RNA-seq Deconvolution Project Report

Yujia Shi

June 19, 2020

Contents

1 Introduction

With the rapid development of single cell RNA-seq technologies, people can easily construct the precise cell-type specific expression profile, which supports the research on cellular heterogeneity. However, single cell RNA-seq technologies has not yet been widely used due to its high cost and complicated experiment procedure. Given the advantages of cost and technical simplicity, bulk RNA-seq is still the mainstream. However, the existence of the cell-type specific single-cell expression profile allows us to extract the latent information within bulk RNA-seq data, such as cell type composition. In this project, we explore the application of weighted least square and iteratively reweighted least square in the bulk RNA-seq deconvolution. We also benchmark our implementation with three mainstream deconvolution methods including deconRNASeq [1], Dtangle[2] and cibersort [3].

2 Method

2.1 Implementation

The IRLS method is implemented using Python 3.6. Source codes are available at https://github.com/yujias424/RNA-seq_deconvolution_project.

2.2 Design of signature matrix and simulated bulk data

To benchmark our approach with other methods using real dataset with non-simulated noise, we included multiple RNA-seq studies selected from metaSRA [4]. Total 263 studies and 4167 RNA-seq experiments were included in this analysis. The count table are normalized to count-per-million (CPM). We also pre-identified the cell type label of each experiment, and there are total 104 cell types.

We propose that the systematic noise can be divided into two parts, including with-in study noise and cross-studies noise. Due to the limitation of our data set, some cell types only have one corresponding study and thus the cross-studies variance is absent. To avoid the potential bias, we exclude the cell types with less than two corresponding studies. There are 48 cell types left for further analysis.

In each simulated experiment, a noise-free signature matrix used for deconvolution and a noisy signature matrix used for generating the bulk data will be built. To generate the noisy signature matrix, we randomly select one experiment from a randomly picked study for each cell type. The experiments from selected study will be excluded during the creation of noise-free signature matrix. For each cell type, the mean profile will be built using the rest of the experiments, which is thus regarded as the signature profile of the corresponding cell type. The final noise-free cell-type signature matrix will be constructed by joining aforementioned cell type specific mean profile.

Cell type composition will be simulated using a Dirichlet distribution, and the bulk data is generated by multiplying the noisy signature matrix and the simulated proportion.

2.3 Marker genes selection

The marker genes will be selected using the similar approach described in Dtangle [2]. Dtangle will look at the ratio of the mean expression for each cell type to the sum of the mean expressions by all other cell types. In our benchmark experiments, multiple thresholds are selected to determine the number of marker genes that are used for deconvolution, ranged from 1% to 90%. The selection of quantile threshold will significantly effect the computational time.

2.4 Weighted least square

Our approach is developed based on the ordinary least squares approach [1]. We first assume a noise-free RNA-seq data. The expression level of gene i in such data then can be modeled as following equation,

$$y_i = \sum_{j=1}^{N} p_j x_{ij} \tag{1}$$

where p_{ij} denotes the proportion of cell type j and x_{ij} denotes the signature expression level of gene i in cell type j. In general, the equation can also be represented in matrix form as

$$Y = \vec{P}X \tag{2}$$

where Y is the bulk RNA-seq data to be deconvoluted, $\vec{P}$ is the relative proportions of cell types and X is the signature matrix.

However, in real situation, the data we want to deconvolute often comes with heavy noise. The expression model in matrix form can then be modeled as

$$Y = \vec{P}X + \epsilon \tag{3}$$

where the ϵ is a noise term matrix contains the noise for all genes. We assume that the noise term is a Gaussian $N(0, \sigma_i^2)$, and thus the expression level of gene i should come from another Gaussian, $N(\vec{P}X, \sigma_i^2)$. The ordinary least squares approach assumes that the noise of each genes is IID Gaussian white noise. In other words, the variance of the Gaussian distribution that different gene's noise comes from is a constant. However, in real situation, the magnitude of different gene's noise is not a constant. For example, only the expression level of housekeeping genes included in some gene ontology groups may hard to be effected by the disturbance of the experiment condition [5]. Therefore, the gene expression data in real situation can often be regarded as heteroskedastic data.

To solve the heteroskedastic data, we can apply the weighted least square approach, which aims at minimizing the weighted sum of squares (WSS) rather than the residual sum of squares (RSS).

$$WSS(X, \vec{P}) = \sum_{i=1}^{N} w_i(y_i - \vec{P} \cdot x_i) \tag{4}$$

We still proposed that the each gene's noise term is a Gaussian, while their variance is not a constant. Assume that the noise variance σ_i^2 of gene i is known, we can give the gene i a weight $1/\sigma_i^2$. The ordinary maximum likelihood estimate (MLE) then can be turned to the heteroskedastic MLE, in which we can achieve a relative precise estimation on the proportion vector $\vec{P}$.

2.5 Noise model

As discussed in weighted least square part, to achieve a proportion estimation with high accuracy, we need to have a relative accurate estimation of the gene specific noise variance. In real situation, we can only observe the bulk data and the signature matrix. Therefore, it is natural for us to propose that there may exist a function to describe the relationship between noise term variance σ_i^2 and the observed expression level y_i.

To train a mean expression versus variance model, the variance for each gene will be firstly calculated by summing up the variance of gene in different cell types. We next built a mean profile based matrix consisting of 253 studies for 48 cell types as discussed in section 2.2, where each column represents the mean expression profile of a specific study containing the experiments labeled with the cell type. The model can be trained using the gene's variance and corresponding mean expression level.

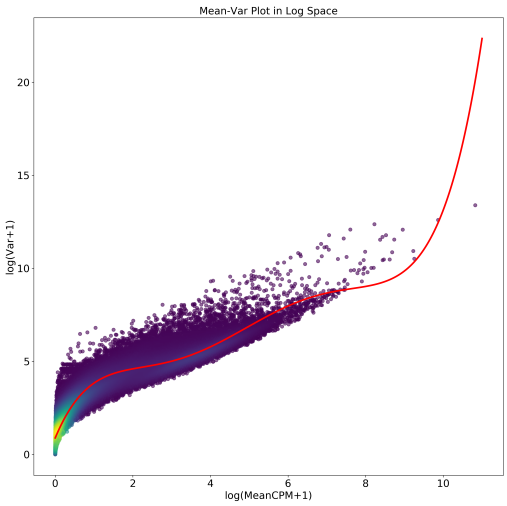


Figure 1: Mean expression level versus Variance model in log space.

Our model shows that the noise variance can be modeled by a function of expression level, which can be directly used as a tool to estimate the gene specific noise variance. Furthermore, we can notice that the regression curve can be approximated as a linear function, indicating that it is possible for us to directly use the expression level as the noise variance in practice.

2.6 Creation of noise variance matrix

Apart from estimating a function between noise term variance σ_i^2 and the observed expression level y_i, we may also propose that the noise variance is independent with the observed expression level. Given a dataset with multiple studies and experiments used for constructing the signature matrix, we can also learn a noise variance matrix withe the same shape as the signature matrix to describe the embedded noise of different cell types.

We assume that different components of noise are independent Gaussian. Therefore, given the variance sum law, the variance of the cell-type specific noise can be calculated through summing up all corresponding within studies variance and the cross studies variance. We can thus obtain the aforementioned variance matrix with same shape as the signature matrix, which can be used for iteratively reweighted least square approach.

To be noticed, the within-study noise variance of gene j is the variance of residual, which is the difference between the mean expression level of gene j and the observed expression level of gene j in the experiment from a certain study. The cross-studies noise variance of gene j is calculated through taking the variance across gene j's mean expression level among selected studies.

2.7 Iteratively reweighted least square

As the discussion about the variance model in the previous section, there are three types of the iteratively reweighted least square (IRLS). The difference between these three types of IRLS approach is that how these methods update the weights in each iteration.

The most simplest approach is to use the observed expression level as an estimation of the noise variance. Aforementioned variance model shows that the variance σ_i^2 can be approximately modeled as a linear function of the observed expression level. Therefore, the variance σ_i^2 can be directly replaced by the observed expression level. The weights are thus an identity function of expression level. Inspired by the iteratively reweighted least squares, we can update the weights by multiplying the latest solution $\vec{P}^{(k)}$ and the signature matrix X.

Instead of using the identity function of the expression level, we can also use the learned univariatespline function to get a more accurate estimation of the noise variance (Figure 1).

Last but not least, we can update the weights by multiplying the latest $\vec{P}$ and the learned variance matrix as well. To be noticed, according to the variance sum law, we need to take a square of $\vec{P}$ before the multiply operation.

Convergence is reached when $||\vec{P}^{(k+1)} - \vec{P}^{(k)}|| \leq 0.01$.

3 Results

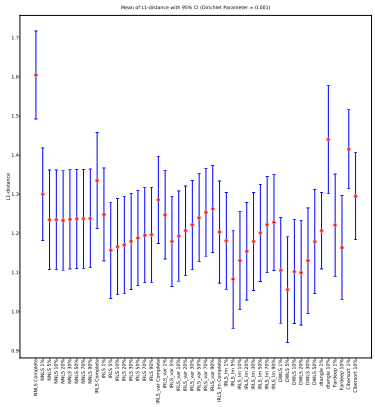
3.1 Learned model function based IRLS performs better than other mainstream methods

We benchmarked our implementation with other mainstream methods. The L1-distance between estimated proportion and true proportion is used as the criterion to compare the performance of different methods. We simulate the cell type composition using Dirichlet distribution with four different parameters 0.001, 0.01, 0.1 and 1. 100 replications were performed for each dirichlet parameter setting, and we compare the mean of the L1-distance results of these experiment to identify the method with best performance.

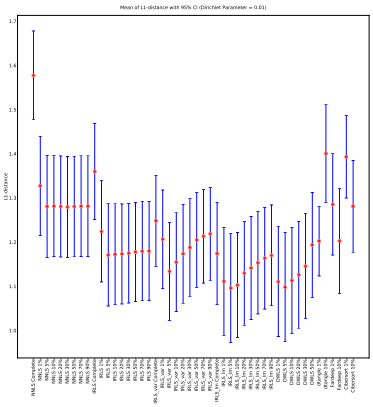
Figure 2 shows that the learned model based IRLS (IRLS_lm) outperforms other approaches including deconRNASeq(NNLS), dtangle, cibersort and identity function based IRLS. As expected, learned model will give a more accurate estimation of the noise variance and thus achieve a better performance compared with the identity function based IRLS.

3.2 Correct estimation of noise variance leads to a better result

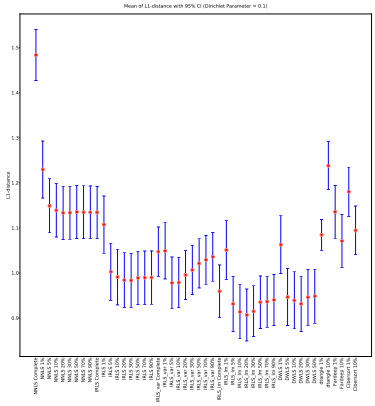
We may also want to see the performance of variance matrix based IRLS. Figure 3 shows that variance matrix based IRLS outperforms other IRLS approaches. Since the studies used for creating the bulk data are also used for generating the variance matrix, the variance matrix can capture the noise within the bulk



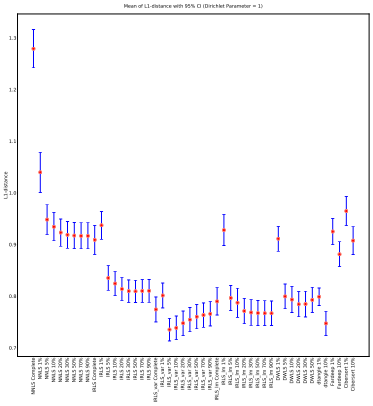
(a) Dirichlet parameter (α) = 0.001



(b) Dirichlet parameter (α) = 0.01



(c) Dirichlet parameter (α) = 0.1



(d) Dirichlet parameter (α) = 1

Figure 2: Benchmark of IRLS approaches with other mainstream methods. Var method can be ignored here since the weights used in this experiment is the sum of variance directly without considering the estimated proportion.

data. This operation makes the variance matrix based IRLS obtain the best performance. However, in real situation, the signature matrix is generated using the dataset which doesn't have any relationship with the bulk data to be deconvoluted. Therefore, aforementioned results can only indicate that the correct estimation of noise variance is a key step of the deconvolution algorithm and will have strong impact on the performance of IRLS approach.

4 Discussion

This project shows that the IRLS is a promising research direction of RNA-seq deconvolution. Our results indicates that a good estimation of gene-specific noise variance is an important compartment of the IRLS algorithm. Although the identity function based IRLS has a better computational time, its performance cannot beat other two IRLS approaches with a more accurate estimation on the noise variance, meaning that expression level is an inaccurate estimation of the noise variance.

In the future, this project may focus on how to obtain an accurate estimation of the gene-specific noise variance from a given dataset.

References

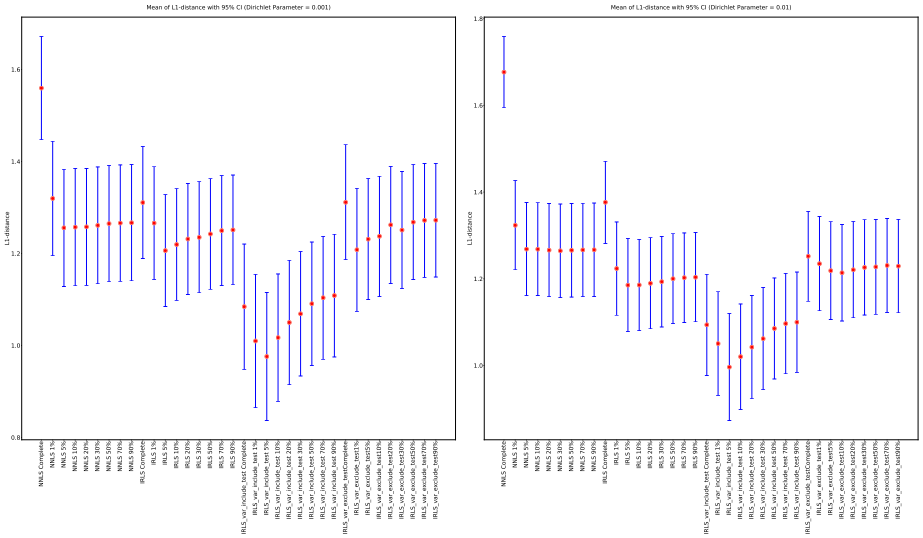
[1] Ting Gong and Joseph D Szustakowski. Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085, 2013.

[2] Gregory J Hunt, Saskia Freytag, Melanie Bahlo, and Johann A Gagnon-Bartsch. dtangle: accurate and robust cell type deconvolution. *Bioinformatics*, 35(12):2093–2099, 2019.

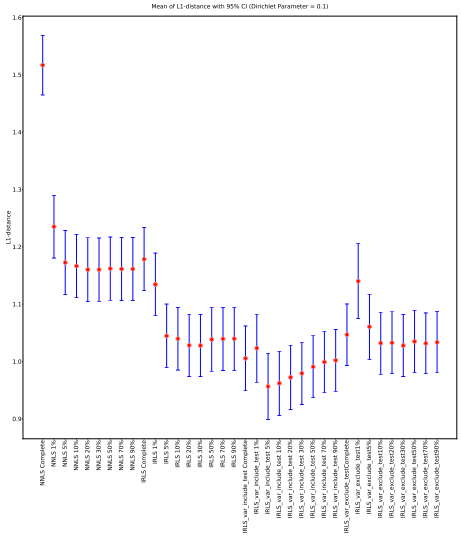
[3] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[4] Matthew N Bernstein, AnHai Doan, and Colin N Dewey. Metasra: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics*, 33(18):2914–2923, 2017.

[5] Chien-Ming Chen, Yu-Lun Lu, Chi-Pong Sio, Guan-Chung Wu, Wen-Shyong Tzou, and Tun-Wen Pai. Gene ontology based housekeeping gene selection for rna-seq normalization. *Methods*, 67(3):354–363, 2014.



(a) Dirichlet parameter (α) = 0.001 (b) Dirichlet parameter (α) = 0.01



(c) Dirichlet parameter (α) = 0.1

Figure 3: Benchmark of variance matrix based IRLS with other IRLS methods. The variance matrix is built using two different dataset. One includes the test studies, the studies that is used for generating the bulk data. The other one excludes the test studies.