Yu-Lab Weekly Report (Update)

Haojun Li

April 17, 2019

1 Introduction

1.1 Background

In tumor microenvironment, the cellular composition of the immune infiltrate of tumors can shed light on the escape mechanisms that tumor cell use to evade the immune response. In clinical trials, it can be used to stratify patients to assign most suitable treatment options depending on the targeted cell type, hence increasing the overall chances of success, and ultimately accelerating access to improved treatment options. Changes in tissue composition are often indicative of disease progression or drug response. As interactions between the immune system and tumour cells are governed by a complex network of cell–cell interactions, especially, knowing the specific immune cell composition of a solid tumour may be essential to predict a patient's response to immunotherapy [20].

With the development of gene sequencing technology, the investigation of the causes and treatment effects of diseases has become more available. Gene expression profiling, RNA-seq and mircoarray has been optimized to be applicable in clinical.

In the past, Fluorescence Activated Cell Sorting (FACS), Immunohistochemistry (IHC) and Immuno Fluorescence (IF) which utilize only a small number of cell type specific markers was used to measure the components in a mixed patient sample. In recent years, more and more biologists use single cell RNA-seq to calculate the cellular composition of a mixed tissue or sample to explore the pathogenic mechanism and therapeutic methods, especially in complex tissue samples, such as tumor tissues. In single cell RNA-seq, individual cells from a mixed sample are labeled and sequenced, and cell components in tissue are determined by grouping and clustering. Both FACS and single cell RNA-seq not only involve significant time, effort, and expense but also result in insufficient RNA abundance for further quantification of gene expression in some cases.

Deconvolution is the method of estimating individual signal components from their mixtures in the area of signal processing. It has diverse applications in fields ranging from hyperspectral imaging to noise cancellation in audio recordings. In the field of biomedical research, we can use this method to estimate the cellular specific components in complex tissues in order to identify the changes in cellular components in tissues. Deconvolving data from biopsy tissues allows further investigation of the interaction between tumor and micro-environmental cells, and role of such interaction in the progression of cancer.

At least seven major issues raise concerns that the *in silico* methods could be prone to errors and cannot reliably portray the cellular heterogeneity of the tumor microenvironment. The main issues are as follows.

- Current techniques depend on the expression profiles of purified cell types to identify reference genes and therefore rely heavily on the data source from which the references are inferred and could this be inclined to overfit these data.
- Current methods focus on only a very narrow range of the tumor microenvironment, usually a subset of immune cell types, and thus do not account for the further richness of cell types in the microenvironment, including blood vessels and other different forms of cell subsets.
- The ability of cancer cells to "imitate" other cell types by expressing immune-specific genes, such as a macrophage-like expression pattern in tumors with parainflammation; only a few of the methods take this into account.

- The ability of existing methods to estimate cell abundance has not yet been comprehensively validated in mixed samples. Cytometry is a common method for counting cell types in a mixture and, when performed in combination with gene expression profiling, can allow validation of the estimations. However, in most studies that include cytometry validation, these analyses were performed on only a very limited number of cell types and a limited number of samples.
- Deconvolution approaches are prone to many different biases because of the strict dependencies among all cell types that are inferred. This could highly affect reliability when analyzing tumor samples, which are prone to form non-conventional expression profiles.
- Deconvolution infers an increasing number of closely related cell types.
- Deconvolution analysis heavily relies on the structure of the reference matrix, which limits its application to the resource used to develop the matrix.

1.2 Prior Art

1.2.1 2019.4.3

(1) This paper [7] uses co-expression patterns in large tumor gene expression datasets to evaluate previously reported candidate cell type marker genes lists, eliminate numerous false positives and identify a subset of high confidence marker genes. The main contributions of this paper include 1) identify a list of 60 marker genes whose expression levels measure 14 immune cell populations; 2) most genes previously reported to be enriched in a single cell type have co-expression patterns inconsistent with cell type specificity. The main disadvantages of this paper include 1) many cell types are short of markers such as CD4 T cell, Monocytes.

1.2.2 2019.3.2

(1) This paper [21] presents a systematic approach for benchmarking such computational methods and assessed the accuracy of tools at estimating nine different immune and stromal cells from bulk RNA-seq samples. The main contributions of this paper include 1) collecting an amount of single cell RNA-seq data set; 2) Benchmark methods performance between cell types, background predictions and spillover to give a guidelines for method selection.

(2) This paper [8] describes the state-of-the-art computational methods for the quantification of immune cells from transcriptomics data and discuss the open challenges that must be addressed to accurately quantify immune infiltrates from RNA sequencing data of human bulk tumors. This review makes a detailed classification and induction of the current deconvolution methods.

1.2.3 2019.2.25

(1) This paper [22] presents immunoStates, a basis matrix built using 6160 samples with different disease states across 42 microarray platforms. The main contributions of this paper include 1) it found that deconvolution accuracy is more dependent on the signature matrix through experiments; 2) they create immunoStates which significantly reduces biological and technical biases by using healthy and disease samples and Hedeg's g effect sizes; 3) This matrix has leads to better deconvolution performance than LM22 and IRIS, independent of the deconvolution algorithms tested.

(2) This paper [18] presents ImSig from tissue transcriptomics data using a network-based deconvolution approach. The main contributions of this paper include 1) it used tissue sample which can reflect the differentiation or activation state of similar cells within tissues; 2) they used gene correlation network, Markov clustering alforithm (MCL), Gene Ontology (GO) in signature creating. The main disadvantage of this paper is the lack of performance comparison of this signature matrix with other common feature matrices such as LM22 and IRIS.

(3) This paper [6] introduced a method called MySort, using microarray data from the public domain to profile gene expression pattern of twenty-two immune cell types. MySort is an improvement over CIBERSORT with a few additional steps and changes in the pipline. The main contributions of this paper include 1) using enrichment score (ES), two-sided unequal variance T-test and differentially expressed genes (DEGs) to define signature matrix; 2) it added clustering,

pearson correlation to remove outliers based on CIBERSORT deconvolution pipeline. The main disadvantage of this paper is the lack of performance comparison with other methods.

1.2.4 2018.12.21

This paper [4] introduced a method called Microenvironment Cell Populations-counter (MCP-counter), which allows the robust quantification of the absolute abundance of eight immune and two stromal cell populations in heterogeneous tissues from transcriptomic data. The main contributions of this paper include 1) the paper has detailed the cell type classification and collection of purified cell data; 2) This method screens the marker for specific cell types. The main disadvantages of this paper include 1) the algorithm has no feature matrix, but only judges the expression of the selected marker; 2) MCP-counter score cannot represent the actual fraction.

1.2.5 2018.11.23

This paper [3] introduced a method called xCell, a novel gene signature-based method, and use it to infer 64 immune and stromal cell types. The main contributions of this paper include 1) collecting the most comprehensive resource to date of primary cell types, spanning the largest set of human cell types; 2) gene signatures are rank-based and are therefore suitable for cross-platform transcriptome measurements. The main disadvantages of this paper include 1) the inferences are strictly enrichment scores, and cannot be interpreted as proportions because of the inability to translate the minimum and maximum scores produced by ssGSEA to clear proportions and 2) Expression of purified cell types measured from different tissues may be different.

1.2.6 2018.11.09

(1) This paper [10] introduced a method called FARDEEP utilizing an adaptive least trimmed square to automatically dectet and removing outliers before estimating the cell compositions. The main contributions of this paper include 1) evaluates all outliers across the datasets and 2) examining the true immune gene signature using non-negative regression and 3) improving prognostic potential when dealing more complex datasets with significant carcinoma cell content. This method has compared with other method including NNLS, PERT [19], DCQ [2], CIBERSORT [15]. But its Performance criteria are sum of squared error(SSE) and codfdicient of determination denoted as R-square. Especially, FAEDEEP has better performance than CIBERSORT through SSE and R-square. But CIBERSORT used RMSE and R to evaluate performance, and the review [14] selects MSE. Different performance criteria will affect the performance comparsion.

(2) This paper [16] discusses the application of Deconvolution in Tumor infiltrating leukocytes (TILs). The main contributions of this paper include 1) Comparing and summarizing the functions of existing deconvolution methods in TILs and 2) Discussing the more significant problems at present.

1.2.7 2018.11.02

(1) This paper [11] introduced a method called Dsection based on probability model. The main contributions of this paper include 1) de-noising uncertain prior information about cell-type proportions and 2) Use prior knowledge to ensure the accuracy of prediction. The shortcoming of this paper is that the number of cell types need to be known.

(2) This paper [12] thinks that CIBERSORT [15] actually has more errors including 1) considering the significant impact of data normalization on deconvolution results, and nonbiological negative correlations mainly due to collinearity in CIBERSORT, 2) using flow cytometry experiment to argue that closely related immune cell types should be positively correlated in abundance, whether in blood or in tumors, 3) claiming that up to 25 percent of LM22 genes are positively correlated with tumor purity.

(3) This paper [17] responds to accusations and gives explanations to another paper [12] including 1) TIMER mixed populations do not sum to 100 percent and are therefore unsuitable for addressing this topic, 2) no significant association between pairwise correlations of leukocyte estimates in tumors and pairwise correlations of corresponding expression profiles in LM22, 3) claiming that exclusion of all significantly positively correlated genes from LM22 had virtually no impact on tumor deconvolution performance.

1.2.8 2018.10.28

I prepare for the presention based on the paper [14], review of deconvolution, the Related method like LLSR [1], QP [9], CIBERSORT [15], DSA, TIMER and so on. What's more, I summarized six points that may improve the doconvolution performance in the future.

1.2.9 2018.10.19

(1) This paper "ML Estimation of Cell Fraction" introduced a method for ML estimation.The main contributions of this paper include 1) Applying decoding from channel coding to Deconvolution and 2) Use prior knowledge to ensure the accuracy of prediction.

(2) This paper [9] introduced a method for quadratic programming. The main contributions of this paper include 1) buliding a system of linear equations through the introduction of a least squares and 2) the explicit modeling of physical constraints in both the description of the problem as well as its solution and 3) suggesting we can only apply on specific tissue to gain the best performance.

1.2.10 2018.10.12

I configure the environment and install CIBERSORT [15] on Ubuntu. I try to run GSE11103, mixtures of four blood cell line. And I also try to run the GEPs from paper of Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing, but errors occur. I think it may due to the format of data. On the other hand, I read the CIBERSORT online method and supplementary in order to find it how to evaluation its performance and compare with other methods. CIBERSORT use Pearson correlation and r.m.s. error to measure linear fit and estimation bias. It also simulates tumors with added noise and cell subset detection limit to compare with other method. But the main problem is the data need to combine two GEPs, but they don't provide the detail methods.

2 Method

2.1 Data preparation

2.1.1 Cancer cell expression profiles

I downloaded about 500 sample LUAD RNA-Seq expression profiles from TCGA. I transformed ensembl id into gene, transformed FPKM format into TPM, removed nulls and duplicates with R code, finally sort TCGA data into available expression profiles.

2.1.2 Purity cell expression profiles

(1) I downloaded the RNA-Seq purified cell matrix including SRR1740038, SRR1740039, SRR1740040, SRR1740041, SRR1740066, SRR1740067, SRR1740068, SRR1740069 in ENCODE project following the paper [3].

(2) I downloaded the microarray purified cell matrix in GEO with cel format following the paper [4].

10 cell types were selected which have a enough amount of purified expression samples including B cells, CD4 T cells, CD8 T cells, NK cells, Dendritic cells, Neutrophils, Monocytes, fibroblast, Macrophages, Endothelium. I collected each cell type purified samples from microarray as follow.

1. B cell are from GSE22886, GSE43677, GSE9378, GSE54017, GSE13411, GSE12845, GSE18723, above 125 sample.
2. CD4 T cell are from GSE22886, GSE2270, GSE8835, GSE7497, GSE36476, GSE20198, GSE13017, GSE17354, above 164 samples.
3. CD8 T cell are from GSE22886, GSE8835, GSE11188, GSE17354, above 44 samples.
4. Endothelium are from GSE9378, GSE9877, GSE32710, GSE22688, GSE14230, above 116 samples.
5. Fibroblast are from GSE40839, GSE30242, above 45 samples.

6. Macrophages are from GSE16972, GSE22528, GSE5099, GSE9874, GSE49072, GSE8515, above 158 samples.
7. Monocytes are from GSE22886, GSE8921, GSE16972, GSE9378, GSE11943, GSE5099, GSE12837, GSE46923, above 170 samples.
8. Myeloid Dendritic Cells are from GSE22886, GSE12259, GSE14816, GSE9946, GSE46923, above 44 samples.
9. Neutrophils are from GSE22886, GSE12837, GSE19556, GSE12484, above 18 samples.
10. NK cells are from GSE22886, above 15 samples.

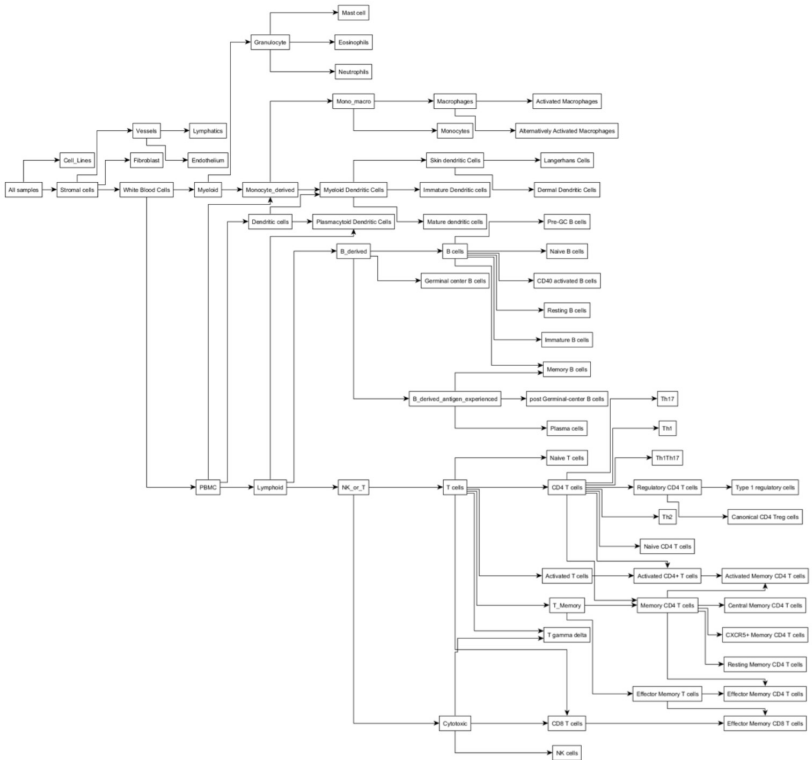

Figure 1: Cell type classification pyramidal graph

2.1.3 True mixture expression profiles with konwn fraction

I downloaded GSE64385 from GEO [4] to gain In vitro RNA mixtures. The dataset is microarray data in Affymetrix Human Genome 133 Plus 2.0 Gene expression platform including 12 mixtures sample and 5 immune cell with known fraction. 5 immune cell populations were sorted from 3 healthy donors' peripheral bloods. Peripheral Blood Mononuclear Cells (PBCMs) and PolymorphoNuclear Cells (PMN) were separated using gradient centrifugation. T cells (DAPI-/CD3+/CD14-/CD19-/CD56-), monocytes (DAPI-/CD3-/CD14+/CD19-/CD56-), B cells (DAPI-/CD3-/CD14-/CD19+/CD56-) and NK cells (DAPI-/CD3-/CD14-/CD19-/CD56+) were FACS-sorted from PBMCs and neutrophils (DAPI-/CD66b+/CD19-/CD3-/CD56-/CD14-) were sorted from PMNs. RNA was extracted from the purified cell population, as well as from the HCT116 colon cancer cell line. RNAs from pure populations were then mixed in various proportions.

2.2 Mixture simulating and marker gene selecting

(1) I used ReadAffy function in R code to read purified cel format microarray data, then standardize it with rma. For subsequent mixing, I made a power to its expression value.

2.3 Algorithm implementation

2.3.1 Supervised learning - mult-output regression

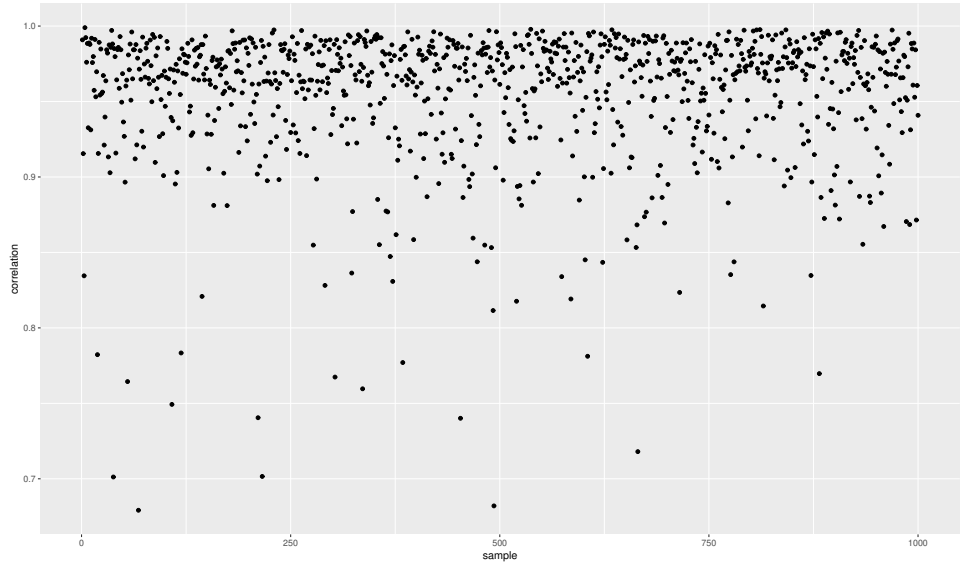

Figure 2: correlation between predicted an true fraction

2.4 Suggestion

Possible ways to improve the deconvolution are as follows:

1. Different methods have different performance on different data, so the method need requires a corresponding prior knowledge of different data [19] [13] [11].
2. Explicit STO can cause lower performance of deconvolution, so the method of filtering violating feature need to be improved [14].
3. Different upper/lower bounds have been proposed in the literature to prefilter expression values prior to deconvolution. Filter lower and upper bounds of gene expressions using adaptive range filtering may improve deconvolution performance[14].
4. Select invariant (among references and between references and samples) cell-type-specific markers to enhance the discriminating power of the basis matrix [14] [15] [23] [13].
5. Solve the regression using the L2 loss function together with an R2 regularizer, or group LASSO if sparsity is desired among groups of tissues/cell types [14] [15].
6. For tumor cells, construct a purification characteristic matrix corresponding to tumor malignant cells, or perform prior probability statistics on the purity of tumor cells to improve the accuracy of deconvolution [23][5].
7. We should consider the performance criteria.

3 Simulations and results

3.1 Purired RNA-seq Data

Dendritic cells RNA-seq TPM data from Quantiseq dataset, sequenced by Quantiseq.

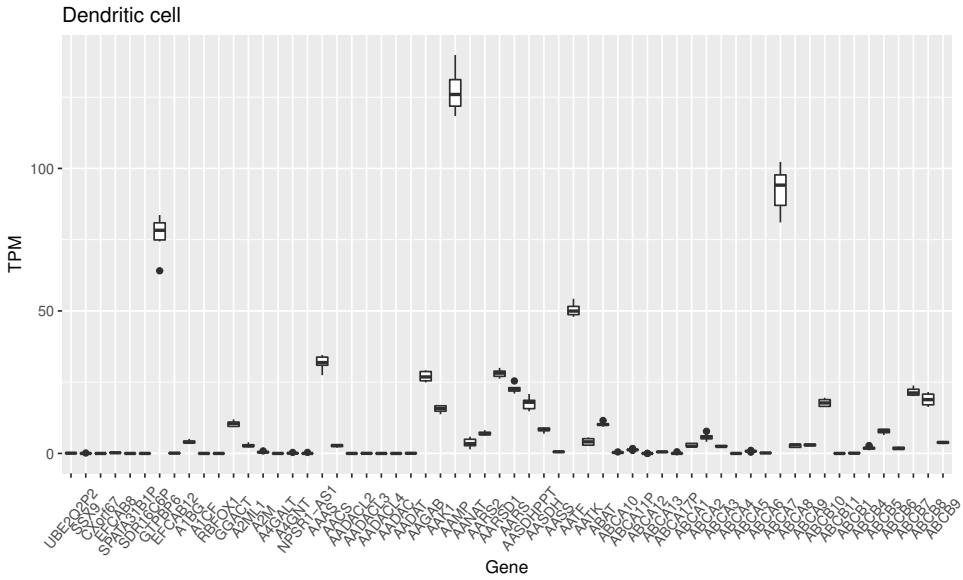

Figure 3: Dendritic cells TPM distribution

3.2 Purified Microarray Data

Data are from GSE473, GSE6740, GSE13017, GSE17354, including 118 samples in order to explore microarray CD4 T cells expression distribution. Different genes in CD4 T cells are clearly distributed in cell expression.

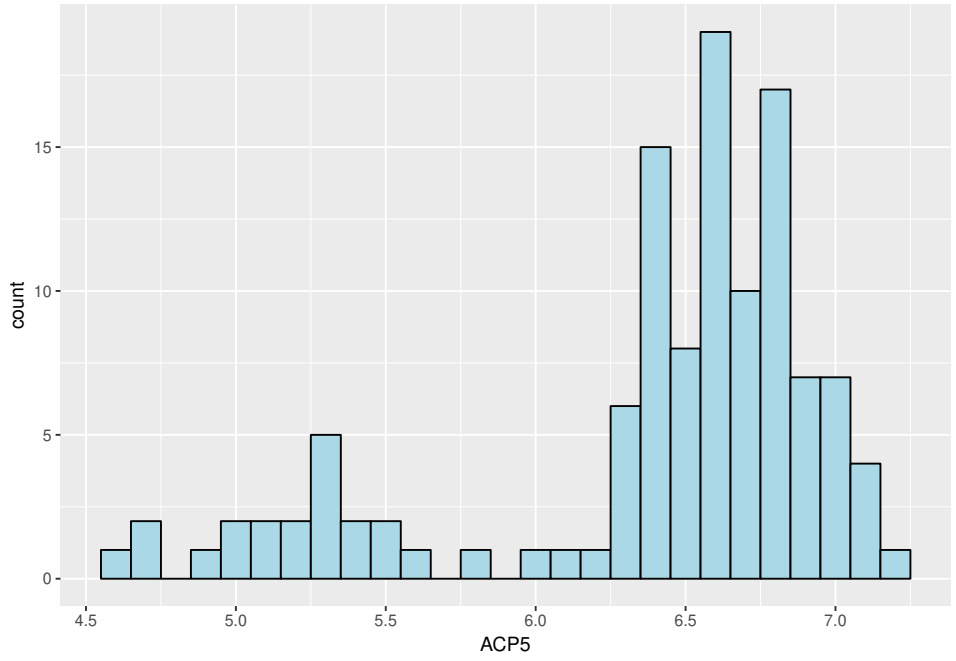

Figure 4: CD4 T cells ACP5

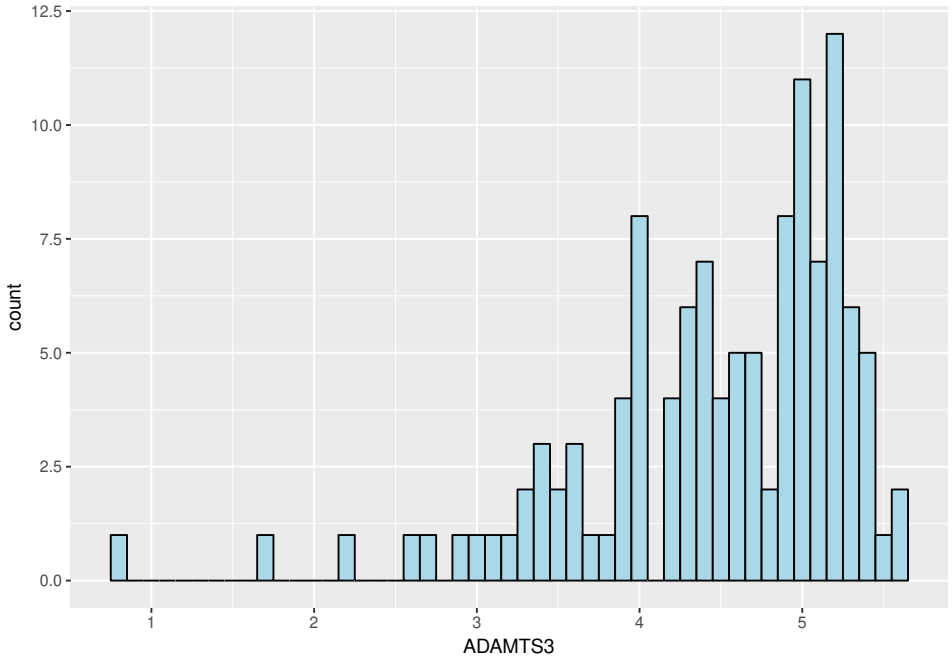

Figure 5: CD4 T cells ADAMDEC1

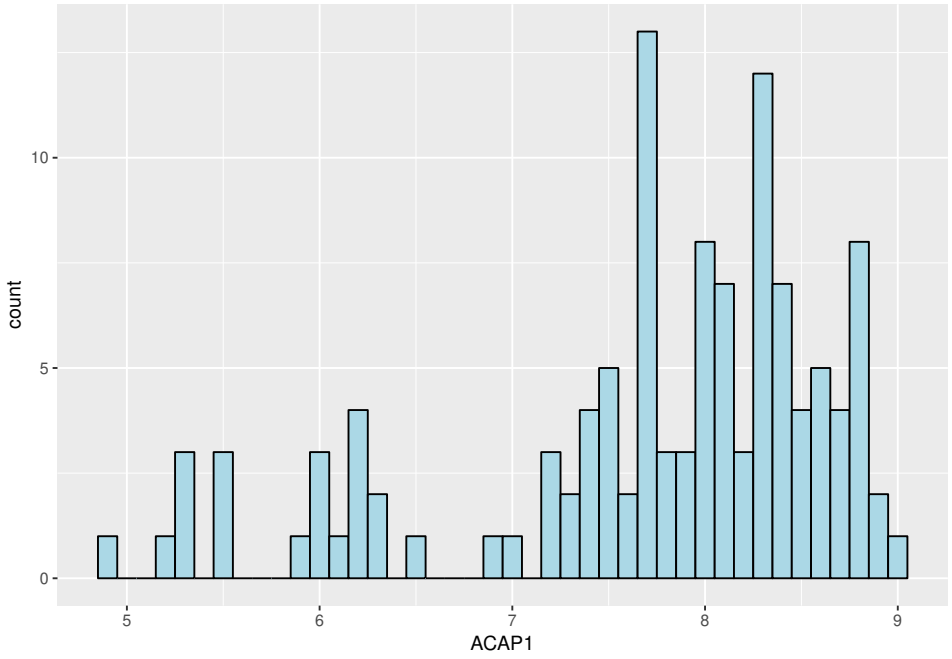

Figure 6: CD4 T cells ACAP1

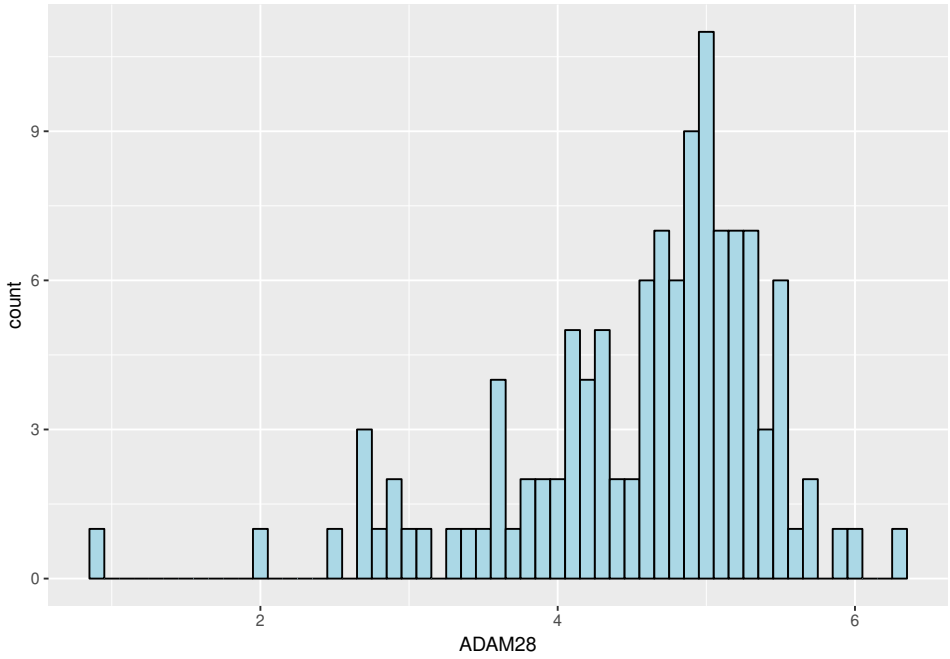

Figure 7: CD4 T cells ADAM28

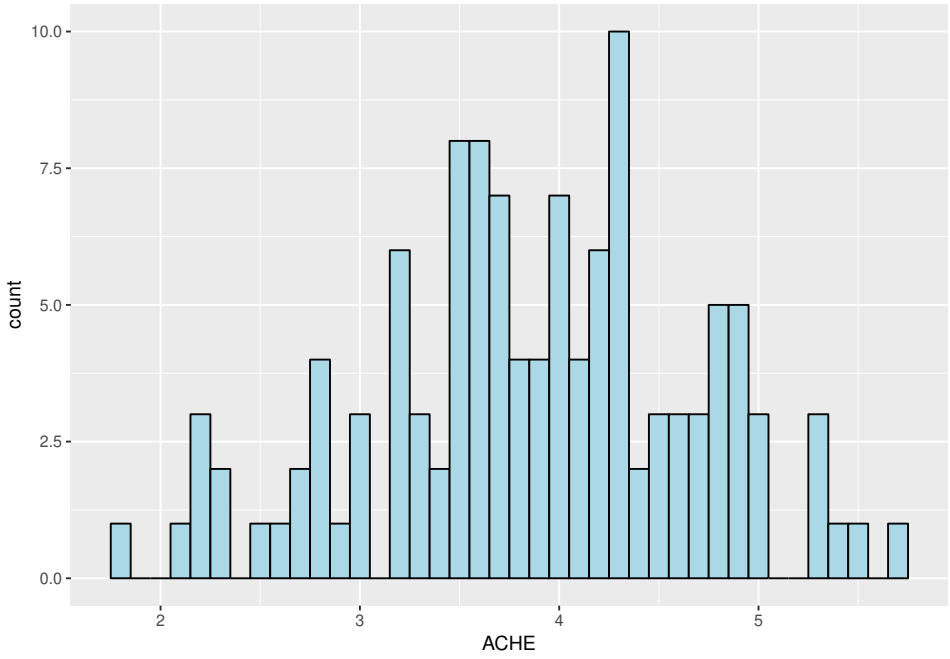

Figure 8: CD4 T cells ACHE

3.3 MCP-count test

I downloaded cel format Micoarray data from GSE22886. Six cell types were selected including B cell, CD8 T cell, NK cell, Myeloid dendritic cell, Monocytes, Neutrophils. I mixed them with random fraction, and ran MCP-count with them.

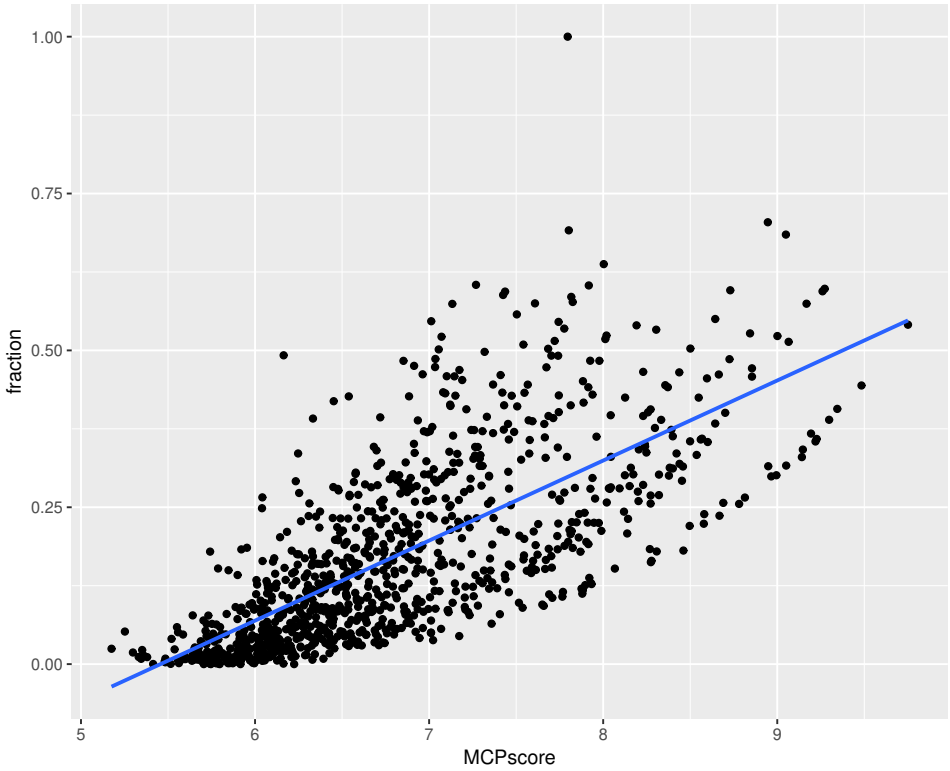

Figure 9: MCP-count CD8 T cell

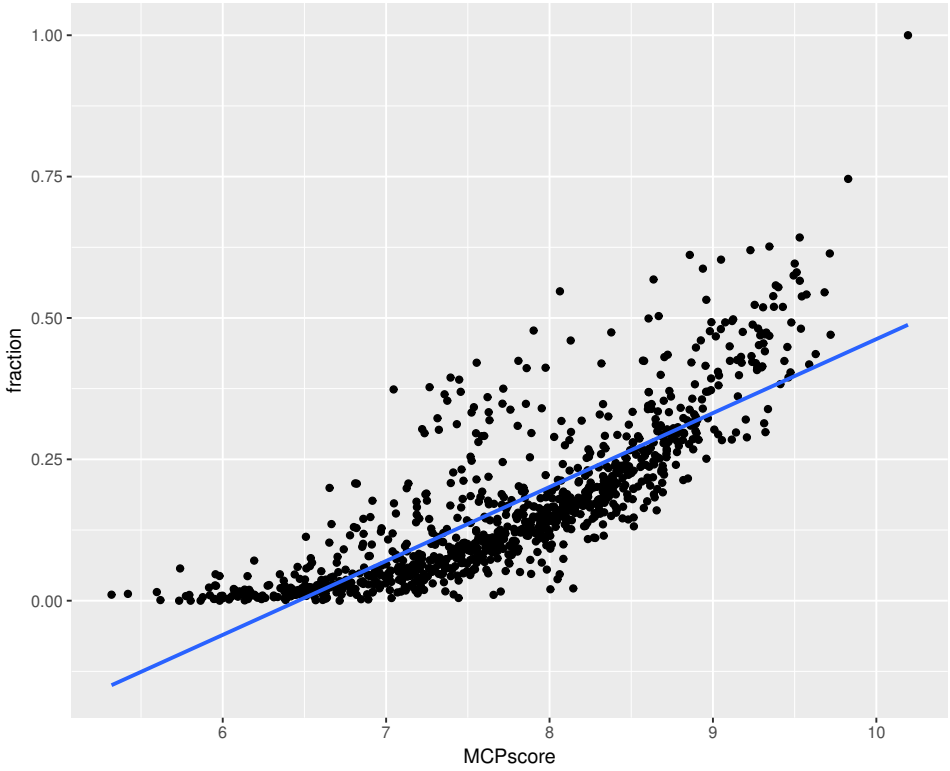

Figure 10: MCP-count B cell

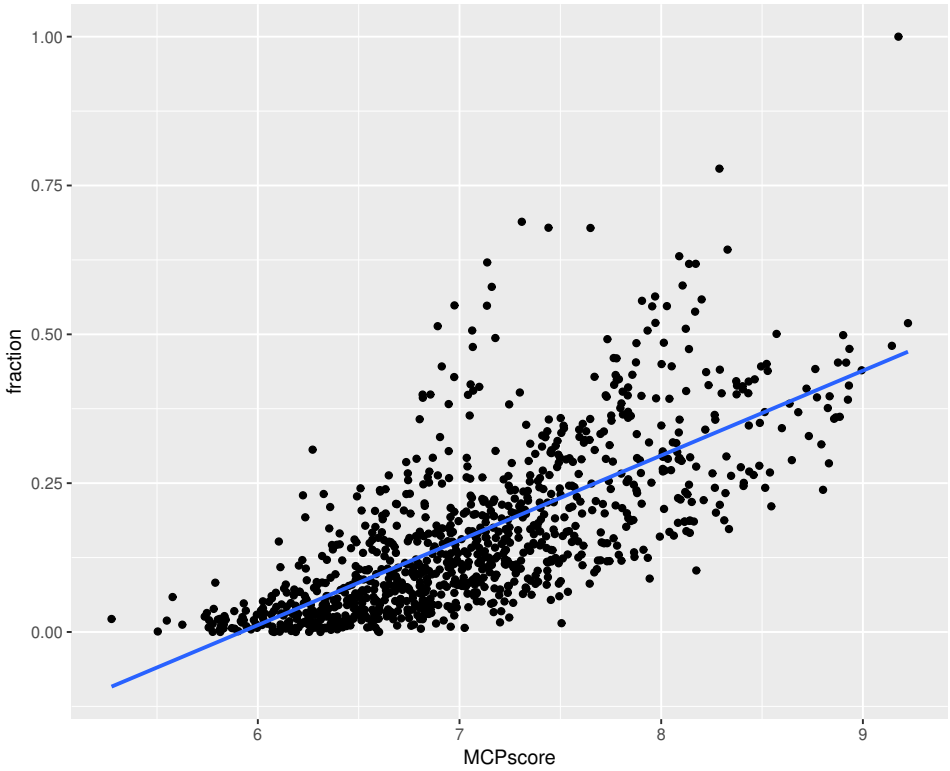

Figure 11: MCP-count NK cell

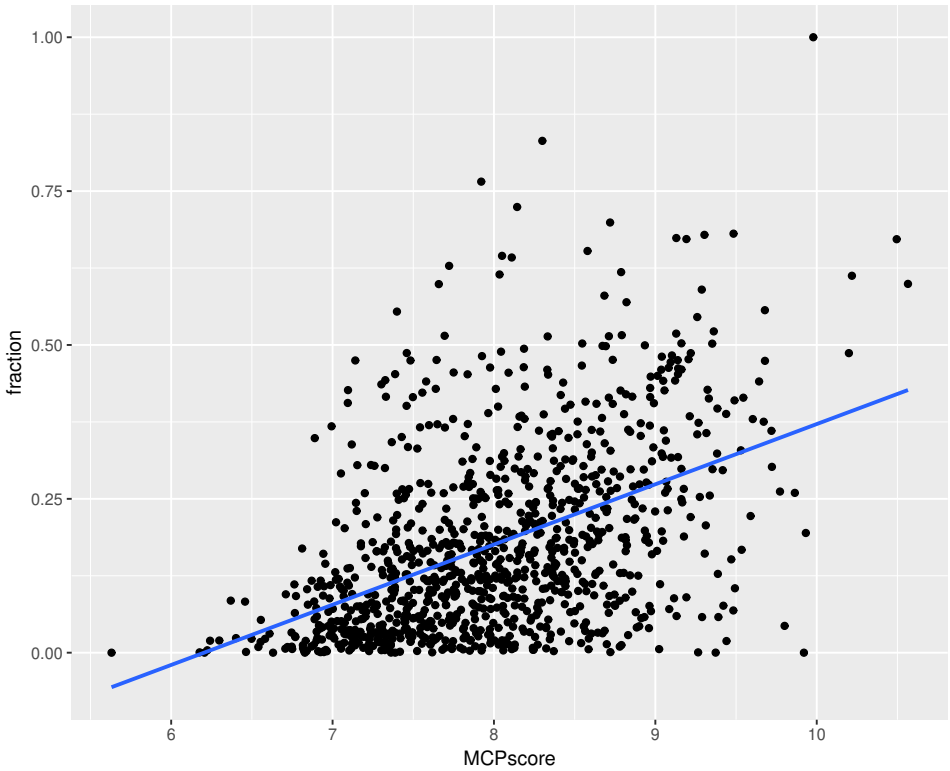

Figure 12: MCP-count Monocytes

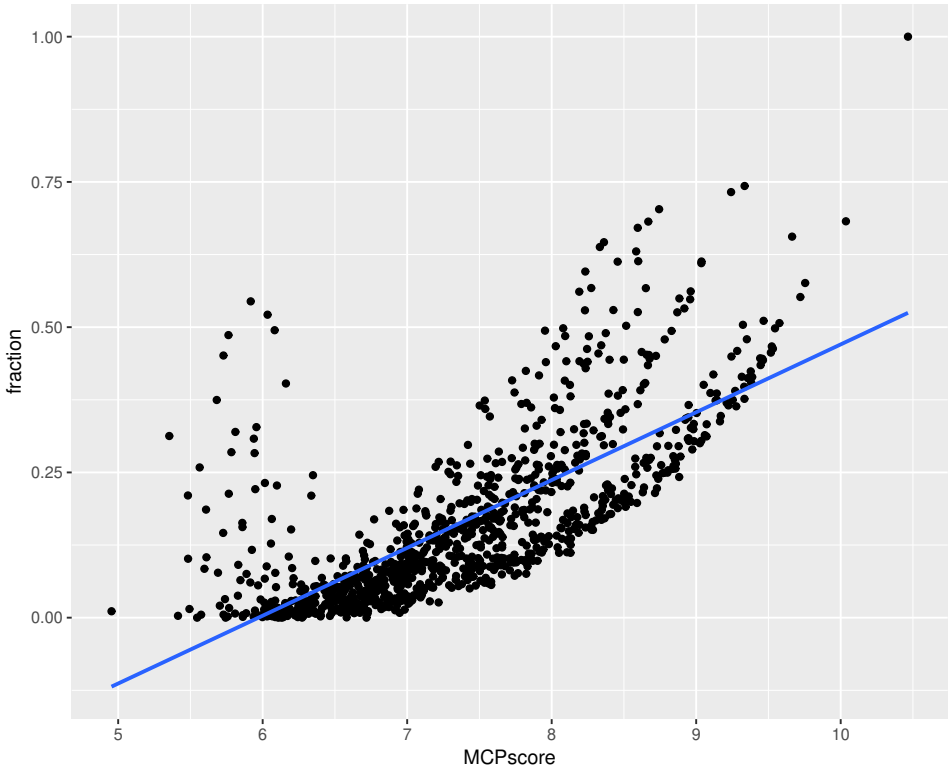

Figure 13: MCP-count Neutrophils

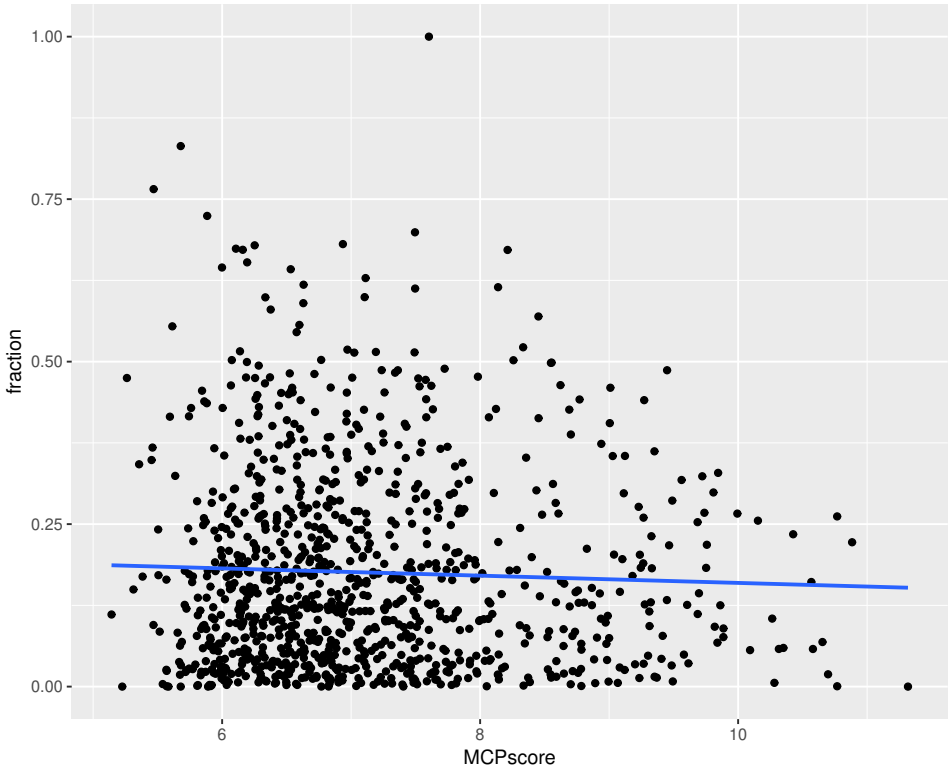

Figure 14: MCP-count Myeloid dendritic cell

It can be seen from the experimental results that the correlation between the MCP-count score and the true mixing ratio of the mixed samples is not high, especially Myeloid dendritic cells and Monocytes. one

3.4 Mixture availability test

I got GSE64385 microarray data (including 10 real-time mixed samples of known proportions) from geo. I extracted a given cell type microarray expression data from a purified data set with a consistent ratio of real samples. I verified the availability of the artificial mixture by correlating the expression of the artificial mixture with the actual mixture.

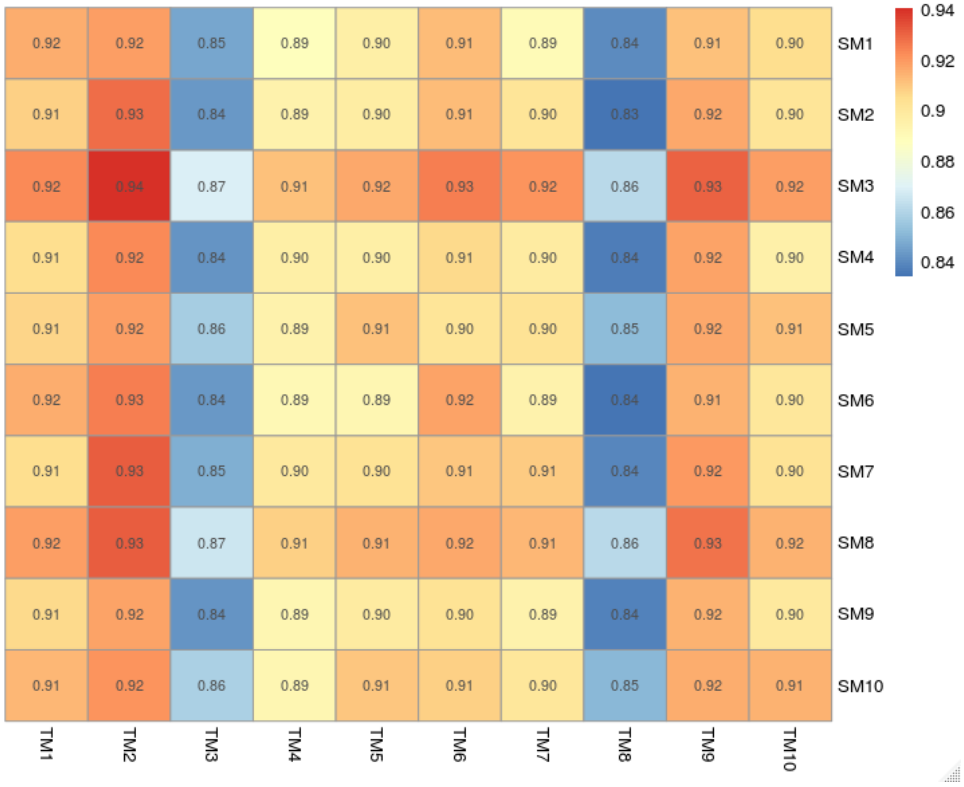

Figure 15: correlation of expression of artificial mixture and actual mixture

It can be seen from the heat map that the correlation between the simulated mixed sample and the real sample is low when the proportion of the sample tumor is high.

3.5 TCGA Data

4 Discussions and future work

List your plan for next week in details. Highlight important issues to take note.
(1) implement algorithm.
(2) create signature.

References

[1] A. R. Abbas, K Wolslegel, D Seshasayee, Z Modrusan, and H. F. Clark. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. Plos One, 4(7):e6098, 2009.

[2] Zeev Altboum, Yael Steuerman, Eyal David, Zohar Barnett-Itzhaki, Liran Valadarsky, Hadas Keren-Shaul, Tal Meningher, Ella Mendelson, Michal Mandelboim, and Irit Gat-Viks. Digital cell quantification identifies global immune cell dynamics during influenza infection. Molecular Systems Biology, 10(2):720, 2014.

[3] Dvir Aran, Zicheng Hu, and Atul J. Butte. xcell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biology, 18(1):220, 2017.

[4] 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(1):218, 2016.

[5] Li Bo, Eric Severson, Jean Christophe Pignon, Haoquan Zhao, Taiwen Li, Jesse Novak, Jiang Peng, Shen Hui, Jon C. Aster, and Scott Rodig. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biology, 17(1):174, 2016.

[6] Shu-Hwa Chen, Wen-Yu Kuo, Sheng-Yao Su, Wei-Chun Chung, Jen-Ming Ho, Henry Horng-Shing Lu, and Chung-Yen Lin. A gene profiling deconvolution approach to estimating immune cell composition from complex tissues. BMC Bioinformatics, 19(4):154, May 2018.

[7] Patrick Danaher, Sarah Warren, Lucas Dennis, Leonard D'Amico, Andrew White, Mary L. Disis, Melissa A. Geller, Kunle Odunsi, Joseph Beechem, and Steven P. Fling. Gene expression markers of tumor infiltrating leukocytes. Journal for ImmunoTherapy of Cancer, 5(1):18, Feb 2017.

[8] Francesca Finotello and Zlatko Trajanoski. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunology Immunotherapy, 67(7):1031–1040, 2018.

[9] T. Gong, N Hartmann, I. S. Kohane, V Brinkmann, F Staedtler, M Letzkus, S Bongiovanni, and J. D. Szustakowski. Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. Plos One, 6(11):e27156, 2011.

[10] Yuning Hao, Ming Yan, Yu Leo Lei, and Yuying Xie. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. bioRxiv, 2018.

[11] Saara Lehmusvaara, Pekka Ruusuvuori, Tapio Visakorpi, and Ilya Shmulevich. Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics, 26(20):2571–2577, 2010.

[12] B. Li, J. S. Liu, and X. S. Liu. Revisit linear regression-based deconvolution methods for tumor gene expression data. Genome Biology, 18(1):127, 2017.

[13] David A Liebner, Kun Huang, and Jeffrey D Parvin. Mmad: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples. Bioinformatics, 30(5):682, 2014.

[14] Shahin Mohammadi, Neta Zuckerman, Andrea Goldsmith, and Ananth Grama. A critical survey of deconvolution methods for separating cell types in complex tissues. Proceedings of the IEEE, 105(2):340–366, 2017.

[15] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods, 12(5):453–457, 2015.

[16] Aaron M Newman and Ash A Alizadeh. High-throughput genomic profiling of tumor-infiltrating leukocytes. Current Opinion in Immunology, 41:77–84, 2016.

[17] Aaron M. Newman, Andrew J. Gentles, Chih Long Liu, Maximilian Diehn, and Ash A. Alizadeh. Data normalization considerations for digital tumor dissection. Genome Biology, 18(1):128, 2017.

[18] Ajit J. Nirmal, Tim Regan, Barbara B. Shih, David A. Hume, Andrew H. Sims, and Tom C. Freeman. Immune cell gene signatures for profiling the microenvironment of solid tumors. Cancer Immunology Research, 6(11):1388–1400, 2018.

[19] Wenlian Qiao, Gerald Quon, Elizabeth Csaszar, Mei Yu, Quaid Morris, and Peter W. Zandstra. Pert: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. Plos Computational Biology, 8(12):e1002838, 2012.

[20] Max Schelker, Sonia Feau, Jinyan Du, Nav Ranu, Edda Klipp, Gavin Macbeath, Birgit Schoeberl, and Andreas Raue. Estimation of immune cell content in tumour tissue using single-cell rna-seq data. Nature Communications, 8(1), 2017.

[21] Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk. Comprehensive evaluation of computational cell-type quantification methods for immuno-oncology. bioRxiv, 2019.

[22] Francesco Vallania, Andrew Tam, Shane Lofgren, Steven Schaffert, Tej D. Azad, Erika Bongen, Meia Alsup, Michael Alonso, Mark Davis, Edgar Engleman, and Purvesh Khatri. Leveraging heterogeneity across multiple data sets increases accuracy of cell-mixture deconvolution and reduces biological and technical biases. bioRxiv, 2017.

[23] Zhong Yi, Ying Wooi Wan, Kaifang Pang, Lionel Ml Chow, and Zhandong Liu. Digital sorting of complex tissues for cell type-specific gene expression profiles. Bmc Bioinformatics, 14(1):89–89, 2013.