0.000	0.177	0.000	0.010
r5	0.001	0.509	0.016	0.120	0.000	0.086
r6	0.134	0.292	0.212	0.377	0.127	0.325

Table S4 P values for correlation results in DNA methylation between CIBERSORT and TOAST method

	mCD4T	mCD8T	mBCell	mMono	mGran	mCD56NK
m1	0.164	0.001	0.018	0.051	0.480	0.013
m2	0.940	0.181	0.942	0.517	0.153	0.713
m3	0.671	0.308	0.490	0.368	0.928	0.137
m4	0.671	0.371	0.457	0.975	0.001	0.004
m5	0.106	0.693	0.194	0.277	0.031	0.023
m6	0.927	0.506	0.290	0.237	0.109	0.950

Table S5 P values for comparison of tumor and normal samples in gene expression with CIBERSORT. tr represents to tumor samples. nr represents to normal samples.

	nrBCell	nrCD8T	nrCD4T	nrCD56NK	nrMono	nrGran
trBCell	0.829	0.000	0.001	0.000	0.000	0.000
trCD8T	0.000	0.024	0.000	0.000	0.000	0.000
trCD4T	0.000	0.000	0.053	0.000	0.000	0.000
trCD56NK	0.000	0.006	0.000	0.005	0.000	0.111
trMono	0.000	0.000	0.000	0.000	0.795	0.000
trGran	0.000	0.012	0.000	0.056	0.000	0.121

Table S6 P values for comparison of tumor and normal samples in DNA methylation with CIBERSORT. tm represents to tumor samples. nm represents to normal samples.

	nmCD4T	nmCD8T	nmBCell	nmMono	nmGran	nmCD56NK
tmCD4T	0.888	0.000	0.000	0.000	0.000	0.000
tmCD8T	0.000	0.030	0.059	0.000	0.000	0.000
tmBCell	0.000	0.007	0.009	0.000	0.000	0.000
tmMono	0.000	0.054	0.000	0.321	0.000	0.000
tmGran	0.000	0.000	0.000	0.000	0.026	0.000
tmCD56NK	0.106	0.000	0.000	0.000	0.000	0.179

Table S7 P values for comparison of tumor and normal samples in gene expression with TOAST. tr represents to tumor samples. nr represents to normal samples.

	nr1	nr2	nr3	nr4	nr5	nr6
tr1	0.000	0.000	0.002	0.332	0.000	0.000
tr2	0.046	0.001	0.000	0.000	0.988	0.000
tr3	0.000	0.000	0.307	0.008	0.000	0.000
tr4	0.000	0.000	0.000	0.000	0.000	0.018
tr5	0.102	0.000	0.000	0.000	0.587	0.000
tr6	0.000	0.000	0.000	0.592	0.000	0.000

Table S8 P values for comparison of tumor and normal samples in DNA methylation with TOAST. tm represents to tumor samples. nm represents to normal samples.

	nm1	nm2	nm3	nm4	nm5	nm6
tm1	0.072	0.000	0.000	0.001	0.014	0.782
tm2	0.001	0.000	0.000	0.000	0.000	0.852
tm3	0.000	0.000	0.000	0.000	0.000	0.084
tm4	0.000	0.000	0.013	0.000	0.000	0.000
tm5	0.000	0.000	0.030	0.000	0.000	0.000
tm6	0.000	0.000	0.000	0.000	0.000	0.001