A CRITICAL REVIEW OF GENE MARKER SELECTION METHODS AND CELL COUNT INFERENCE TOOLS

by

Muying Wang

Bachelor of Science, China Pharmaceutical University, 2015

Submitted to the Graduate Faculty of

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science in Chemical Engineering

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Muying Wang

It was defended on

March 13, 2017

and approved by

Jason E. Shoemaker, PhD, Assistant Professor, Department of Chemical & Petroleum Engineering

Ipsita Banerjee, PhD, Associate Professor, Department of Chemical & Petroleum Engineering

Robert S. Parker, PhD, Professor, Department of Chemical & Petroleum Engineering

Thesis Advisor: Jason E. Shoemaker, PhD, Assistant Professor, Department of Chemical & Petroleum Engineering

A CRITICAL REVIEW OF GENE MARKER SELECTION METHODS AND CELL COUNT INFERENCE TOOLS

Muying Wang, M.S.

University of Pittsburgh, 2017

Seasonal influenza virus is a threat for human being. Understanding dynamical change of immune response induced by influenza infection could benefit diagnosis and drug development, using transcriptome analysis. But transcriptomic data is often complicated by the changing cell makeup of the tissue during disease. It's difficult to distinguish between gene regulations and cell proliferation or migration. Therefore inference of the change in cell counts is necessary, and computational models for cell count inference are introduced in this thesis. Besides, in most models related to prediction of cell quantities, gene marker selection is used as the first step. Thus computational methodology concerning gene marker selection for cell count inference is also reviewed.

Different gene marker selection methods are applied to a common dataset to evaluate their behaviors. The uniqueness and expression intensity are the key properties for evaluation of obtained markers. As for predicting cell enrichment, principles of three kinds of schemes are explained. Computational algorithms named CTen and CIBERSORT are introduced as examples of them. Estimation behaviors of these tools are tested by a microarray dataset. Analysis of the estimations shows that they may provide good estimation but are not suitable for careful study of complex problems, e.g. dynamical samples.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my advisor, Dr. Jason Shoemaker, for his help and guidance. I am also grateful to the members of my committee for their patience and support. In addition I would like to thank other members in our lab, Emily Ackerman and Robert Gregg, for their kind help in the research. I would also like to thank my parents for their support of my graduate study.

1.0 INTRODUCTION

Seasonal influenza infection is a threat for human beings. Induced by Influenza A (H1N1) viruses, the most recent global pandemic happened in 2009. Yearly influenza viruses could be lethal for people at high risk: people of 65 years old or older, people with chronic diseases, pregnant women or young children [1]. According to Centers for Disease Control and Prevention, H3N2 viruses and H1N1 viruses circulated in the United States during the 2015-2016 influenza season, which caused an estimated 25 million influenza illnesses, and 12,000 pneumonia & influenza (P&I) deaths (Suggested by past data, the total number of influenza-associated respiratory and circulatory (R&C) deaths may be 2-4 times greater than estimates using only P&I deaths) [2]. Though the numbers vary, influenza viruses result in millions of sicknesses, hundreds of thousands of hospitalizations, and thousands or tens of thousands of deaths every year in the United States [3]. Infection with influenza raises challenges in clinical diagnosis and treatment. It is very difficult to distinguish the influenza-viral infection from respiratory illnesses induced by other viral or bacterial causes merely on the basis of symptoms [4]. Although a number of rapid tests are available to detect influenza virus induced illnesses, their accuracy of detection is nevertheless not guaranteed [4] [5]. There are several antiviral drugs available which benefit people at high risk. But they usually work only if treatment starts within the first 2 days of illness [5] [6]. Part of these drugs are not available for young children [5] [7].

Clinical diagnosis and drug development for influenza resulted illnesses are limited because of the lack of characterization of the immune response during influenza infection. Researchers have understood the general procedure of viral-induced immune response, including recognition of virus, activation of lymphocytes and elimination of virus and host cells [8]. However, we are not aware of the precise regulatory pathways and signaling mechanisms happening within and among the associated immune cells. The effects of cellular composition change in the process of immune response are not well explained. We also don't fully understand how the system controls the balance between restrained and excessive immunity [9].

Transcriptome analysis may help to characterize regulations and dynamics of the immune response, and associated influence in infection pathology. DNA microarray and high throughput sequencing technology have been well developed. Their generated data quantify gene expressions of samples in a genome wide. In addition time-series experiments of transcriptome analysis can characterize global dynamical change of a system, biological functions of genes and might indicate the concerning regulation pathways. Plenty of bioinformatics tools have been developed to analyze these transcriptomic data. These tools are not only meaningful in understanding of regulations of immune response, but also could lead to clinical improvements in the future. Potential applications of computational algorithms comprise of faster and more accurate diagnosis of influenza infection, suggestion of personal and highly-targeting treatment, and monitoring of trend in immunopathology (Figure 1). Although these tools are designed for immune response after influenza infection, similar mechanisms may also be applicable in other cases of disease.

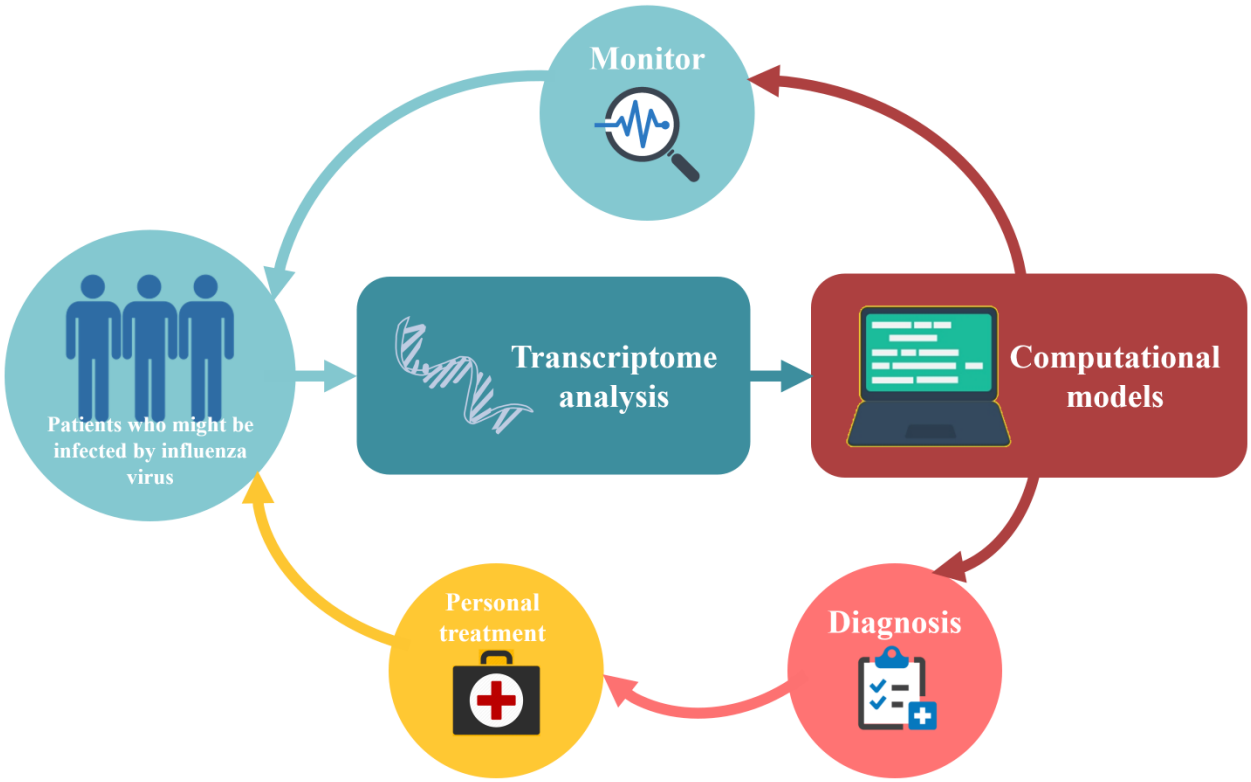

Figure 1. An example of clinical application of computational tools using transcriptomic data in diagnosis and treatment of influenza infection

But dynamical gene expression data of complex samples by transcriptome techniques also bring several confounding factors. One significant confusion is the understanding of gene regulation. Expression signals of genes from virus-infected samples might increase or decrease during a period of time. But this might not be an indication that the genes have been activated or suppressed during the immune response, since changes of cellular composition in the biological sample might also induce the fluctuations of gene expressions. Possible composition change of immune cells induced by influenza infection includes migration, proliferation, and differentiation of lymphocytes [8] [10].

Researchers have been trying to clarify causes of differential expression and identify cellular composition by bioinformatics tools and developing cell count inference algorithms.

Fisher's exact test has been used in comparison of a sample with biological signatures of potential cell types to figure out existing immune cells responsible for differential gene expressions [11] [12] [13]. Clustering analysis of time-course transcriptomic data helps to distinguish different groups of genes with separate expression patterns. Study of each group and associated leukocyte cell types benefits understanding of regulations of immune response [14].

Expression deconvolution algorithms attempt to estimate cell counts or cell-specific gene expression from transcriptomic data. Computational tools using this concept have been utilized to predict cell abundances in different stages of a cell cycle [15], and fractions of distinguished tissues or cells [16] [17] [10] [18]. In recent years they are further developed for more complex situations. Subdivisions of one kind of cell [19] [10] [20], or a cell at different states [19] [20], as well as cell lines in different circumstances [19] could be predicted by deconvolution tools. If applying time-series transcriptomic data, dynamics of immune cells during influenza infection can be de-convoluted as well [10]. The concept of deconvolution is also applicable to analysis of differential expressions, when aware of cell proportions [21] [18]. The general principle of expression deconvolution is to assume the expression profile of a cell mixture is linear to the expression profile of pure cells, and the coefficients are cell fractions based on the whole sample, as shown by the equation below [15].

$$B = A \cdot X$$

The vector of B represents expressions of the cell mixture, with the number of genes as its length. While the matrix of A refers to expressions of pure cells, with the number of genes as the number of rows, and the number of cell candidates as the number of columns. As per the vector of X, it is solved by a linear regression model. The linear regression models that have

been used for expression deconvolution include: linear least squares regression [19], elastic net regularization [10], support vector regression in CIBERSORT [20] and so on [18] [15] [17].

But deconvolution tools also face challenges in construction of the matrix A and application of linear regression method. To construct the matrix A (expression profiles of pure cells) is equal to designating the standard expression behaviors of each cell type. It assumes expression intensities are constant and unified in all cells of the same type/state, which seems inaccurate. And the matrix A can't include unknown cell populations or populations without published transcriptomic data. As per the challenge of linear regression, it often provides negative estimations of cell proportions, or estimations more than 1. This is not realistic, thus for most tools there is an additive step for removal of negative values [19] [21] [20], and sometimes an additional step to normalize estimations to sum to one [20]. Furthermore, common linear regression is reported to work well for samples comprised of 3-4 cell populations or distinguished tissues [17] [18] [19]. But estimations are less perfect for samples of complex compositions [10] [18] [19]. And it provides bad predictions for correlated samples, because it doesn't take connections between the samples into account. This will be proved in the example of CIBERSORT, applying temporal microarray data of lung tissue.

In the present thesis, I evaluate currently existing cell count inference techniques using gene expression data derived from several collections of tissue-derived or artificially mixed cell populations. Principles, procedures and estimation behaviors of two computational algorithms, CTen and CIBERSORT, are carefully analyzed. The temporal microarray data of infected lung tissue [14] are used to evaluate predictions of these tools.

2.0 GENE MARKER SELECTION METHODS

Since gene marker selection is always included as the first step of cell count inference, different gene marker selection methods are also reviewed. Gene markers are genes that could help to identify and distinguish between different tissue/cell types using transcriptomic data. They are essential because they represent expression signatures of different cell types. Gene markers are utilized as reference in comparison with differentially expressed genes in samples to identify existence of cell populations [11] [12] [13]. Or expression intensities of gene markers are used as cell signatures to quantify cell proportions [19] [10] [18] [15] [20] [16] [17]. Gene markers could also give a clue of the core biological functions. Genes are chosen as markers because of their properties of expression intensities. Those generally expressed in all cell types, e.g. housekeeping genes, are beyond our interest for cell identification. Ideally, a good marker is unique for a certain cell type, and with a strong signal. It is highly expressed in one kind of cell and lowly expressed in all other cells. While in reality, this perfect marker is not abundant, especially in comparison of cell lines connected with each other. As a compromise, researchers often use a group of genes as markers to represent a cell type. The combination of their expression intensities forms the signatures of this specific type of cell.

To test performances of different gene marker selection methods, I searched in public microarray data and constructed a dataset associated with 15 immune cells and lung tissue (see Methods). It integrates data of subtypes of lymphocytes, or immune cells in different states.

Formerly reported gene marker selection methods are rebuilt and are applied to the dataset. Uniqueness and intensity level of obtained markers are quantified and compared among different selection methods. To address concerns about how many gene markers to be selected, the connection between threshold value and inference of cell quantities is also examined.

2.1 GENE MARKERS OBTAINED FROM SURFACE MARKERS

Cell surface markers are proteins expressed on the surface of cells and could be used for recognition of cell subtypes [22]. They pioneered study of cell type identification and separation, e.g. FACS, and are still widely used. A cell surface marker's gene could be used as a marker for expression analysis too. Altboum et al [10] developed a tool, DCQ, that uses the gene expression of well-established cell surface markers of immune cells to infer changes in the numbers of immune cells in a sample. Relative cell quantities are estimated by expression profiles of both pure and admixed cells for these markers. In total, 60 surface markers are gathered for DCQ and 41 of them are mapped to genes in our library of pure cells (listed in Appendix A).

Expression profiles of these 41 markers are shown in Figure 1. Several markers are of high intensities in a plenty of cell types, like Cd48, also known as BLAST-1. It is a cell surface marker for T cells, B cells, NK cells, stem cells, macrophage and monocyte [23]. This is not surprising since related cell subtypes, or cells at variant states are often closely related in function. Correlation coefficients of macrophage in two states and DC subtypes are near to 1, and the same situation for stimulated B cells and non-stimulated ones (Figure 2). In conclusion, surface markers are not suitable for separation of dependent or relevant cell subtypes, the same as cells in variant states.

Reversely some markers are of low values in nearly all cell types. A part of them are surface markers of cell types included in DCQ but not in the library of pure cells, such as Cd28, a cell surface marker for pre T cells [10]. Then it wouldn't be strange to find Cd28's expression value is small. But other part of markers, e.g. Cr2, is a marker for both B cells and dendritic cells, but is lowly expressed in all cells. A protein may be an excellent surface markers for experiments, but its coding gene can be a bad gene marker for expression analysis. This weakens further cell quantity estimation by DCQ, because DCQ counts on variation of expressions for different cell types.

In a word, the limited cell surface markers are far from satisfactory in cell identification and count inference.

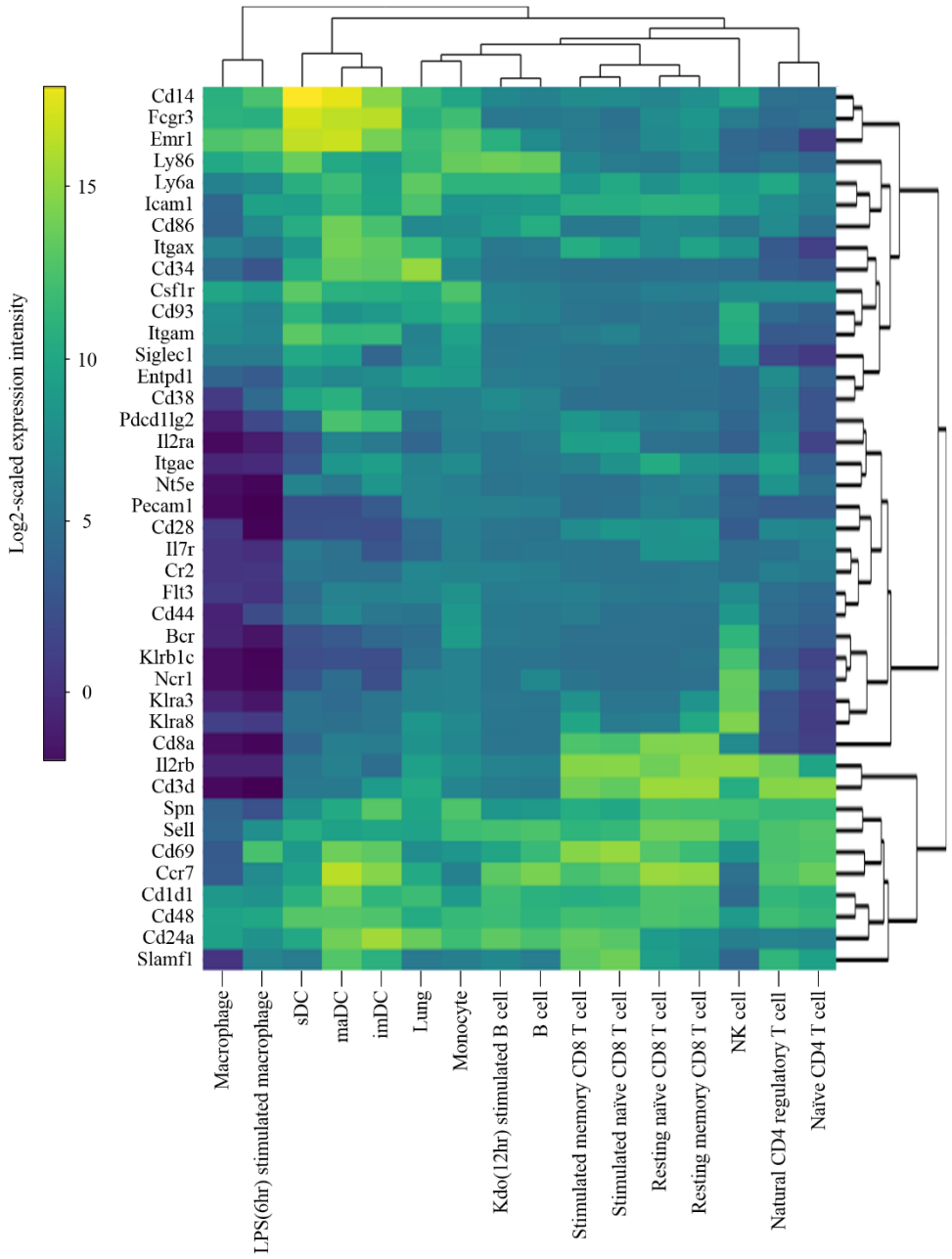

Figure 2. Log-scaled gene expression intensities of 41 cell surface markers for immune cells and lung tissue. 41 cell surface markers are obtained from surface markers used in a tool named DCQ **[10]**. They are mapped to genes in our library of pure cells and referring log-scaled expression intensities are plotted.

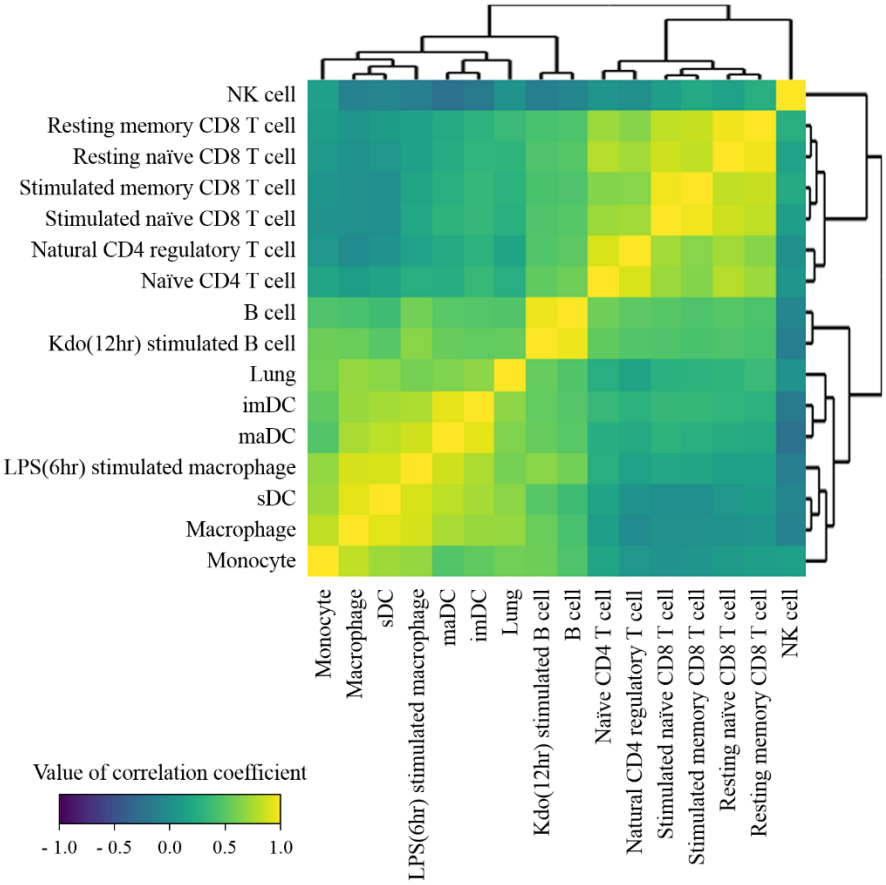

Figure 3. Correlation coefficients of expression intensities of 41 surface markers for 16 cell types. 41 cell surface markers are obtained from surface markers used in a tool named DCQ **[10]**. They are mapped to genes in our library of pure cells for their intensities. Correlation coefficients of these intensity values are calculated for each pair of cell types.

2.2 INTENSITY-BASED GENE MARKER SELECTION

It is straightforward to obtain gene markers by this approach. It assumes that highly expressed genes indicate information of cell functions, which therefore indicate cell types. For example, it's reported that genes of expression values 15 times or 10 times greater than the median are

selected as gene markers for each cell population [13]. This is equal to sorting expressions of a cell type and then applying a threshold.

Unfortunately its disadvantages are: first, deviations of replicates are intensity-dependent for microarray data (less of a problem for RNA-Seq [24]), which could make an effect on further analysis, like deconvolution; besides, similar cell types, e.g. immune cells, might share a large number of genes that are highly expressed, since related or the same cell functions and pathways are involved in these cell types.

To evaluate performance of intensity-based gene marker selection, expression intensities of 15 immune cells and lung tissue are gathered together from separate public datasets, as our library of pure cells. These intensities are sorted from the largest to the smallest for each cell/tissue type. At a threshold of 100, top 100 genes are selected as gene markers for their referring cell type. With the dataset of gene markers for all cell populations, the uniqueness of gene markers for each cell population is analyzed. Since some of the markers show in more than one cell types, gene markers of one cell type is compared with markers of all other cell types. The number of gene markers unique in one specific cell type is computed (unique genes obtained at other thresholds in Appendix A). As displayed in Figure 3, most gene markers are overlapped among different cell types. For 9 cell types out of 16, ratios of unique gene markers out of total 100 markers for each cell type are under 10 percent. The most extreme example is resting memory CD8 T cell, which only have 1 unique gene marker among 100 markers in total. These results agree with our expectations. Some cell types have rare unique markers because they are functionally-related, e.g. subtypes of CD4 and CD8 T cells. While functions of T cells, macrophage, NK cells are less similar, thus easier for separation.

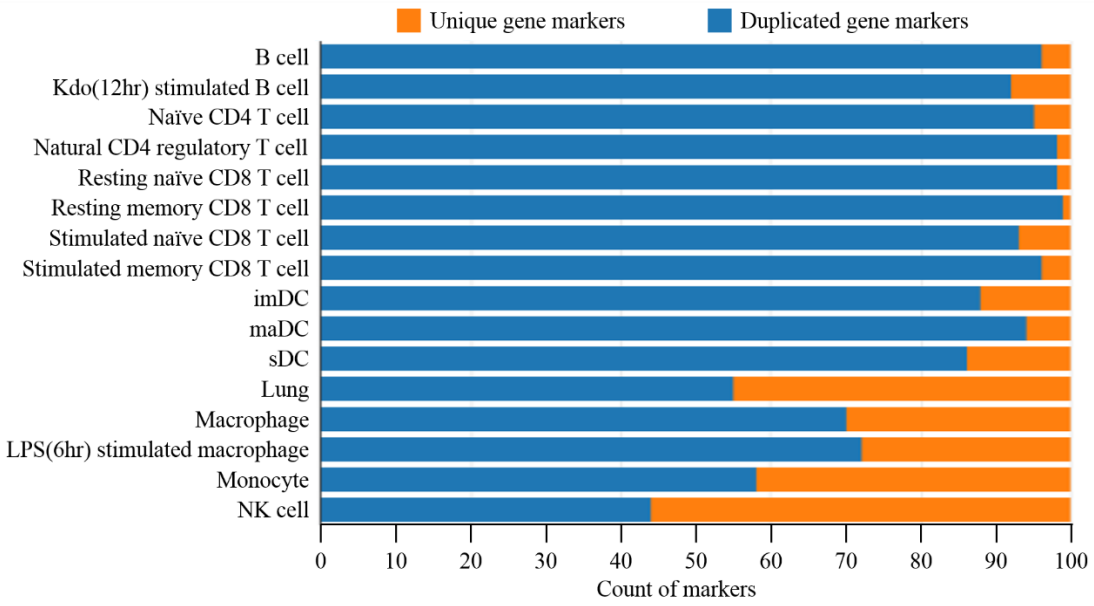

Figure 4. The number of unique gene markers by sorting expression intensities (threshold = 100). Expression intensities of genes are obtained from the library of pure cells. Each cell type's intensities are sorted in a decreasing manner. Top 100 genes are selected as markers for each cell type, and are compared to markers of other cells to compute unique gene markers, indicated as orange bars. Duplicated markers are indicated as blue bars.

2.3 HIGHEST RATIOS

In the intensity-based gene marker selection, intensity values of genes for the same cell type are compared with each other to filter out genes of highest intensities for this specific cell type. But for approaches in this section, intensities of one cell type is compared with intensities of other cell types to find differential expression values. It's reported that ratios of a gene's intensity in one cell population to its intensity in another population are utilized to describe expression variability of this gene among different cell populations [12]. The use of fold change is in the same manner [19]. These methods aim at finding gene markers uniquely highly-expressed in one or two [19] specific kinds of cells. The resulted markers are believed to be more typical and

"personal" for a certain cell type. Although well accepted, these approaches couldn't guarantee intensity level of selected markers, if without limits on expressions. This hurts when performing quantitative analysis based on expressions of gene markers, e.g. deconvolution.

2.3.1 Highest ratios

The first algorithm to introduce is the most elementary one. Ratios of one cell type's expression to the average of all other cell types are calculated and then sorted, based on the library of pure cells (Methods). Gene markers are defined by picking up the highest ratio values.

As shown by Figure 4, there still exist overlapped gene markers among different cell types (results at other cut-offs in Appendix A). Lung is the only cell type to have 100% of uniqueness, and 4 cell types, B cell, Kdo(12hr) stimulated B cell, naïve CD4 T cell and natural CD4 regulatory T cell, show uniqueness of less than 50%. This is because most cell types come from immune system, which share common molecular pathways and biological functions. Lung cells vary a lot with immune cells, so do their expression profiles. Thus it is more obvious to distinguish between lung cells and immune cells. On the contrary, immune cells share similar expression behaviors. There could be a gene that is highly expressed in more than one cell types but lowly expressed in others. But in general, the uniqueness of markers is largely improved.

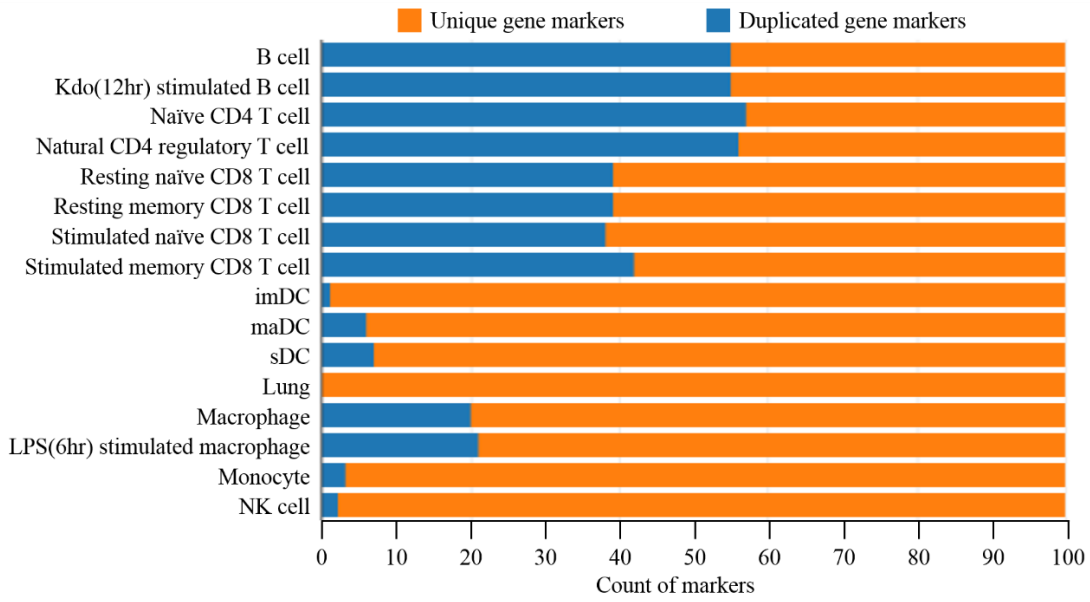

Figure 5. The number of unique gene markers by calculating expression ratios (threshold = 100). For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated, and then sorted in a decreasing manner. Top 100 genes are selected as markers for each cell type, and are compared to markers of other cells to compute unique gene markers, indicated as orange bars. Duplicated markers are indicated as blue bars.

The actual problem is markers' intensities. As discussed before, gene marker selection method is often the first step of cell count inference tools. As inference of cell counts linearly depend on behaviors of expression intensities, the higher the intensities, the more sensitive the inference tool could be. However for the approach of highest ratios, it doesn't give a hint of expression level of these markers. The log-scaled expression values for selected markers vary in a wide range (Figure 5, results of other thresholds in Appendix A). Most are at a reasonable level but a part of their intensities are too low and useless for cell count inference tools.

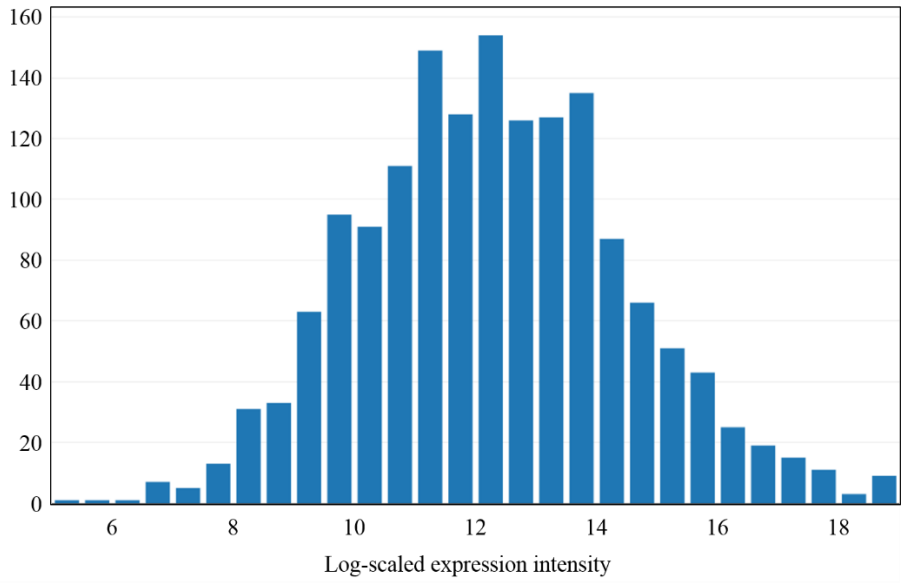

Figure 6. Distribution of expression intensities of all gene markers obtained by calculating expression ratios (threshold = 100). For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated, and then sorted in a decreasing manner. Top 100 genes are selected as markers for each cell type, and their log-scaled intensities are plotted.

2.3.2 Fold change

In the algorithm of the highest ratios, expressions of one cell type are compared with all others. While for the algorithm of the fold change, merely the top expressed cell type is compared with the second cell type. Finally fold change values are sorted, and genes of largest folds are adopted. In this manner, any genes with fold change values are unique markers. Because among expressions of all cell types for a gene, there's solely one highest expressed cell type. And only this cell type is given a fold value to characterize how the gene is differentially expressed in all cell types.

But out of the gain comes its loss. The numbers of resulted gene markers at a threshold of 1 are plotted (Figure 6). The NK cell could have up to 1514 markers in total but the B cell and

the naïve CD4 T cell have as low as 7 markers. If to further estimate cell counts by these markers, different cell types have to be weighed to the same degree, which means no more than 7 gene markers are available for each cell type (112 for all 16 cell types).

What's more, the highest expressed cell population for a gene may not show a high expression intensity value. The log-scaled expression values of obtained markers are aggregated in a lower level, comparing to results of highest ratios. The quality of these markers are questionable.

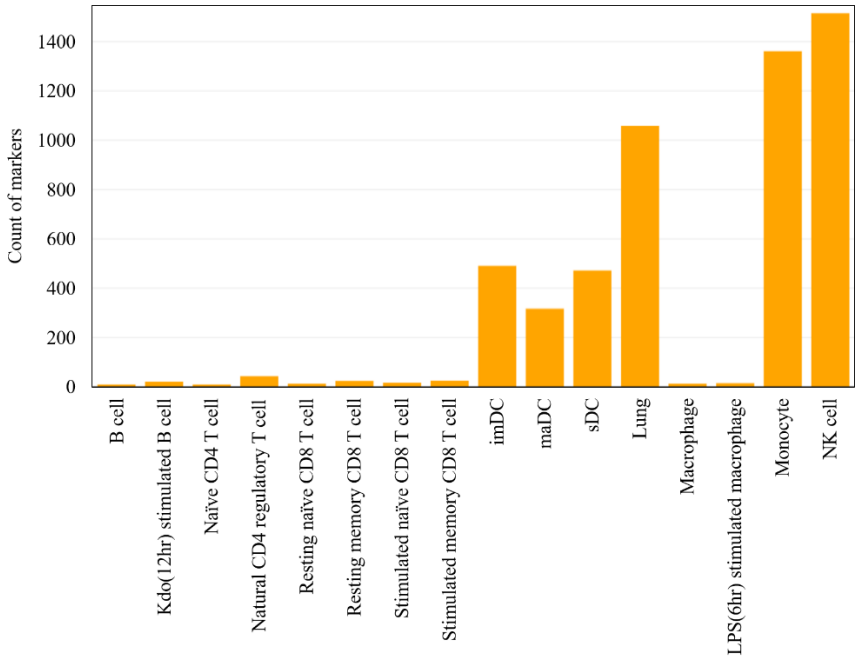

Figure 7. The number of gene markers obtained by sorting fold changes (threshold = 1). Each gene's highest expressed cell population is compared with the next highest expressed population, and referring fold change is calculated. Fold change values under 1 are removed, and remained genes are selected as markers. The numbers of obtained markers for each cell type are as the yellow bars.

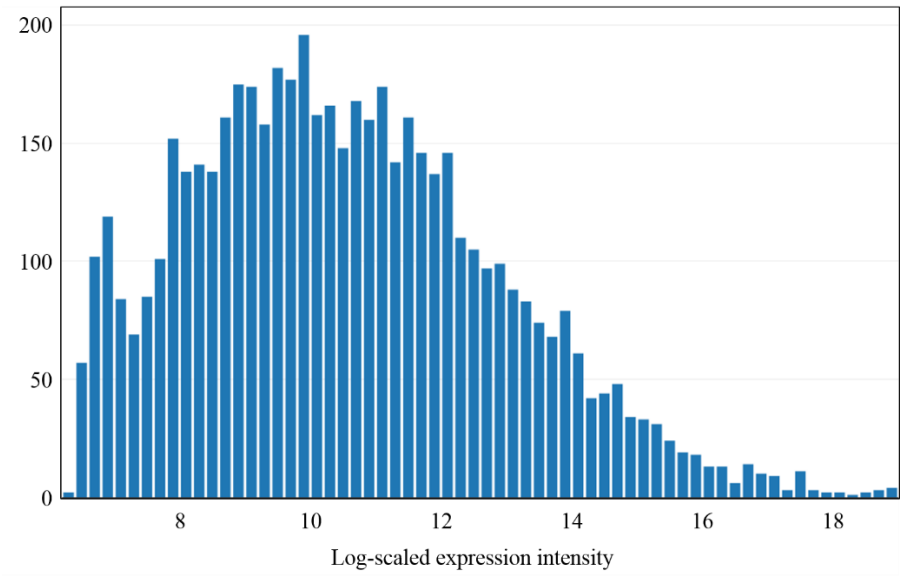

Figure 8. The distribution of expression intensities of gene markers obtained by sorting fold changes (threshold = 1). Each gene's highest expressed cell population is compared with the next highest expressed population, and referring fold change is calculated. Fold change values under 1 are removed, and remained genes are selected as markers. The distribution of their log-scaled intensities are plotted.

2.4 COMBINATORIAL METHOD

To remove genes of low expression values and better serve examples of deconvolution in Section 3, a simple improvement of "highest ratios" is: to filter out genes of expressions beneath a lower bound. In this thesis, genes of log2-scaled expression intensities below 8 are removed. Resulted markers show that most unique markers are saved (Figure 8), while successfully improves the general expression intensity level (Figure 9). The generated gene markers of 16 cell lines in the library are implemented in the example of CIBERSORT in Section 3.

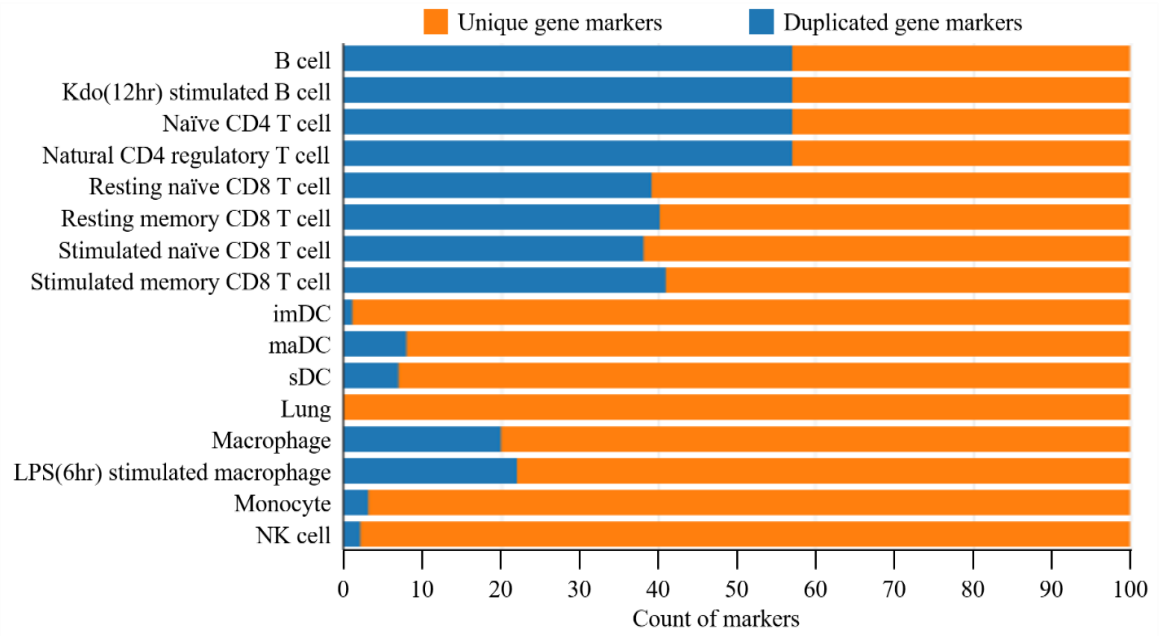

Figure 9. The number of unique gene markers and duplicated markers obtained by combinatorial method (threshold = 100). First step of combinatorial method is the same with the approach of highest ratio. For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated. Then genes of log2-scaled expression intensities below 8 are removed. Finally the ratios of remained genes are sorted in a decreasing manner. Top 100 genes are selected as markers for each cell type, and are compared to markers of other cells to compute unique gene markers, indicated as orange bars. Duplicated markers are indicated as blue bars.

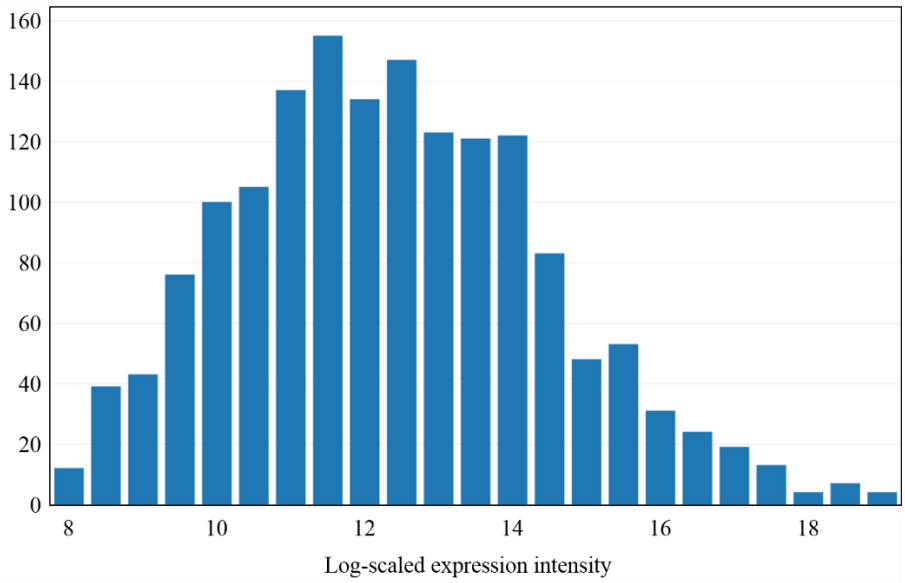

Figure 10. The distribution of expression intensities of gene markers obtained by combinatorial method (threshold = 100). First step of combinatorial method is the same with the approach of highest ratio. For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated. Then genes of log2-scaled expression intensities below 8 are removed. Finally the ratios of remained genes are sorted in a decreasing manner. Top 100 genes are selected as markers for each cell type, and distribution of their log-scaled intensities are plotted.

2.5 STATISTICAL METHODS

Lastly, statistical tests have are widely used to obtain differentially expressed genes. These resulted genes could also be regarded as markers of associated samples. When applying samples of single cell types, statistical tests could help define gene markers too. Wang et al reported to use two-tailed student *t*-test as part of criterion for marker selection [16]. Commonly used statistical tools in analysis of expressions are student *t*-test, FDR, Welch *t*-test and so on. They are so well known that details will not be discussed in this thesis.

2.6 THRESHOLD SELECTION

A significant question is ignored for convenience when explaining the gene marker selection methods above: how many gene markers should be selected? For cell surface markers, the amount of well accepted markers are limited thus all markers are favored. But in other algorithms, intensity-based gene marker selection, highest ratio and statistical method, the number of gene markers is set manually. This might strongly influence markers' quality.

The maximum number of markers that fold change offers is illustrated in Figure 10. However applying a threshold of 5 for fold changes removes all potential markers of 5 cell types (Figure 11). The numbers of gene markers at different cut-offs by sorting expression intensities are also calculated, as shown in Appendix A. If the cut-off is 10, the top 10 highest expressed genes will be selected as markers for the specific cell type. Comparing the amount of markers at cut-off of 20 with 10, the fraction of unique gene markers of most cell types seems to increase as the increment of cut-off value. However for several cell types, the fraction actually decreases. In comparison of 50 with 20, the fraction of unique markers decreases although for a small amount of cell lines the fraction increases. In general, when above 50, the proportion of unique gene markers decreases as the cut-off value grows.

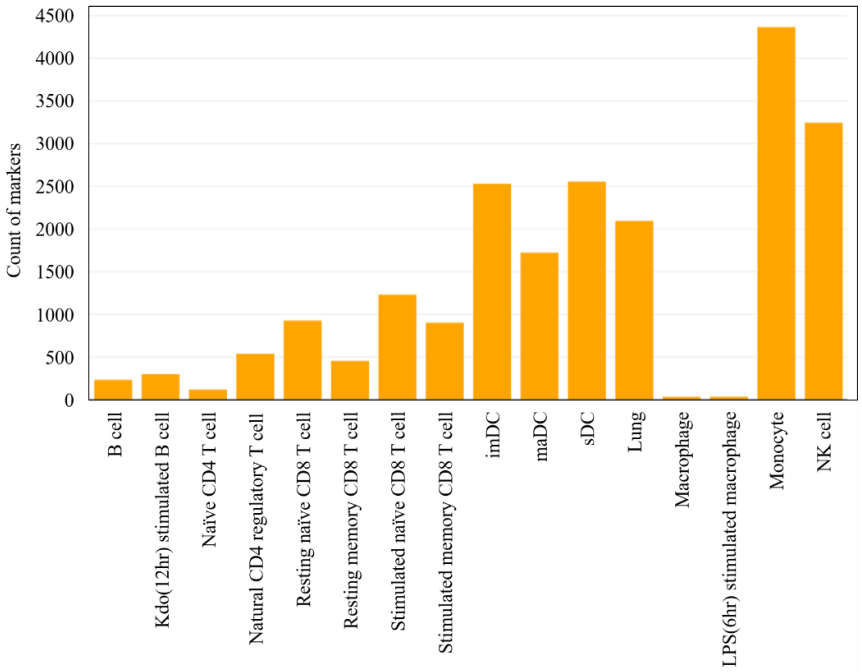

Figure 11. The number of gene markers obtained by sorting fold changes (threshold = 0). Each gene's highest expressed cell population is compared with the next highest expressed population, and referring fold change is calculated. The numbers of obtained markers for each cell type are as the yellow bars.

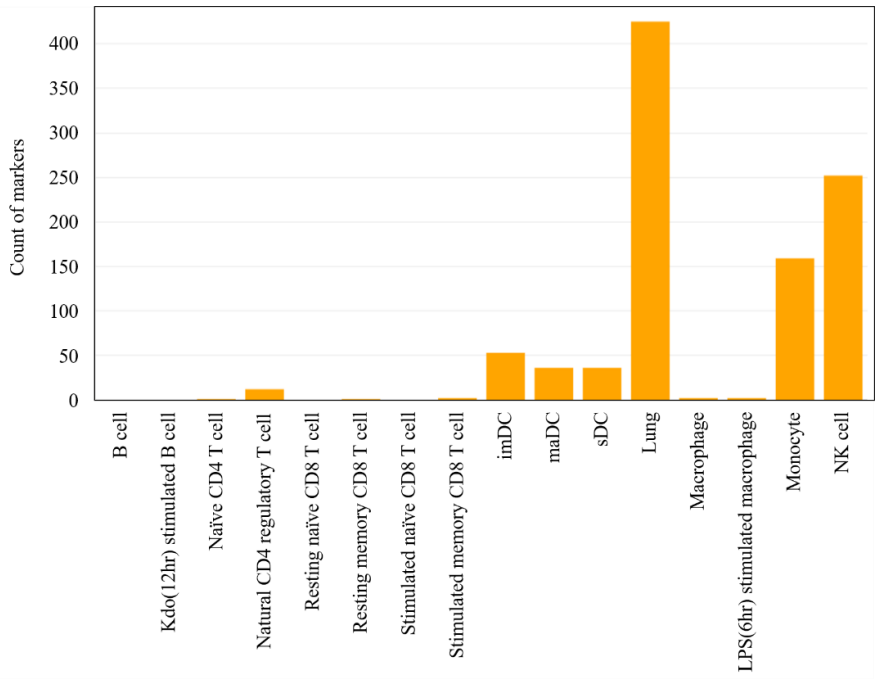

Figure 12. The number of gene markers obtained by sorting fold changes (threshold = 5). Each gene's highest expressed cell population is compared with the next highest expressed population, and referring fold change is calculated. Fold change values under 5 are removed, and remained genes are selected as markers. The numbers of obtained markers for each cell type are as the yellow bars.

As a result, there seems to exist an optimal threshold to obtain the largest fraction of unique markers, which might have the best behavior for further analysis. It's reported that for deconvolution analysis, minimizing condition number gives the optimum. The matrices of gene markers' expression values under different thresholds are generated. Then their condition numbers are sorted to find out the minimum one. The referring threshold and matrix will be adopted [19] [20]. Condition number is popular because in deconvolution, linear relation between expression intensities of cell mixtures (the vector B in the equation below) and pure cells (the matrix A) are assumed. Condition number is a property of matrix A. It's calculated as the product of the 2-norm of the matrix and the 2-norm of its inverse (or pseudo-inverse), and is

equal to or greater than 1. It's meaningful because it approximately tells how big the error of the estimated X (cell fractions in deconvolution problem) in comparison with error of vector B. If condition number is just a little larger than 1, then matrix A could be well inversed, and the error of X is not tremendously increased by the error of B. Otherwise, matrix A can't be well inversed and the error of X might dramatically increases as the increase of B. But it's hard to set a boundary of the value of condition number to define a bad condition number.

$$B = A \cdot X$$

In order to prove whether condition number helps to achieve the optimum, a small-scale of deconvolution is implemented. First, the expression profiles of pure liver, brain and lung tissues and their mixtures are obtained from the literature [21]. Next gene markers and associated condition numbers are computed at different thresholds. Expression profiles of these markers are then used for deconvolution. The resulted cell proportions are compared with cell abundances reported in the literature to calculate correlation coefficients. Condition numbers and correlation coefficients at threshold of 1 to 1000 are plotted in Figure 12 and Figure 13. The condition number decreases to the minimum, oscillate a little bit and then keeps increasing. In fact, the variation of condition number under different thresholds is very tiny in this instance. On the other hand, correlation coefficient rapidly increases, oscillates and then turns stable at a high level. The minimum condition number doesn't give minimum correlation coefficient. A high condition number might not be significant for deconvolution. Although condition number gives guidance to threshold selection, deconvolution behavior is not entirely correlated to it.

To ensure stability of both matrix of expressions and deconvolution results, and considering efficiency of computation, a cut-off of 100 is applied in this thesis.

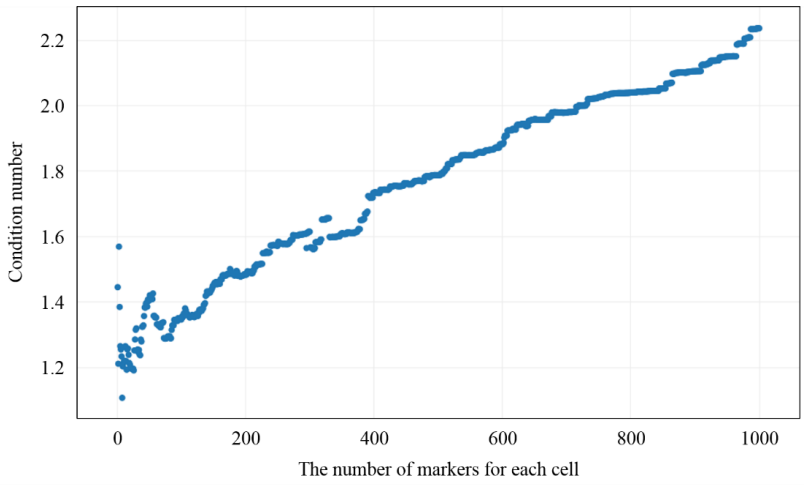

Figure 13. The condition number of the matrix of gene markers' intensities (threshold = 1-1000). Combinatorial method is applied to expression profiles of pure liver, brain and lung tissues to obtain potential gene markers. Then markers for each tissue type are obtained at different thresholds, from 1 to 1000. The associated condition numbers of the matrix of gene markers' intensities are calculated.

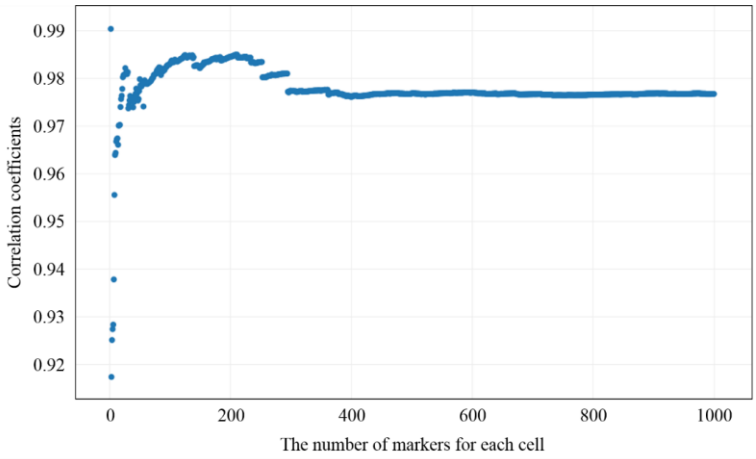

Figure 14. The correlation coefficients of deconvolution results at different thresholds (threshold = 1-1000). Combinatorial method is applied to expression profiles of pure liver, brain and lung tissues to obtain potential gene markers. Then markers for each tissue type are obtained at different thresholds, from 1 to 1000. The expression profiles of 3 kinds of tissues are used for deconvolution. The referring results are compared with actual proportion of these tissues, and correlation coefficients are calculated.

2.7 SUMMARY OF APPROACHES ABOVE

Table 1. Properties of gene markers obtained by different methods

Gene marker selection method	Behavior of obtained gene markers		
	Intensity	Uniqueness	Number of obtained gene markers
Cell surface markers	In a wide range	Low	Small
Intensity-based gene marker selection method	High	Low	Depend on thresholds
Highest ratio	In a wide range	High	Depend on thresholds
Fold change	In a wide range	High	In a wide range
Statistical method	In a wide range	High	Depend on thresholds

3.0 EXAMPLES OF CELL COUNT INFERENCE TOOLS

Cell count inference tool is a computational algorithm to predict existence and abundance of cell populations (including tissue types, and cells at different states). It is on the basis of transcriptome analysis, since microarray or RNA-Seq data of samples, or differentially expressed genes generated from these data are required as input. In this section, two cell count inference tools will be introduced: CTen [13] and CIBERSORT [20]. CTen compares differentially expressed genes, or clustered genes with its own database of gene markers for different cell/tissue types, indicating cell/tissue types associated with these genes and relative enrichments of the cells or tissues. CTen is especially valuable when lack of knowledge of samples although it can't provide exact cell quantities. While CIBERSORT utilizes microarray or RNA-Seq data of samples and user-generated dataset of possibly existing cell types to quantify related cell proportions in the samples. Like many other deconvolution tools, CIBERSORT is applicable only when aware of potential cell composition and given enough data of pure cells. Its estimation depends on quality of the data.

The time-series microarray data of infected mouse lung tissue are applied to both CTen and CIBERSORT. They both accurately predict existence of important immune cells. However, CIBERSORT can't compute connections between samples at different time points, which gives rise to discontinuity and oscillation in its final estimations, which will be described in details in this section.

3.1 CTEN

As mentioned in Introduction, CTen uses a Fisher's exact test to associate a set of genes with a specific cell type [13]. It recommends dynamic clustering to identify the tested gene sets. Its database is comprised of 96 mouse and 84 human tissue/cell types. User's list of differentially-expressed genes, or preferentially dynamically clustered gene sets, is uploaded to the website for comparison with this database. Enrichment scores of each cell type are computed as –log10 of BH-adjusted p-values.

CTen could be used in an elementary level and an advanced level.

3.1.1 The elementary level

The elementary case of application is to upload a sample's differentially expressed genes, which are received by any statistical test, e.g. student T test and FDR. Enrichment scores of possibly existing cell populations will be returned.

To simply test CTen's performance, top 100 markers for B cells, generated by combinatorial method, are uploaded. The full table of results are listed in Appendix B. Enrichment scores above 2 are recommended as significant [13]. The most enriched cells predicted by CTen are B cells from different lineages or subtypes under different states, with the No.1 as "Follicular B cells". Spleen and lymph nodes are also among the top 10 enriched tissue/cell types. CTen accurately predicts existence of B cells.

3.1.2 The advanced level

CTen also suggests a workflow for advanced use-case [13](Figure 14). When studying dynamics of a complex biological process, e.g. dynamical changes of immune system after infection, this workflow helps to distinguish between regulations of genes and changes of cell populations. In an immune process, there could happen migration, proliferation, differentiation and activity transition. These dynamical changes in the cell level may enormously influence referring expression profiles.

Reported by Shoemaker et al, murine lung tissues infected by different influenza strains are analyzed by microarray, at separate time points [14]. After normalization and quality control, differential expression is assessed by application of a linear model [14]. Probes of fold-change less than 2 for infected-to-control comparison are all removed before clustering analysis. Different settings are reported to be tested for robust clustering by WGCNA, resulting in 45 co-expression modules. In these modules, a submodule of N2 is analyzed by CTen. The most enriched cell types are LPS-stimulated macrophages (details in Appendix B). This implies that genes in this submodule of N2 are associated with migration and activation of macrophages during immune response of infected lung tissue.

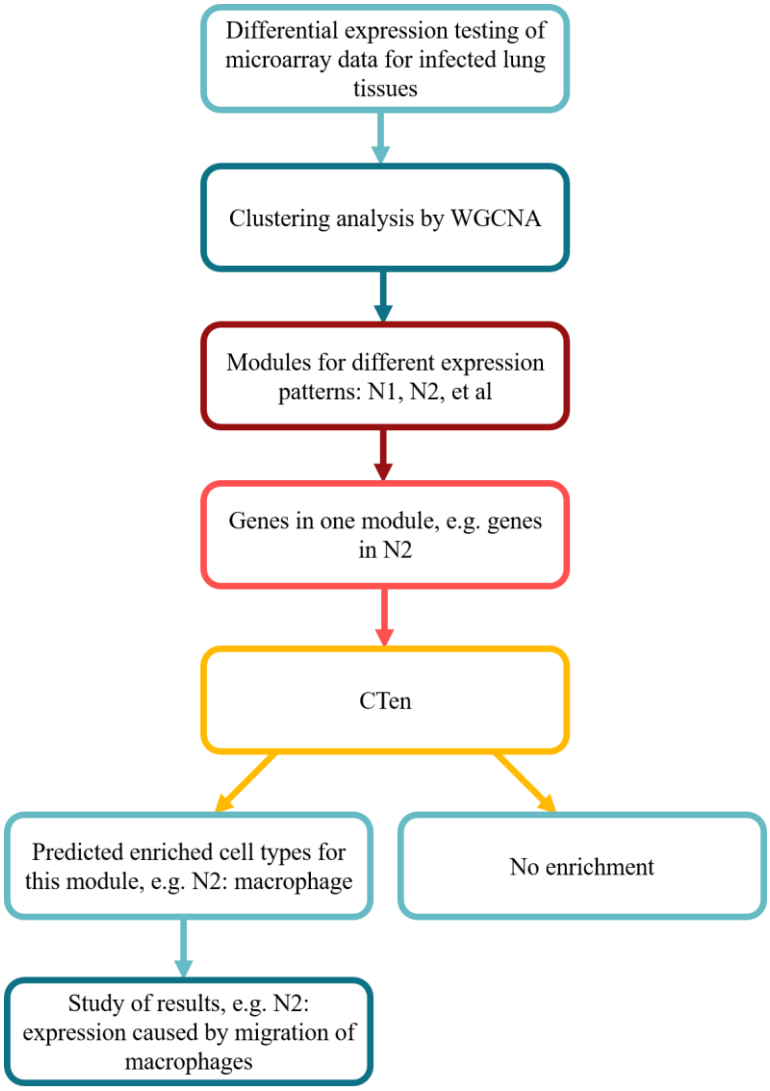

Figure 15. The workflow of CTen for advanced use-case **[13] [14]**

CTen presumes that the whole picture of cellular composition is unknown, thus it is not able to predict quantitative cell abundances for careful study. But CTen is very helpful when tissue or cell types in the sample are beyond researcher's knowledge. It could satisfy needs of understanding relative change of environment in the cell level. But in this advanced case, results of CTen depends on pre-analysis of differentially expressed genes implemented by WGCNA in the example.

3.2 CIBERSORT

As discussed in the last section, CTen is not capable of providing cell quantities. CIBERSORT is a deconvolution tool developed by Newman et al [20]. Given expression profiles of pure cells and cell mixtures, CIBERSORT could predict cell proportions.

First, files of gene markers' expressions for pure cells, and cell mixtures are required as inputs. The former one is named Signature Matrix in CIBERSORT, and the latter one named Mixture. These two matrices are then normalized to zero mean and unit variance for preprocessing. Then nu-support vector regression (ν-SVR), a machine learning tool is applied for optimization to compute regression coefficients. If there exist negative coefficients, they will be set to zero, while other coefficients will be normalized to sum to one. The resulted coefficients are provided as proportions of associated cell populations (Figure 15).

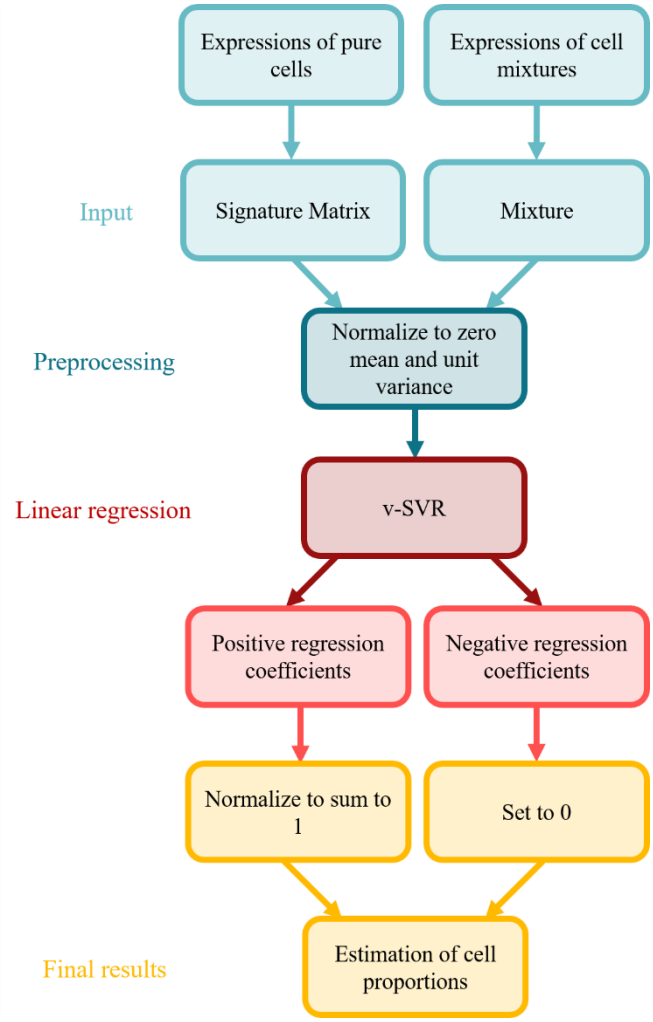

Figure 16. The algorithm of CIBERSORT **[20]**

To promote efficiency of deconvolution analysis, merely a part of the entire expression profiles of pure and admixed cells are performed in CIBERSORT. In this thesis, temporal expression profiles of infected murine lung tissues are gathered from the research by Shoemaker et al [14]. Gene markers are defined by combinatorial method, using the library of pure cells. These markers' expression intensities in the library of pure cells and lung samples are uploaded to CIBERSORT, serving as Signature Matrix and Mixture respectively. Computed cell fractions of replicates are averaged at each time point.

As expected, the proportion of lung tissue keeps decreasing during the immune response to infection (Figure 16). Original lung tissue is damaged after infection. And at the same time, proliferation of existing lymphocytes and migration from other murine tissues increase the total number of cells in the samples. The abundance of macrophages grows rapidly (macrophage LPS-6hr as Figure 17 and macrophage as Figure 27(a) in Appendix B). Macrophage LPS-6hr is predicted to be the most enriched cell type. This matches with prediction by CTen [14], as discussed in the example of advanced use-case. Thus the proportion of lung tissue in samples is reduced.

But surprisingly, the enrichment of lung tissue turns to be 20% - 40% of total cells in samples of pH1N1 and H5N1, since 72 hours after infection. Simultaneously, macrophage LPS-6hr turns to 40% - 60% of total cells. Such dramatic decrease of lung tissue and increase of macrophage LPS-6hr haven’t been reported before. This behavior doesn’t seem realistic since mice could hardly survive with strongly weakened lung tissues.

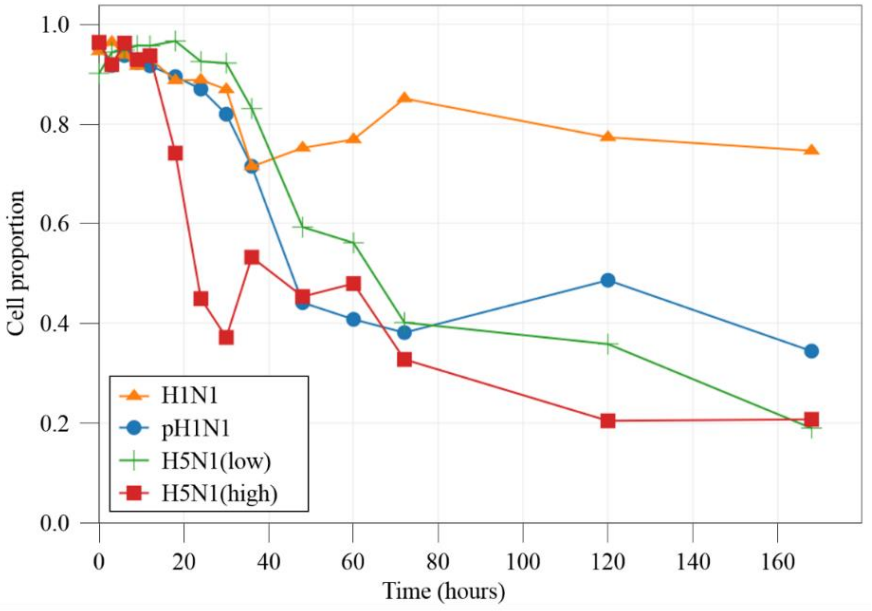

Figure 17. Proportions of lung tissue at 14 time points for different virus strains. Microarray data of murine lung tissues, infected by H1N1, pH1N1, H5N1 of low and high dose at different time points are obtained from literature **[14]**. The library of pure cells serves as Signature Matirx. Both datasets are applied to CIBERSORT to estimate cell proportions. Estimated proportions of 3 replicates at 14 time points, 0h, 3h, 6h, 9h, 12h, 18h, 24h, 30h, 36h, 48h, 60h, 72h, 120h, 168h, are averaged respectively.

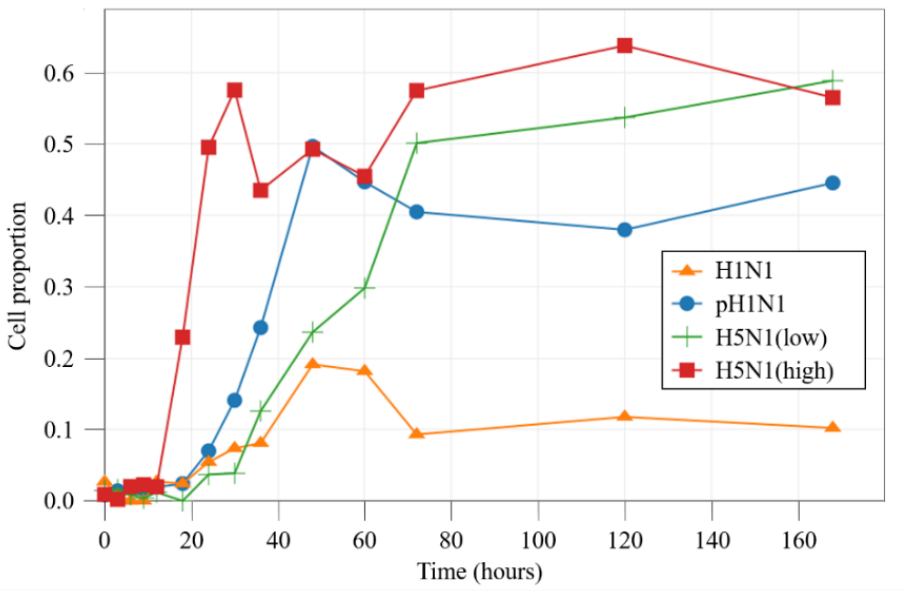

Figure 18. Proportions of macrophage LPS-6hr at 14 time points for different virus strains. Microarray data of murine lung tissues, infected by H1N1, pH1N1, H5N1 of low and high dose at different time points are obtained from literature **[14]**. The library of pure cells serves as Signature Matirx. Both datasets are applied to CIBERSORT to estimate cell proportions. Estimated proportions of 3 replicates at 14 time points, 0h, 3h, 6h, 9h, 12h, 18h, 24h, 30h, 36h, 48h, 60h, 72h, 120h, 168h, are averaged respectively.

As shown by Figure 18, the dynamical change of resting memory CD8 T cells is compatible with common understanding of memory T cells, which are largely proliferated at the last stage of immune reponse after infection. Enrichment of resting memory CD8 T cells at the beginning could be neglected. However it is driven to a high level especially after 120 hours, predicted as 5% - 12% by CIBERSORT. The amount of memory T cells has rarely been reported before. It's questionable whether resting memory CD8 T cells could be as much as 12% of total cells.

The NK cells are not abundant in samples of infected lung tissues. The largest estimation of NK cells is less than 1% of total cells for all virus strains. Figure 19 shows a peak and also the

maximum of enrichment for each strain within the first day after infection. Dynamics of NK cells in the first 24 hours haven't been well studied before, thus prediction of the peak can't be demonstrated. Most researches are based on level of days. According to Gazit et al [25], NK cells in the lung keep increasing in a period of 5 days after infection by influenza A PR8 virus. This is contradictory to estimation by CIBERSORT in the period from Day 1 to Day 7.

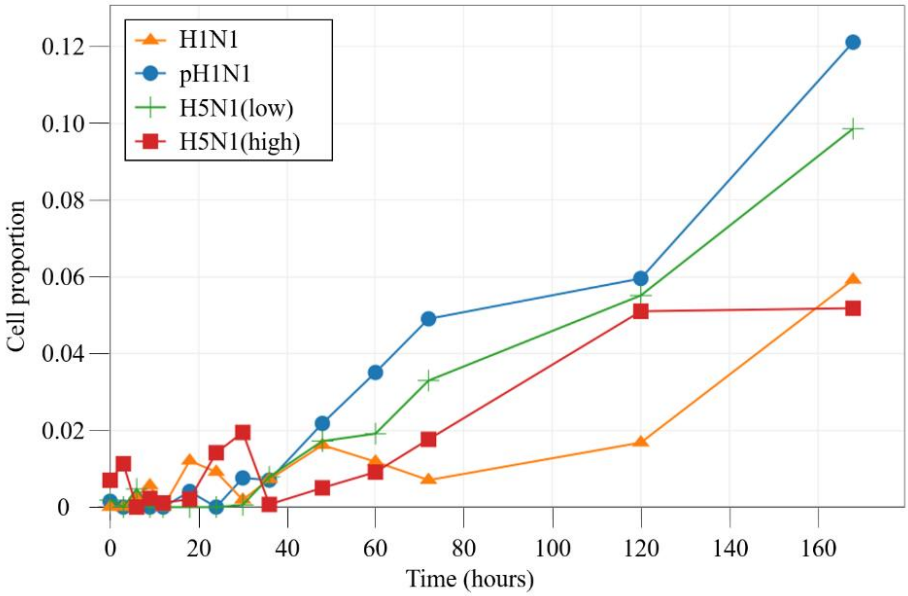

Figure 19. Proportions of resting memory CD8 T cells at 14 time points for different virus strains. Microarray data of murine lung tissues, infected by H1N1, pH1N1, H5N1 of low and high dose at different time points are obtained from literature **[14]**. The library of pure cells serves as Signature Matirx. Both datasets are applied to CIBERSORT to estimate cell proportions. Estimated proportions of 3 replicates at 14 time points, 0h, 3h, 6h, 9h, 12h, 18h, 24h, 30h, 36h, 48h, 60h, 72h, 120h, 168h, are averaged respectively.

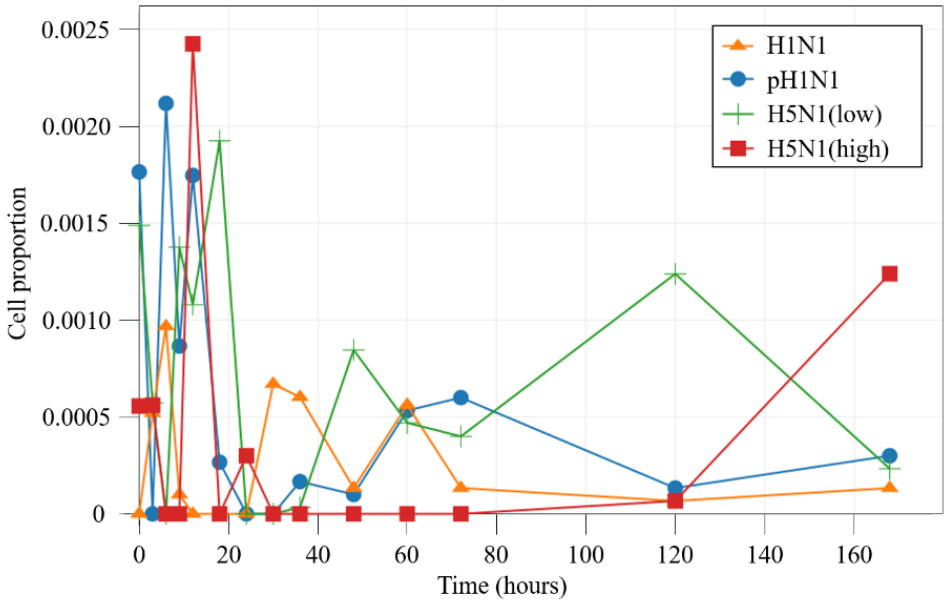

Figure 20. Proportions of NK cells at 14 time points for different virus strains. Microarray data of murine lung tissues, infected by H1N1, pH1N1, H5N1 of low and high dose at different time points are obtained from literature **[14]**. The library of pure cells serves as Signature Matirx. Both datasets are applied to CIBERSORT to estimate cell proportions. Estimated proportions of 3 replicates at 14 time points, 0h, 3h, 6h, 9h, 12h, 18h, 24h, 30h, 36h, 48h, 60h, 72h, 120h, 168h, are averaged respectively.

Cell dynamics of other immune cells are listed in Appendix B. In all of these plots, fractions of dynamical cell abundances vibrate intensely. In reality, cells can't be entirely eliminated and suddenly grow to the peak. Cell fractions at different time points are dependent on each other instead of being irrelevant. So CIBRSORT is not suitable for temporal estimation because it completely neglects the connections.

Different from Fisher's exact test, CIBERSORT and other deconvolution tools provide quantitative predictions of cell enrichment. This is their strong suit. However to achieve this, detailed information of every possible cell type is required. First of all, this means cell lines not well studied are not applicable at all. Even a qualitative estimation is not available by

CIBERSORT. Except for unknown cell types, the expression profiles of well-known cell types might also be of good/bad quality. Since deconvolution results largely depend on these expression values, reproducibility of predicted cell enrichment might be low when changing expression profiles.

4.0 CONCLUSIONS AND FUTURE WORK

4.1 CONCLUSIONS

For the gene marker selection methods, quality of obtained markers are analyzed by applying a microarray dataset, in the aspects of uniqueness and intensity level. As a result, cell surface markers couldn't guarantee neither intensity level nor uniqueness of markers. The intensity-based selection approach provides markers of high intensities, while the number of unique markers is not satisfactory. The highest ratio approach seems to be much better than other approaches. The uniqueness of markers is ensured and the intensities are generally acceptable, although a part of the gene markers' expression intensities are too low. Threshold selection might also significantly effects the quality of markers. There seems to exist an optimal threshold, and condition number is reported to achieve this optimum. While we demonstrate that condition number merely gives evidence for threshold selection, and is not directly related to deconvolution results. For most cell count inference tools, a gene marker selection is usually included or suggested as the first step [19] [10] [20] [13] [16]. Because gene markers provide a concise version of the information of pure cells, which removes noise and improves efficiency of study.

As per any kind of cell count inference tools, the principle is to compare transcriptomic data of a sample with those of pure cells. Therefore information of pure cells is indispensable.

CTen is capable of analysis of both genes expressed in different conditions, and dynamic clustering of time-course transcriptomic data. It accurately predicts existence of cell populations. While it doesn't provide quantitative predictions since it assumes cellular composition of a sample is not entirely known. Another scheme, deconvolution provides quantitative predictions. It performs great for simple combination of cell populations. However as the sample becomes complex, e.g. dynamical microarray data of a tissue, part of the predictions might be completely inaccurate.

4.2 FUTURE WORK

We see from the foregoing discussion that simple deconvolution tools neglect connections between time-series transcriptomic data, and give rise to discontinuity of their predictions of cell dynamics. But dynamical change of immune cell populations is one of our main concerns, and we want to understand their roles in infection pathology. Therefore, we are building a novel model to better predict cell dynamics of temporal samples.

Based on the original assumption of deconvolution concept, expression profiles of cell mixtures are linear to expression profiles of pure cells, with coefficients as the estimated cell proportions (see the equation below). For time-series transcriptomic data, B becomes a matrix of temporal transcriptomic data, whose rows relate to different genes and columns relate to different samples. X is also a matrix, which contains estimations of associated cell populations at different time points. We further postulate dynamics of a specific cell population is a function of time, e.g. $x = k_1 t^3 + k_2 t^2 + k_3 t^1 + k_4 t^0 + k_5 t^{-1} + k_6 t^{-2}$, and the matrix X could be described as the multiplication of a matrix K and a vector T (shown in the following equation), where T contains

variations of time (e.g. t^3, t^2, …) and K is a collection of parameters for each cell type and each element in T (e.g. k_1, k_2, …). With given A, B and T, as long as K is estimated, proportions of each cell type at each time point will be obtained.

$$B = A \cdot X$$

$$X = K \cdot T$$

If the model above is demonstrated effective, it can be helpful in discovering significant genes as well. In the analysis of the simple deconvolution example (deconvolution of liver, lung and brain mixtures) in Section 2.6 Threshold Selection, we note the estimation results might be stable when the number of gene markers is large enough (Figure 14). Applying our new model to temporal samples making use of different amount of gene markers, we might figure out which genes significantly deviate the estimations.

5.0 METHODS

5.1 LIBRARY OF PURE CELLS

Microarray data of 16 cell types are gathered in this library. These cell types are: B cell [26], Kdo(12hr) stimulated B cell [26], naïve CD4 T cell [27], natural CD4 regulatory T cell [27], resting naïve CD8 T cell [28], resting memory CD8 T cell [28], stimulated naïve CD8 T cell [28], stimulated memory CD8 T cell [28], imDC [29], maDC [29], sDC [29], lung [14], macrophage [30] [31], LPS(6hr) stimulated macrophage [31], monocyte [32], NK cell [33].

To be compatible with infected lung tissue data [14] analyzed by deconvolution tools in this thesis, only microarray analyses implemented under procedures of Agilent-014868 Whole Mouse Genome Microarray 4x44K are considered candidates. When there are more than one datasets available for a cell type, cluster analysis of them is done to compare their quality. After datasets of different cell types are settled down, annotation is unified and data are non-log scaled. At last, replicates of each cell types are averaged to form the final library of pure cells.

Table 2 Sources of microarray data for library of pure cells

Cell type	GEO accession	Platform for microarray
B cell	GSE23620	GPL7202
Kdo(12hr) stimulated B cell	GSE23620	GPL7202
Naïve CD4 T cell	GSE17166	GPL4134
Natural CD4 regulatory T cell	GSE17166	GPL4134
Resting naïve CD8 T cell	GSE16145	GPL4134
Resting memory CD8 T cell	GSE16145	GPL4134
Stimulated naïve CD8 T cell	GSE16145	GPL4134
Stimulated memory CD8 T cell	GSE16145	GPL4134
imDC	GSE31273	GPL7202
maDC	GSE31273	GPL7202
sDC	GSE31273	GPL7202
Lung	GSE63786	GPL7202
Macrophage	GSE16180, GSE20207	GPL4134
LPS(6hr) stimulated macrophage	GSE20207	GPL4134
Monocyte	GSE14850	GPL7202
NK cell	GSE30629	GPL4134

5.2 GENE MARKER SELECTION METHODS

5.2.1 Cell surface markers

60 cell surface markers of immune cells are obtained from the dataset used in a deconvolution tool named DCQ, as reported by Altboum et al [10]. 41 of them are mapped to the library of pure cells and expression values of these markers are gathered for analysis (listed in Appendix A).

5.2.2 Intensity-based gene marker selection method

Expressions of each cell type are sorted in a decreasing manner. Then each probe is mapped to referring gene symbol (if available). With NAs and duplicates removed, final list of gene symbols for a cell type could be less than the number of original probes. Finally different thresholds, t = 10, 20, 50, 100, 200, 500, are set up manually to obtain lists of gene markers for every cell type.

Duplicated gene markers are obtained by comparing markers of one cell type with all other cell types. Thus unique genes are equal to the number of total gene markers subtracted by duplicated markers. These also apply to the following gene marker selection methods.

5.2.3 Highest ratios

In Section 2, ratios of one cell type's expression to average of all other cell types are calculated under the equation below (where r refers to expression intensity, e is expression intensity of a probe for a cell type, i and m represent probes, and j and n represent cell types). Then they are

sorted in a decreasing manner and mapped to gene symbols. Thresholds of 10, 20, 50, 100, 200, 500, are also applied to generate lists of gene markers.

$$r_{i\in P,j\in C} = \frac{e_{i\in P,j\in C}}{mean\ \ of\ \ \{e_{m,n}, m \neq i, n \neq j, m \in P, n \in C\}}$$

$$P = \{all\ \ probes\}, C = \{all\ \ cell\ \ types\}$$

As per fold change approach, it's said that student T test with 95% confidence intervals are implemented [19]. Each probe's highest expressed cell population is compared with the next highest expressed population to find good markers. And this step is repeated by comparing the top cell population with the thirdly highest expressed population. At last, the number of markers to be selected is determined by optimization of condition number.

Considering there are no replicates for each cell type in the library of pure cells, student T test is not performed in this thesis. And merely top cell population and the second highest population are compared with each other by calculating fold changes. The equation of fold change is as follows.

$$Fold\ change = \frac{The\ highest\ expression - The\ second\ highest\ expression}{The\ second\ highest\ expression}$$

Finally, genes are sorted by their fold change values for each cell population. The largest genes for a cell population are determined as good markers for this population. Thresholds of fold change values, instead of the number of markers, are performed in this approach. They are: 0.5, 1, 1.5, 2, 5, and 10.

5.2.4 Combinatorial method

The algorithm of combinatorial method is comprised of three steps: (1) to compute ratios of expressions as explained before; (2) to filter out low expression values as per cell type; (3) to sort ratios and set the threshold.

First, ratios of one cell type are equal to its expression values divided by the average of all other cell types, as described before. In order to make sure gene markers for certain cell types have expression levels in a reasonable range, a lower bound of log2-scaled expression is defined as 8. For each cell population, any probes with expressions under the lower bound will be removed from the list of potential markers for this cell population. But this removed gene might be qualified for other cell types. At last, remaining probes for every cell type are sorted in a decreasing manner and mapped to gene symbols. Thresholds of 10, 20, 50, 100, 200, 500, are also applied to potential genes and then generate lists of markers for all cell types.

5.3 CELL COUNT INFERENCE TOOLS

5.3.1 Application of CIBERSORT to infected lung data

Expression dynamics of murine lung tissues, infected by different virus strains are analyzed in this thesis [14], serving as matrix of Mixture. The virus strains are reported to be: A/Kawasaki/UTK-4/09 H1N1 virus (H1N1), A/California/04/09 H1N1 virus (pH1N1), and A/Vietnam/1203/04 H5N1 virus (H5N1). The library of pure cells serves as Signature Matirx. Quantile normalization of CIBERSORT is disabled for this analysis. Following suggestion given

by Zhong and Liu [34], both expressions of Mixture and Signature Matrix are in linear space, i.e. non-log-scaled.

Estimated cell enrichments of 3 replicates at 14 time points, 0h, 3h, 6h, 9h, 12h, 18h, 24h, 30h, 36h, 48h, 60h, 72h, 120h, 168h, are averaged respectively. Ratios of a particular immune cell among total immune cells are calculated as follows.

$$Ratio\ of\ an\ immune\ cell\ to\ total\ immune\ cells = \frac{Proportion\ of\ an\ immune\ cell}{1 - Proprotion\ of\ lung\ cells}$$

APPENDIX A

GENE MARKER SELECTION METHODS

A.1 CELL SURFACE MARKERS

60 cell surface markers applied in DCQ are: Bcr, Ccr7, Cd14, Cd19, Cd1d1, Cd207, Cd24a, Cd27, Cd28, Cd34, Cd38, Cd3d, Cd4, Cd44, Cd48, Cd5, Cd69, Cd74, Cd86, Cd8a, Cd93, Cr2, Csf1r, Cxcr2, Emr1, Enpep, Entpd1, Epcam, Fcer1g, Fcgr3, Flt3, Foxp3, Icam1, Il2ra, Il2rb, Il7r, Itga2, Itgae, Itgam, Itgax, Kit, Klra3, Klra8, Klrb1c, Ly6a, Ly6c1, Ly86, Ncr1, Nt5e, Pdcd1lg2, Pdgfra, Pdpn, Pecam1, Ptprc,Sdc1, Sell, Siglec1, Siglec5, Slamf1, and Spn.

The 41 markers mapped to genes in the library of pure cells are: Bcr, Ccr7, Cd14, Cd1d1, Cd24a, Cd28, Cd34, Cd38, Cd3d, Cd44, Cd48, Cd69, Cd86, Cd8a, Cd93, Cr2, Csf1r, Emr1, Entpd1, Fcgr3, Flt3, Icam1, Il2ra, Il2rb, Il7r, Itgae, Itgam, Itgax, Klra3, Klra8, Klrb1c, Ly6a, Ly86, Ncr1, Nt5e, Pdcd1lg2, Pecam1, Sell, Siglec1, Slamf1, and Spn.

A.2 INTENSITY-BASED GENE MARKER SELECTION METHOD

The numbers of unique gene markers resulted from sorting expressions under thresholds of 10, 20, 50, 200 and 500 are shown below.

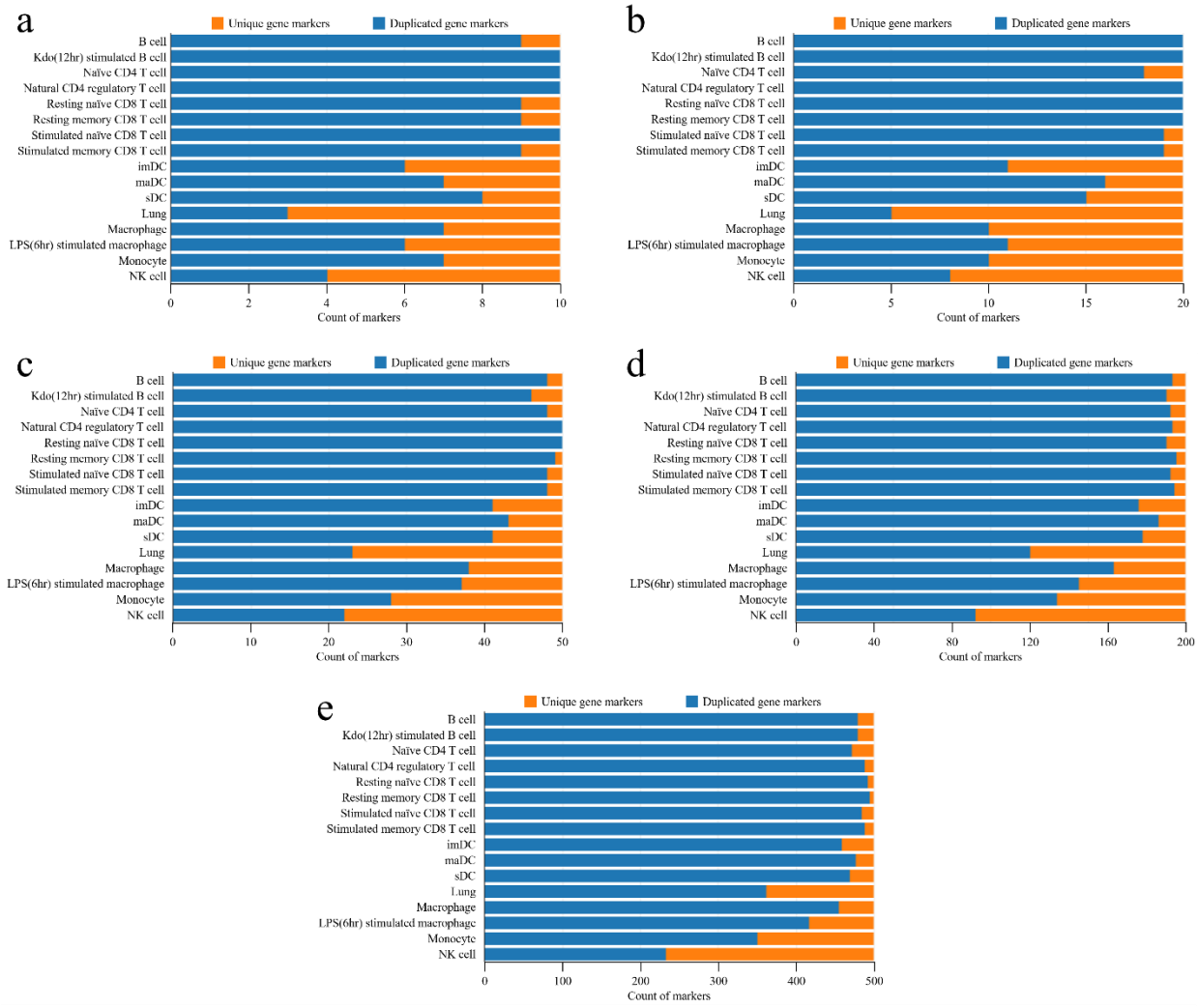

Figure 21. The number of unique gene markers and duplicated markers obtained by sorting expression intensities at different thresholds. Expression intensities of genes are obtained from the library of pure cells. Each cell type's intensities are sorted in a decreasing manner. Selected gene markers for each cell type are compared to markers of other cells to compute unique gene markers, indicated as orange bars. Duplicated markers are indicated as blue bars. (a) Threshold = 10. (b) Threshold = 20. (c) Threshold = 50. (d) Threshold = 200. (e) Threshold = 500.

A.3 HIGHEST RATIO

A.3.1 The number of unique gene markers at thresholds of 10, 20, 50, 200, 500

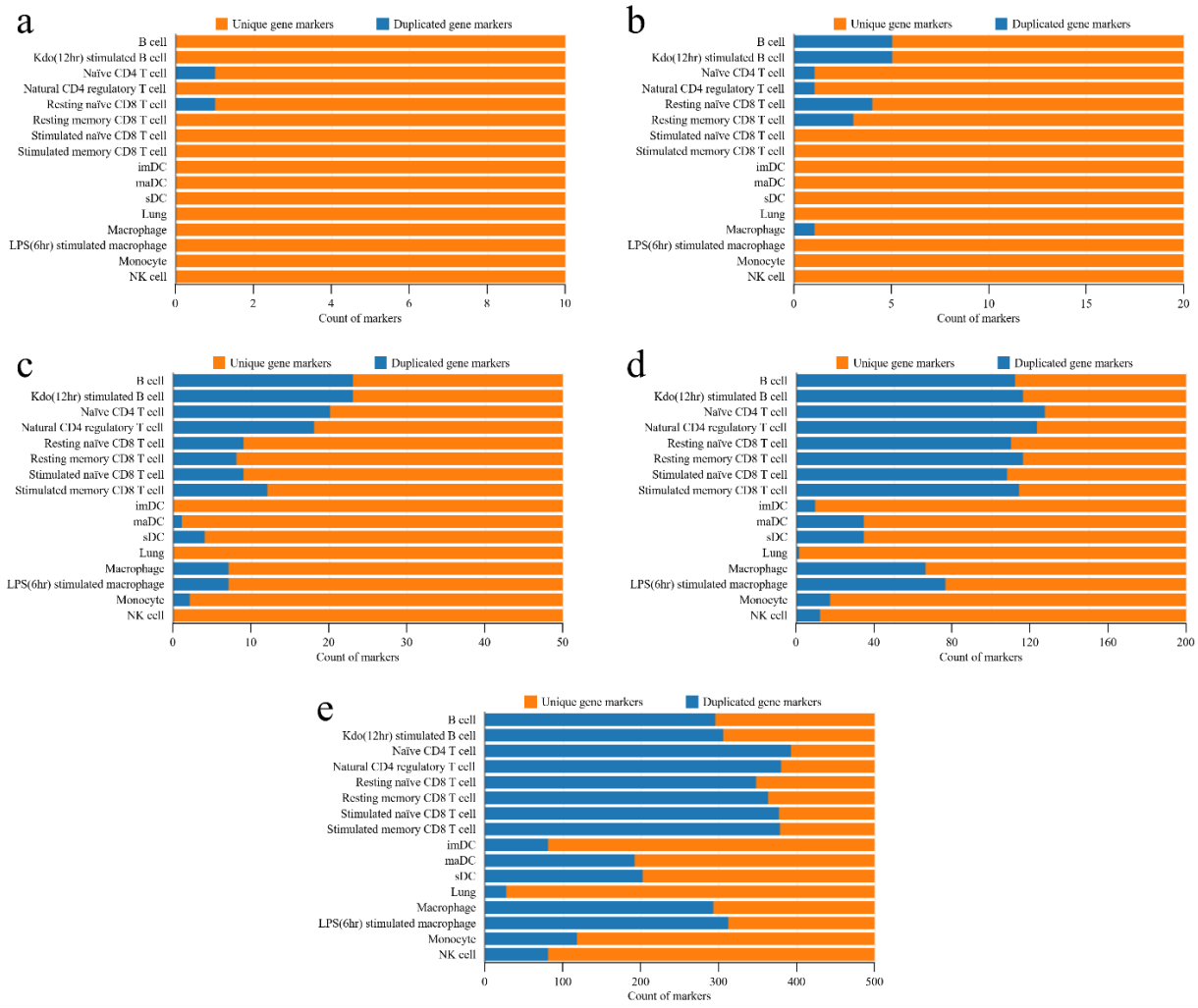

Figure 22. The number of unique gene markers and duplicated markers obtained by calculating expression ratios at different thresholds. For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated, and then sorted in a decreasing manner. Selected gene markers are compared to markers of other cells to compute unique gene markers, indicated as orange bars. Duplicated markers are indicated as blue bars. (a) Threshold = 10. (b) Threshold = 20. (c) Threshold = 50. (d) Threshold = 200. (e) Threshold = 500.

A.3.2 Intensity level of selected markers at thresholds of 10, 20, 50, 200, 500

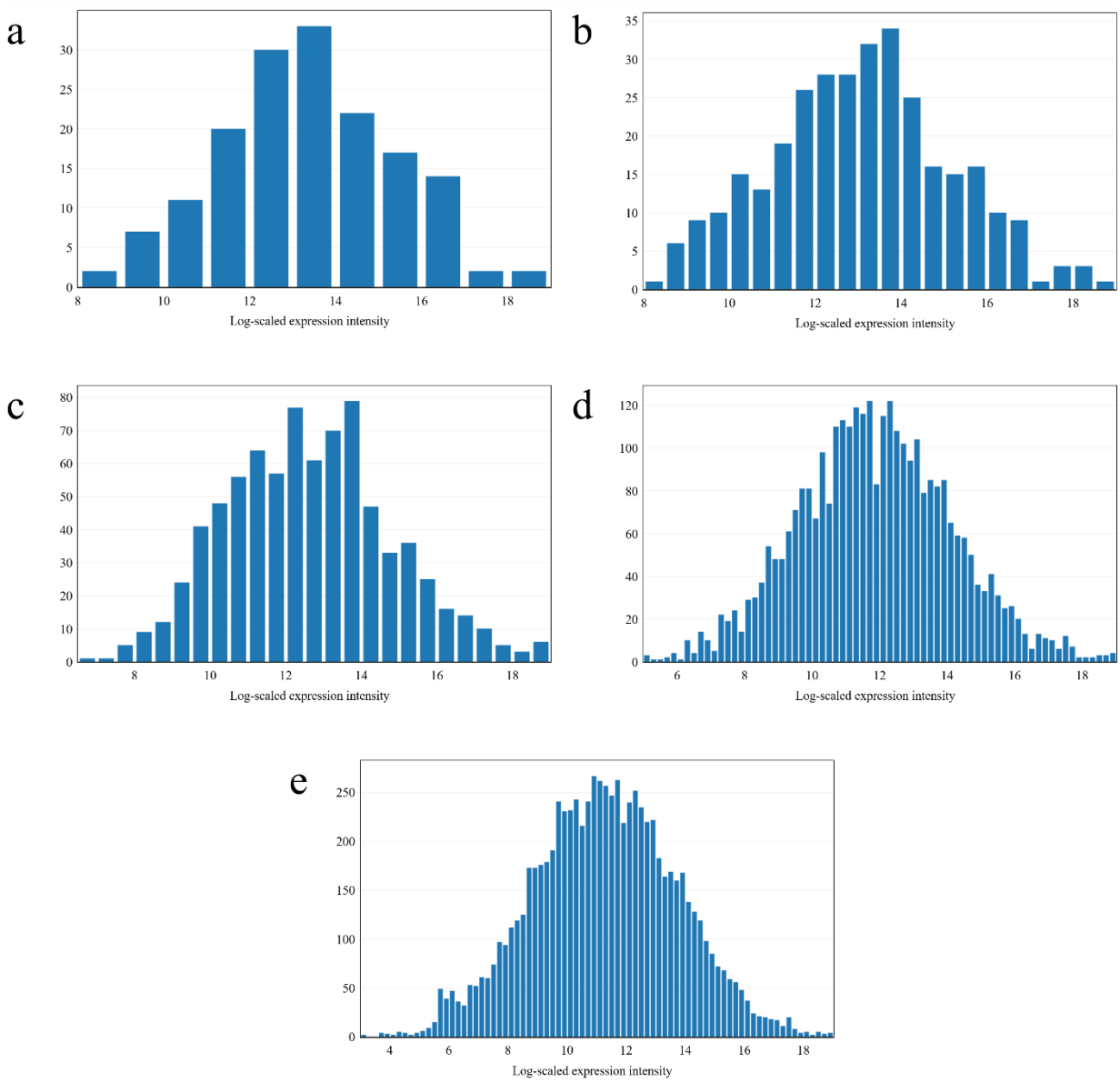

Figure 23. Distribution of expression intensities of all gene markers obtained by calculating expression ratios at different thresholds. For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated, and then sorted in a decreasing manner. The distribution of log-scaled intensities of selected markers are plotted. (a) Threshold = 10. (b) Threshold = 20. (c) Threshold = 50. (d) Threshold = 200. (e) Threshold = 500.

A.4 FOLD CHANGE

A.4.1 The number of unique gene markers at thresholds of 0.5, 1.5, 2, 10

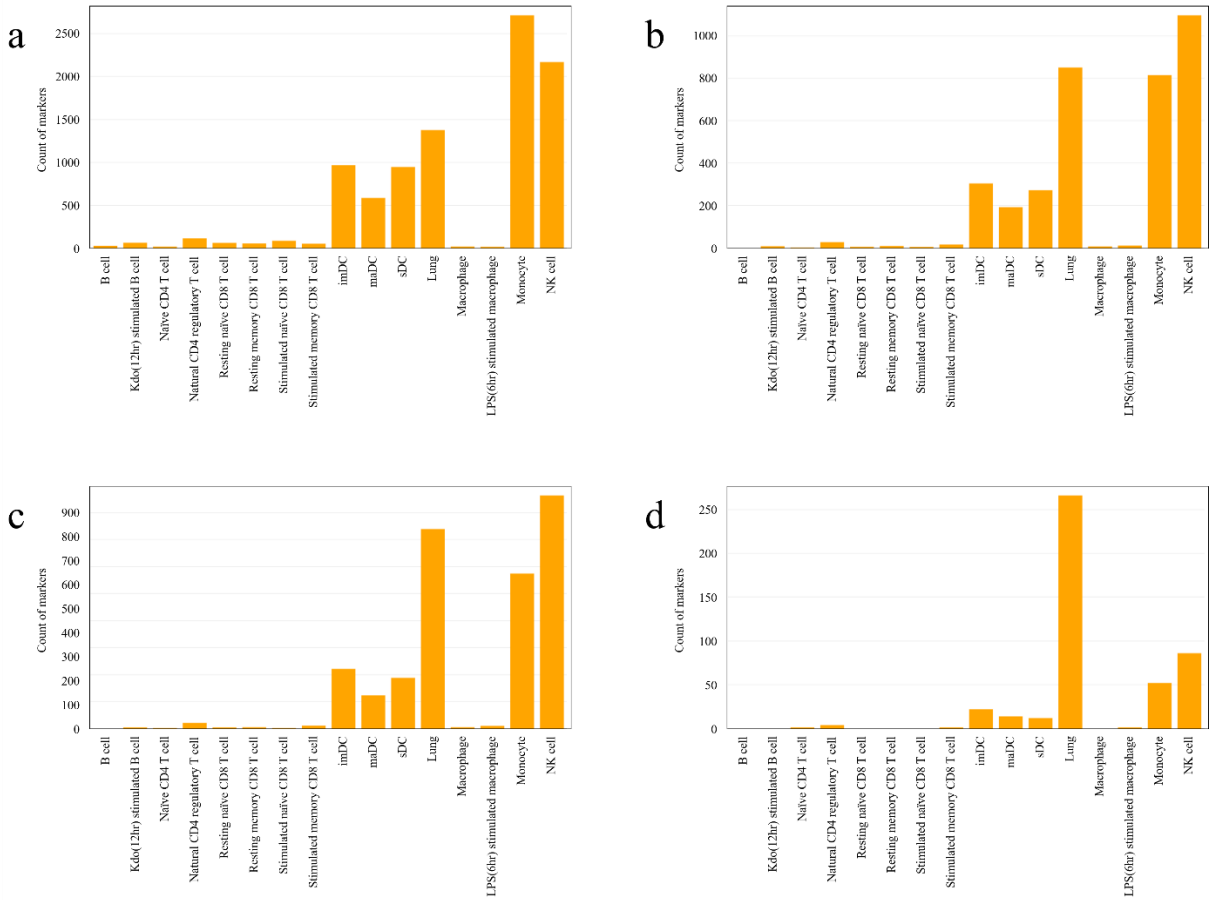

Figure 24. The number of gene markers obtained by sorting fold changes at different thresholds. Each gene's highest expressed cell population is compared with the next highest expressed population, and referring fold change is calculated. Fold change values under 1 are removed, and remained genes are selected as markers. The numbers of obtained markers for each cell type are as the yellow bars. (a) Threshold = 0.5. (b) Threshold = 1.5. (c) Threshold = 2. (d) Threshold = 10.

A.4.2 The intensity level at thresholds of 0, 0.5, 1.5, 2, 5 and 10

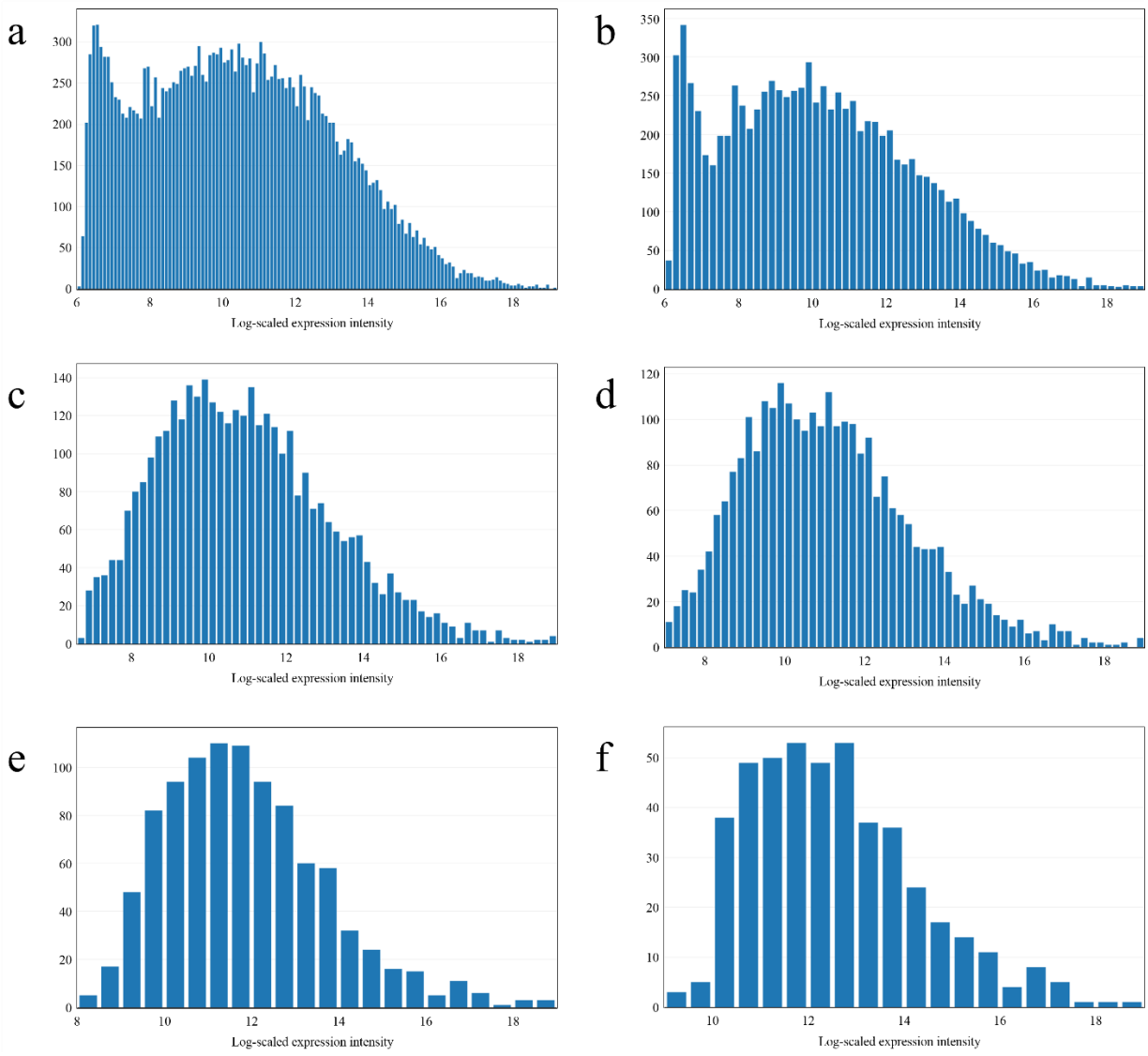

Figure 25. The distribution of expression intensities of gene markers obtained by sorting fold changes at different thresholds. Each gene's highest expressed cell population is compared with the next highest expressed population, and referring fold change is calculated. Fold change values under the threshold are removed, and remained genes are selected as markers. The distribution of their log-scaled intensities are plotted. (a) Threshold = 0. (b) Threshold = 0.5. (c) Threshold = 1.5. (d) Threshold = 2. (e) Threshold = 5. (f) Threshold = 10.

A.5 COMBINATORIAL METHOD

A.5.1 The number of unique gene markers at thresholds of 10, 20, 50, 200 & 500

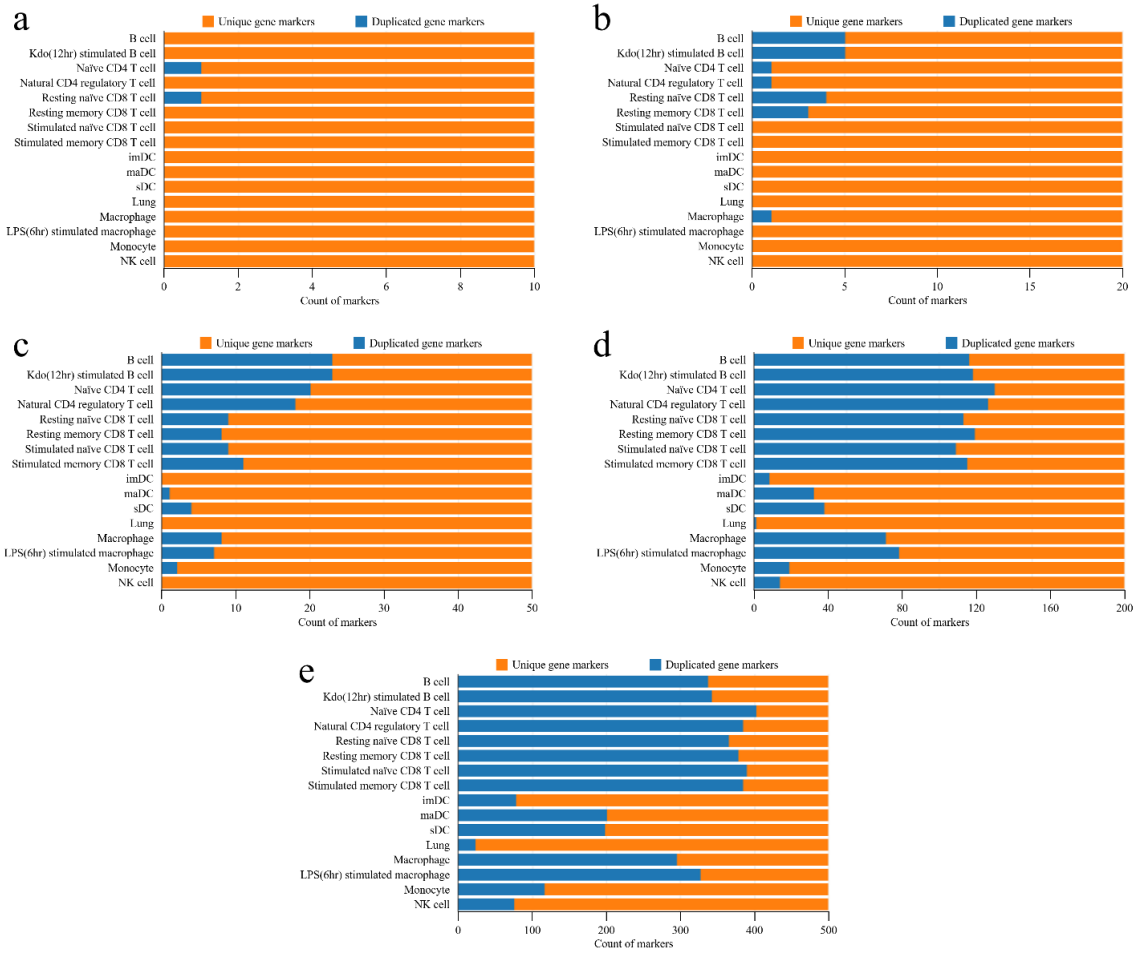

Figure 26. The number of unique gene markers and duplicated markers obtained by combinatorial method at different thresholds. First step of combinatorial method is the same with the approach of highest ratio. For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated. Then genes of log2-scaled expression intensities below 8 are removed. Finally the ratios of remained genes are sorted in a decreasing manner. The selected genes are compared to markers of other cells to compute unique gene markers, indicated as orange bars. Duplicated markers are indicated as blue bars. (a) Threshold = 10. (b) Threshold = 20. (c) Threshold = 50. (d) Threshold = 200. (e) Threshold = 500.

A.5.2 The distribution of intensities at thresholds of 10, 20, 50, 200 & 500

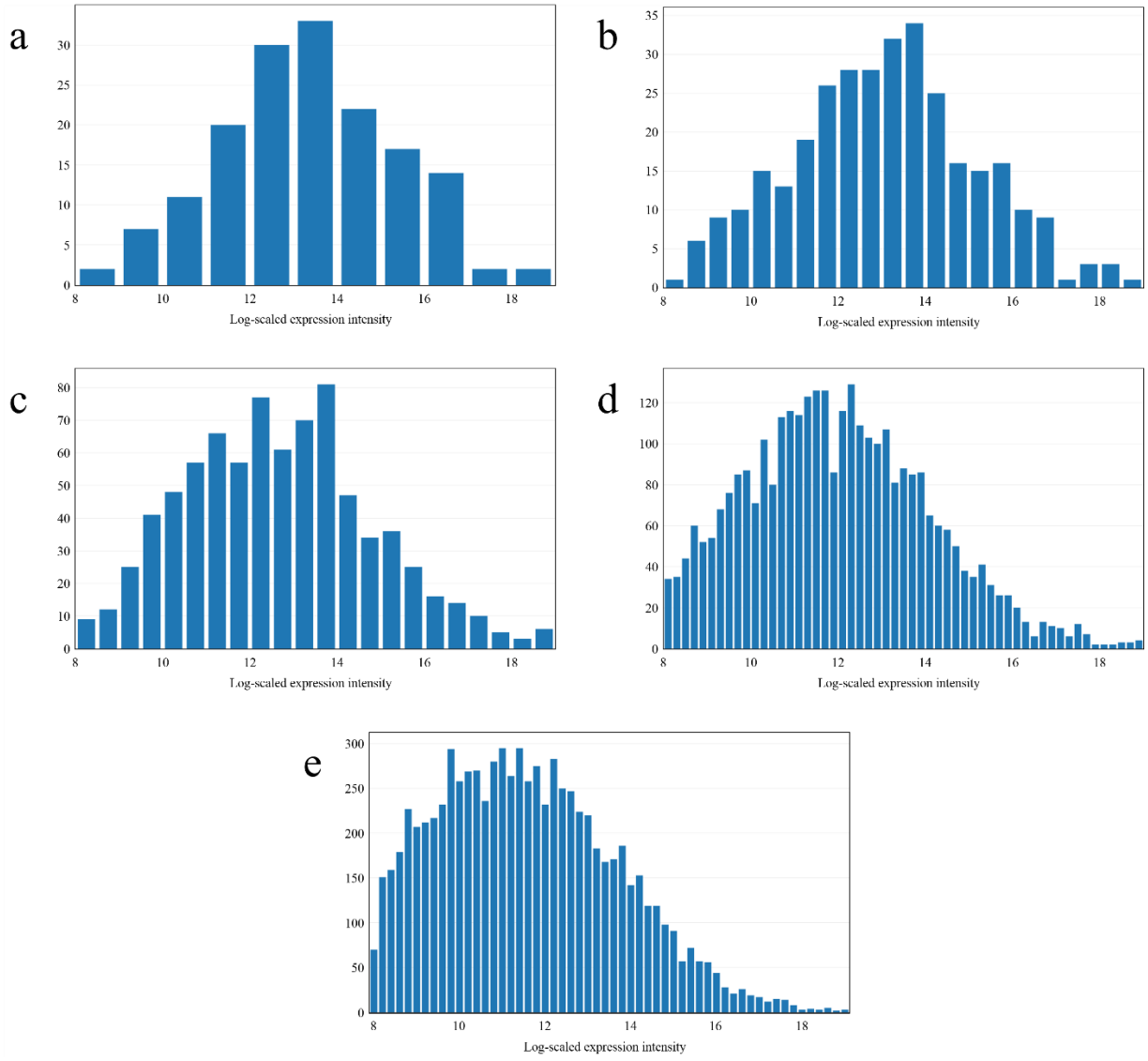

Figure 27. The distribution of expression intensities of gene markers obtained by combinatorial method at different thresholds. First step of combinatorial method is the same with the approach of highest ratio. For each gene and each cell type included in our library of pure cells, the ratio of a certain cell type's expression intensity to the average of all other cell types are calculated. Then genes of log2-scaled expression intensities below 8 are removed. Finally the ratios of remained genes are sorted in a decreasing manner. The distribution of selected markers' log-scaled intensities are plotted. (a) Threshold = 10. (b) Threshold = 20. (c) Threshold = 50. (d) Threshold = 200. (e) Threshold = 500.

APPENDIX B

CELL COUNT INFERENCE TOOLS

B.1 CTEN

B.1.1 100 markers of B cells uploaded to CTen

Kcnj1, Fcer2a, Spib, H2-Ab1, H2-Aa, Cecr2, Pla2g2d, LOC675694, Pou2af1, Pgls, Dtx1, Dnase1l3, Ms4a1, H2-Ob, Hip1r, Sfn, Chst3, Tnfrsf13c, Igh-6, 2010309G21Rik, Zfp318, Igk-V1, H2-Oa, Vpreb3, H2-Eb1, Ccr6, Bfsp2, Faim3, Cerk, Ell3, Igl-V1, Pck2, Slc23a1, Cxxc5, Igh-VJ558, A530040E14Rik, Myo1c, Neil1, Fcrl1, Sbk1, Ighg, Blnk, Igk-V19-14, Mif4gd, LOC629915, Mapk11, Siglecg, Bmf, 1700021K19Rik, Igk-V21-12, LOC380824, Trib3, Tmem163, Eaf2, Bcl11a, Cd22, Ier5l, LOC434638, Lynx1, AI324046, 2010007H06Rik, Bank1, 4930566A11Rik, C2ta, Cabc1, B3gnt5, Igkv1-133, Arid5a, Crip3, Vps13a, LOC544905, Bach2, Mical1, Ralgps2, Aars, Stk23, 1500001A10Rik, D6Ertd456e, Tubb2b, Pou2f2, LOC619916, LOC669091, Icosl, Eif2ak3, LOC619994, Amn, Slc4a1, Cnr2, Rasal1, 2310015N21Rik, Itpr3, Bcl3, 0610039H22Rik, Igk-V8-16, Vars2, Ero1lb, Igkv4-63, Igk-V38, Kcnb1, and LOC668544.

B.1.2 Enrichment scores resulted from CTen for gene markers of B cells obtained by combinatorial method

Table 3. Cell types with enrichment scores more than two for gene markers of B cells obtained by combinatorial method

Tissue/Cell type	Enrichment score
Follicular B Cells	36.46890902
B Cells (GL7 pos; KLH)	32.85258623
B Cells (GL7 pos; Alum)	31.89924
B Cells (GL7 neg; KLH)	28.00279686
Spleen	27.62642604
B Cells (GL7 neg; Alum)	26.41449094
B Cells Marginal Zone	22.96848378
Lymph Nodes	22.3594034
Foxp3+ Tcells	10.46394055
Bone Marrow	9.132079099
CD8a+ Dend. Cells Myeloid	9.008724547
Bone	8.563331846
B220+ Dend. Cells	7.018300587
CD8a+ Dend. Cells Lymphoid	4.615836135
CD8+ T cells	2.358491917
Mast Cells IgE	2.060619809

B.1.3 Enrichment scores resulted from CTen for the submodule of N2 [14]

Table 4. Enrichment scores of top 10 enriched cell types for the submodule of N2 **[14]**

Cell Type	Enrichment Score
Macrophage Bone Marrow Lps 24 Hrs	103.9353717
Macrophage Bone Marrow Lps 6Hrs	94.19054067
Macrophage Bone Marrow Lps 2Hrs	84.48263487
Macrophage Peri Lps 7Hrs	81.26686847
Osteoclasts	79.83518315
Microglia	78.39783479
Macrophage Peri Lps 1Hrs	75.22997402
Macrophage Bone Marrow	75.15725021
Macrophage Peri	72.82573035
CD8a+ Dend. Cells Myeloid	63.17519274

B.2 CIBERSORT

B.2.1 Cell fractions of influenza infected lung tissues predicted by CIBERSORT

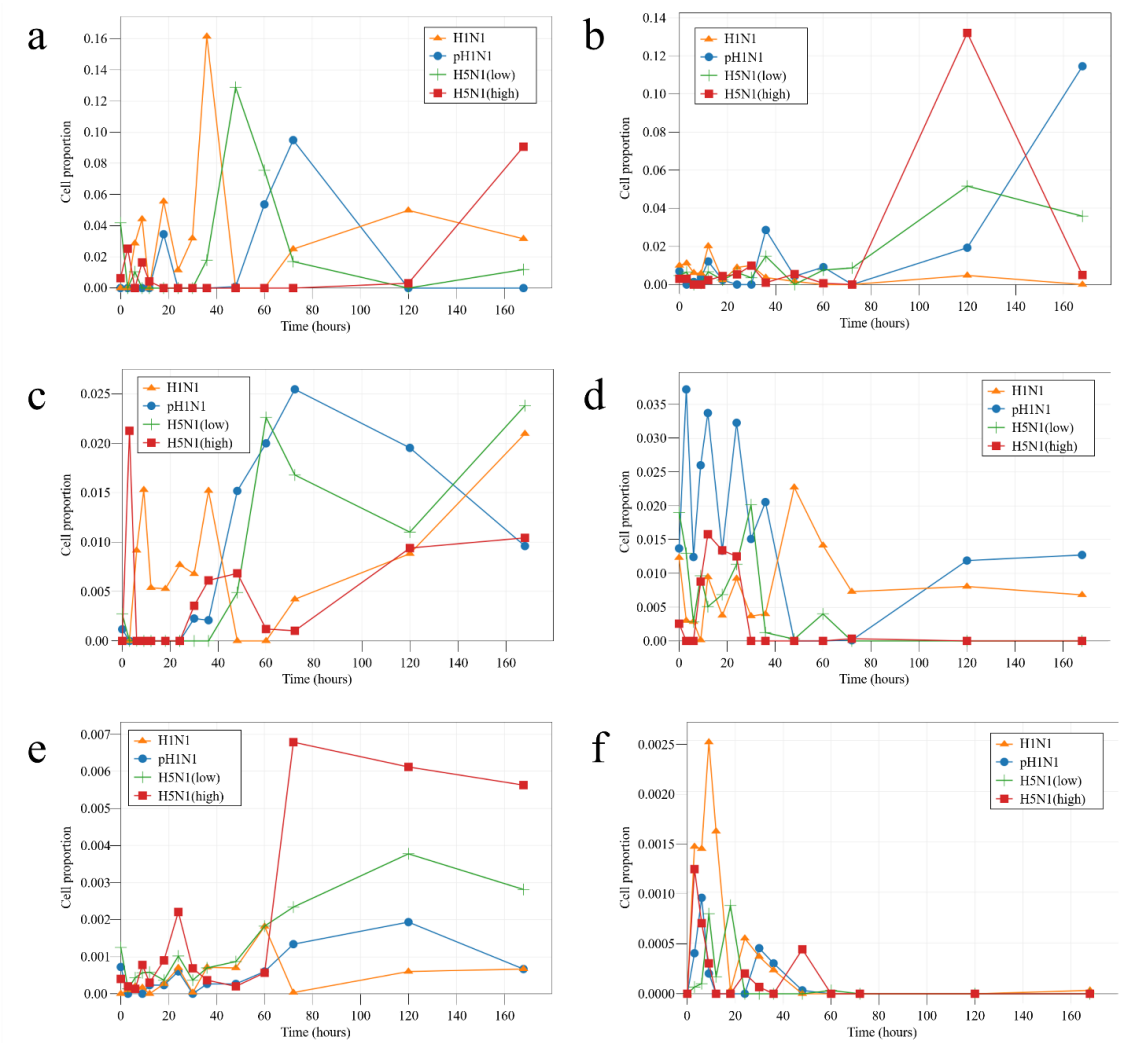

Figure 28. Proportions of (a) macrophage; (b) stimulated memory CD8 T cells; (c) B cells; (d) stimulated B cells; (e) imDCs; (f) maDCs at 14 time points for different virus strains. Microarray data of murine lung tissues, infected by H1N1, pH1N1, H5N1 of low and high dose at different time points are obtained from literature **[14]**. The library of pure cells serves as Signature Matirx. Both datasets are applied to CIBERSORT to estimate cell proportions. Estimated proportions of 3 replicates at 14 time points, 0h, 3h, 6h, 9h, 12h, 18h, 24h, 30h, 36h, 48h, 60h, 72h, 120h, 168h, are averaged respectively.

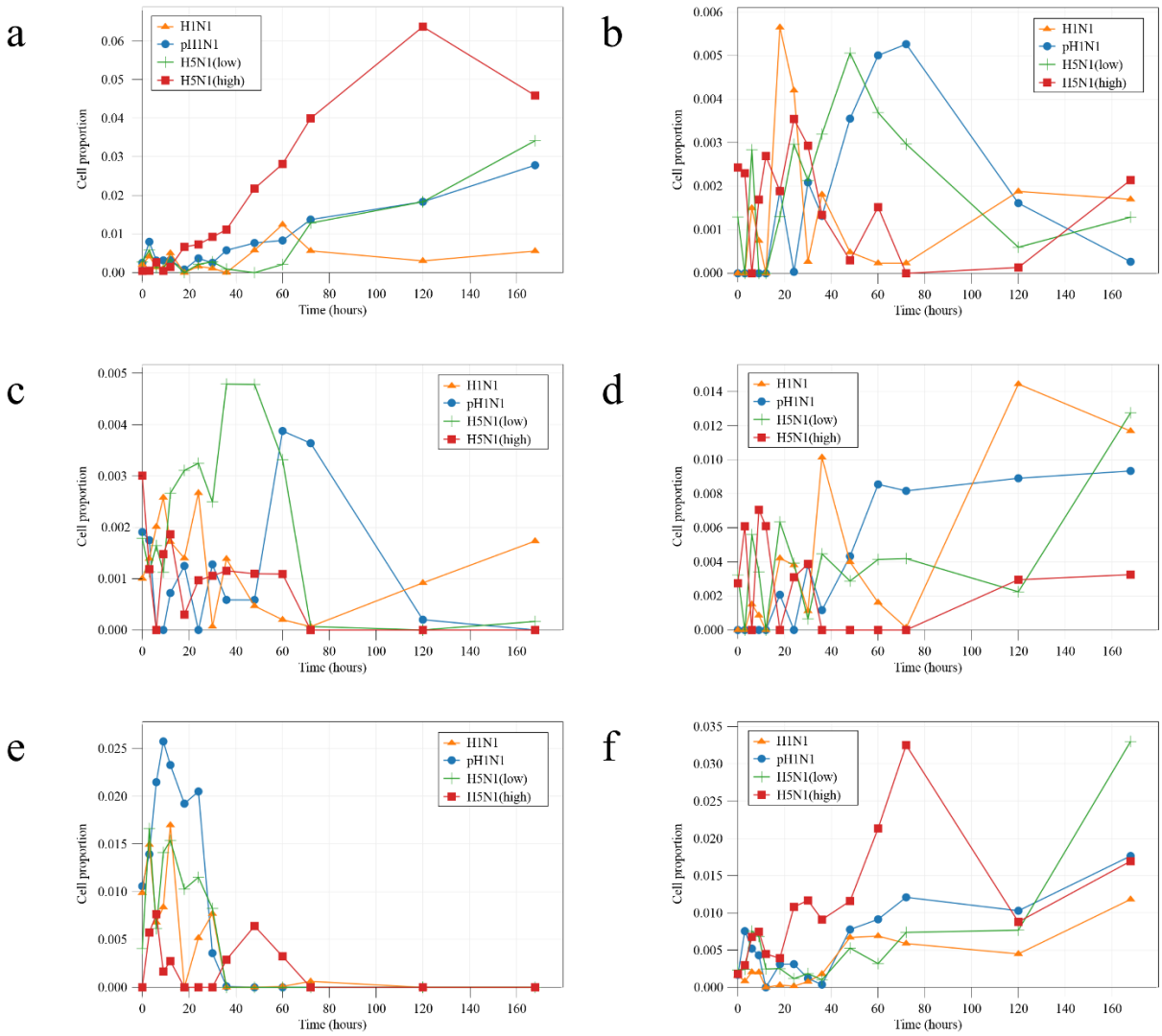

Figure 29. Proportions of (a) sDCs; (b) monocyte; (c) naïve CD4 T cells; (d) natural CD4 Tregs; (e) resting naïve CD8 T cells; (f) stimulated naïve CD8 T cells at 14 time points for different virus strains. Microarray data of murine lung tissues, infected by H1N1, pH1N1, H5N1 of low and high dose at different time points are obtained from literature **[14]**. The library of pure cells serves as Signature Matirx. Both datasets are applied to CIBERSORT to estimate cell proportions.

BIBLIOGRAPHY

[1] World Health Organization, "WHO," 11 2016. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs211/en/. [Accessed 3 2017].

[2] Centers for Disease Control and Prevention, "CDC," 13 12 2016. [Online]. Available: https://www.cdc.gov/flu/about/disease/2015-16.htm. [Accessed 3 2017].

[3] Centers for Disease Control and Prevention, "CDC," 4 5 2016. [Online]. Available: https://www.cdc.gov/flu/about/qa/disease.htm. [Accessed 3 2017].

[4] Centers for Disease Control and Prevention, "CDC," 15 8 2016. [Online]. Available: https://www.cdc.gov/flu/about/qa/testing.htm. [Accessed 3 2017].

[5] R. Dolin, "Chapter 187. Influenza," in Harrison's Principles of Internal Medicine, New York, McGraw-Hill, 2012.

[6] Centers for Disease Control and Prevention, "CDC," 5 1 2017. [Online]. Available: https://www.cdc.gov/flu/antivirals/whatyoushould.htm. [Accessed 3 2017].

[7] Centers for Disease Control and Prevention, "CDC," 25 1 2017. [Online]. Available: https://www.cdc.gov/flu/professionals/antivirals/links.htm. [Accessed 3 2017].

[8] G. R. Klimpel, "Immune Defenses," in Medical Microbiology, Galveston, University of Texas Medical Branch at Galveston, 1996.

[9] T. Wang, T. Town, L. Alexopoulou, J. F. Anderson, E. Fikrig and R. A. Flavell, "Toll-like receptor 3 mediates West Nile virus entry into the brain causing lethal encephalitis," Nature Medicine, vol. 10, pp. 1366 - 1373, 2004.

[10] Z. Altboum, Y. Steuerman, E. David, Z. Barnett‐Itzhaki and L. Valadarsky, "Digital cell quantification identifies global immune cell dynamics during influenza infection," Molecular Systems Biology, vol. 10, no. 2, 2014.

[11] H. I. Nakaya, J. Wrammert, E. K. Lee, L. Racioppi and S. Marie-Kunze, "Systems Biology of Seasonal Influenza Vaccination in Humans," Nat Immunol, vol. 12, no. 8, pp. 786-795, 2011.

[12] L. Josset, J. A. Belser, M. J. Pantin-Jackwood and J. H. Chang, "Implication of Inflammatory Macrophages, Nuclear Receptors, and Interferon Regulatory Factors in Increased Virulence of Pandemic 2009 H1N1 Influenza A Virus after Host Adaptation," Journal of Virology, vol. 86, no. 13, pp. 7192-7206, 2012.

[13] J. E. Shoemaker, T. J. Lopes, S. Ghosh, Y. Matsuoka, Y. Kawaoka and H. Kitano, "CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data," BMC Genomics, 2012.

[14] J. E. Shoemaker, S. Fukuyama, A. J. Eisfeld, D. Zhao, E. Kawakami and S. Sakabe, "An Ultrasensitive Mechanism Regulates Influenza Virus-Induced Inflammation," PLoS Pathogens, 2015.

[15] P. Lu, A. Nakorchevskiy and E. M. Aleksey, "Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations," PNAS, vol. 100, no. 18, p. 10370–10375, 2003.

[16] M. Wang, S. Master and L. A. Chodosh, "Computational expression deconvolution in a complex mammalian organ," BMC Bioinformatics, vol. 7, no. 328, 2006.

[17] Y. Zhong, Y.-W. Wan, K. Pang, L. M. Chow and Z. Liu, "Digital sorting of complex tissues for cell," BMC Bioinformatics, vol. 14, no. 89, 2013.

[18] D. A. Liebner, K. Huang and J. D. Parvin, "MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples," Bioinformatics, vol. 30, no. 5, pp. 682-689, 2014.

[19] A. Abbas, K. Wolslegel, D. Seshasayee, Z. Modrusan and H. Clark, "Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus," Plos One, 2009.

[20] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu and C. D. Hoang, "Robust enumeration of cell subsets from tissue expression profiles," Nature Methods, p. 453–457, 2015.

[21] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian and F. Staedtler, "Cell type–specific gene expression differences in complex tissues," Nature Methods, vol. 7, no. 4, 2010.

[22] "QIAGEN," QIAGEN, [Online]. Available: https://www.qiagen.com/us/shop/genes-and-pathways/complete-biology-list/cell-surface-markers/. [Accessed 2017].

[23] B. Biosciences, "BD Biosciences," 2010. [Online]. Available: https://www.bdbiosciences.com/documents/cd_marker_handbook.pdf. [Accessed 2017].

[24] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad, "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," Genome Research, p. 1509–1517, 2008.

[25] R. Gazit, R. Gruda, M. Elboim, T. I. Arnon, G. Katz and H. Achdout, "Lethal influenza infection in the absence of the natural killer cell receptor gene Ncr1," Nature Immunology, vol. 7, pp. 517-523, 2006.

[26] L. Escoubet-Lozach, C. Benner, M. Kaikkonen and J. Lozach, "Mechanisms establishing TLR4-responsive activation states of inflammatory response genes," PLoS Genetics, 2011.

[27] F. Pan, H. Yu, E. Dang and J. Barbi, "Eos mediates Foxp3-dependent gene silencing in CD4+ regulatory T cells," Science, vol. 325, no. 5944, pp. 1142-1146, 2009.

[28] D. Joanna and S. Hao, "NCBI," 18 May 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16145. [Accessed 6 January 2017].

[29] X. Liu, X. Qu, Y. Chen and L. Liao, "Mesenchymal stem/stromal cells induce the generation of novel IL-10-dependent regulatory dendritic cells by SOCS3 activation," The Journal of Immunology, vol. 189, no. 3, pp. 1182-1192, 2012.

[30] K. Al Moussawi, E. Ghigo, U. Kalinke and L. Alexopoulou, "Type I interferon induction is detrimental during infection with the Whipple's disease bacterium, Tropheryma whipplei," PLoS Pathogens, 2010.

[31] E. Ghigo, A. Barry, L. Pretat and K. Al Moussawi, "IL-16 promotes T. whipplei replication by inhibiting phagosome conversion and modulating macrophage activation," PLoS One, 2010.

[32] F. Swirski, M. Nahrendorf, M. Etzrodt and M. Wildgruber, "Identification of splenic reservoir monocytes and their deployment to inflammatory sites," Science, vol. 325, no. 5940, pp. 612-616, 2009.

[33] A. Latorre, B. Caniceiro, H. Fukumasu and D. Gardner, "Ptaquiloside reduces NK cell activities by enhancing metallothionein expression, which is prevented by selenium," Toxicology, vol. 304, pp. 100-108, 2013.

[34] Y. Zhong and Z. Liu, "Gene expression deconvolution in linear space," Nature methods, vol. 9, no. 1, pp. 8-9, 2012.

[35] S. F. Ibrahim and G. van den Engh, "Flow cytometry and cell sorting," Adv Biochem Eng Biotechnol, vol. 106, pp. 19-39, 2007.

[36] S. C. Bendall, E. F. Simonds, P. Qiu, A. D. Amir and P. O. Krutzik, "Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum," Science, vol. 332, pp. 687-696, 2011.

[37] T. F. Kong, W. Ye, W. K. Peng, H. W. Hou and Marcos, "Enhancing malaria diagnosis through microfluidic cell enrichment and magnetic resonance relaxometry detection," Scientific Reports, 2015.

[38] C. Mueller, A. deCarvalho, T. Mikkelsen, N. Lehman and V. Calvert, "Glioblastoma cell enrichment is critical for analysis of phosphorylated drug targets and proteomic-genomic correlations," Cancer Res, vol. 74, no. 3, 2014.

[39] R. R. Jahan-Tigh, C. Ryan, G. Obermoser and K. Schwarzenberger, "Flow Cytometry," Journal of Investigative Dermatology, vol. 132, 2012.

[40] T. Mauad, L. A. Hajjar, G. D. Callegari, L. F. da Silva and D. Schout, "Lung Pathology in Fatal Novel Human Influenza A (H1N1) Infection," American Journal Of Respiratory And Critical Care Medicine, vol. 181, pp. 72-79, 2010.

[41] J. S. Peiris, C. Y. Cheung, C. Y. H. Leung and J. M. Nicholls, "Innate immune responses to influenza A H5N1: friend or foe?," Trends in Immunology, vol. 30, no. 12, pp. 574-584, 2009.