UNDO: An open-source software tool for unsupervised deconvolution of tumor-stromal mixed expressions

Niya Wang

April 26, 2023

The R package UNDO (Unsupervised Deconvolution of Tumor-stromal Mixed Expressions) provides a completely unsupervised way to deconvolute tumor-stroma mixed expressions and exploit the existence of marker genes that do not need to be known in advance.

1 Introduction

Tumor-stroma interactions serve as both a major confounding factor and an underexploited information source in studying carcinogenesis [1]. Experimental solutions to isolate pure cells have many limitations [2]. All computational alternatives are limited by their need for prior knowledge of cell proportions or marker signatures to support a supervised deconvolution [3][4].

Within a well-grounded mathematical framework, we report a completely unsupervised method for deconvoluting tumor-stroma mixed expressions, exploiting the existence of marker genes that do not need to be known in advance. Fundamental to the success of our approach is the geometric identifiability of marker genes warranted by expression non-negativity.

UNDO begins by detecting the marker genes (genes whose expressions are exclusively enriched in tumor/stroma) located on scatter radii of mixed expressions. Based on the expression values of detected marker genes, UNDO estimates cell proportions by standardized average, and deconvolute the mixed expressions into tumor/stroma profiles via matrix inversion.

2 Overview of UNDO Package

This package includes five functions, which help realize the selection of marker genes and the calculation of mixing matrix and pure expression profiles. A single successful deconvolution only requires one input – mixed gene expression profiles, but the users can also input the real mixing matrix if it is known beforehand, so they can compare the ground truth with the deconvolution results from UNDO package.

1. `two_source_deconv`: This is the main function that call all the subfunctions to implement the whole deconvolution process. It requires a n-by-m

gene expression data matrix or ExpressionSet object as input. Users can choose to input the percentage of genes with minimum and maximum norms they want to move. Depend on the data quality, users can also choose to input the epsilon 1 and epsilon 2 which control the number of marker genes. The argument return decides whether to return the estimated pure expression profiles. These arguments, except expression matrix, will be set as default values if the users did not specify them. The output of this function are estimated mixing matrix and estimated pure expression profiles when return is set to 1. If the real mixing matrix is provided, the output will also include the value of E1 measurement. If the pure expression profiles are available, the function will also calculate the correlation between the pure and estimated expression profiles.The estimated mixing matrix and expression profiles will be saved in a newly generated folder under the workspace;

2. `gene_expression_input`: This function is called by two_source_deconv() to detect whether the input gene expression data matrix is valid. If the input is null, or contains negative values, the algorithm will terminate. If the input contains missing value, the correspondence rows will be deleted and warning information will be given. If the correlation coefficient between two samples are equal to one, which means the two samples are from the same source, the algorithm will terminate. The output of the function is the gene expression data matrix after passing the validity check;

3. `dimension_reduction`: When the input expression data contains more than 2 samples, principle component analysis will be used to reduce the sample dimension m to 2. It returns the expression values and dimension reduction matrix used to recover the mixing matrix for all m samples;

4. `marker_gene_selection`: This function implements the selection of marker genes. The output contains the marker gene list in two sources, and the slopes of marker genes which are used to calculate the estimated mixing matrix;

5. `mixing_matrix_computation`: This function computes the mixing matrix and pure expression levels based on the output of the function marker_gene_selection();

6. `calc_E1`: This function calculates the E1 measurement when the real mixing matrix is provided. When E1 is closer to 0, the estimated mixing matrix is closer to the real one. When the real mixing matrix is unknown, E1 is set to null.

Note that the input expression data should be after normalization, but without logarithmic transformation. Users can select the normalization method they prefer to normalize the raw data.

3 Theoretical Basis

Fundamental to the success of our approach is a geometric discovery of tumor or stroma-specific marker genes and expression non-negativity. We adopt the

linear latent variable model of raw measured expression data, given by (bold font indicates column vectors):

$$\mathbf{x}(i) = \mathbf{a}_1 s_{tumor}(i) + \mathbf{a}_2 s_{stroma}(i) \quad (1)$$

where $s_{tumor}(i)$ and $s_{stroma}(i)$ are the expression values for pure tumor and stroma tissues. $\mathbf{x}$ are the expression values for heterogeneous samples for genes $i = 1, ..., n$.

Since raw measured gene expression values are non-negative, when cell-specific marker genes exist for each cell type, the linear latent variable model (1) is identifiable using two or more mixed expressions, as we will elaborate via the following theorems.

`Theorem 1` (Scatter compression). Suppose that pure tissue expressions are non-negative and $\mathbf{x}(i) = \mathbf{a}_1 s_{tumor}(i) + \mathbf{a}_2 s_{stroma}(i)$ where $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly independent, then, the scatter plot of mixed expressions is compressed into a scatter sector whose two radii coincide with $\mathbf{a}_1$ and $\mathbf{a}_2$.

`Theorem 2` (Unsupervised identifiability). Suppose that pure tissue expressions are non-negative and cell-specific marker genes exist for each constituting tissue type, and $\mathbf{x}(i) = \mathbf{a}_1 s_{tumor}(i) + \mathbf{a}_2 s_{stroma}(i)$ where $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly independent, then, the two radii of the scatter sector of mixed expressions coincide with $\mathbf{a}_1$ and $\mathbf{a}_2$ that can be readily estimated from marker gene expression values with appropriate rescaling.

4 Deconvolution Analysis

In this section, we use a simple example to show how to use UNDO package to perform tumor-stroma deconvolution. We use the numerically mixed samples in this example so that we can compare the estimated mixing matrix with real one.

```
> library(UNDO)
> #load tumor stroma mixing tissue samples
> data(NumericalMixMCF7HS27)
> X <- NumericalMixMCF7HS27
> #load mixing matrix for comparison
> data(NumericalMixingMatrix)
> A <- NumericalMixingMatrix
```

The mixtures are from MCF7 and HS27 cell line expression with varying mixing proportions. There are 22215 probe sets in total. We use PLIER to perform normalization. The mixing proportions we used are shown as below:

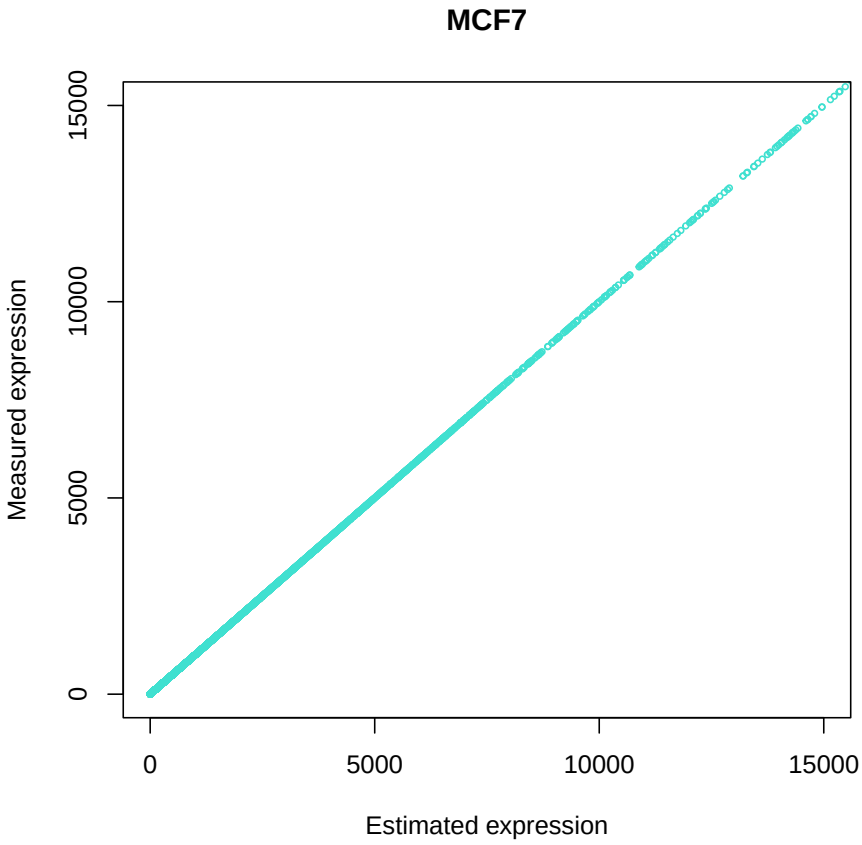
```
> A

            V1        V2
[1,] 0.7747503 0.2252497
[2,] 0.1501265 0.8498735
```

In the gene expression data, genes with extremely small norms are removed since they are easily to be influenced by noise, and genes with extremely large norms may be outliers, so we also remove them in the following analysis.

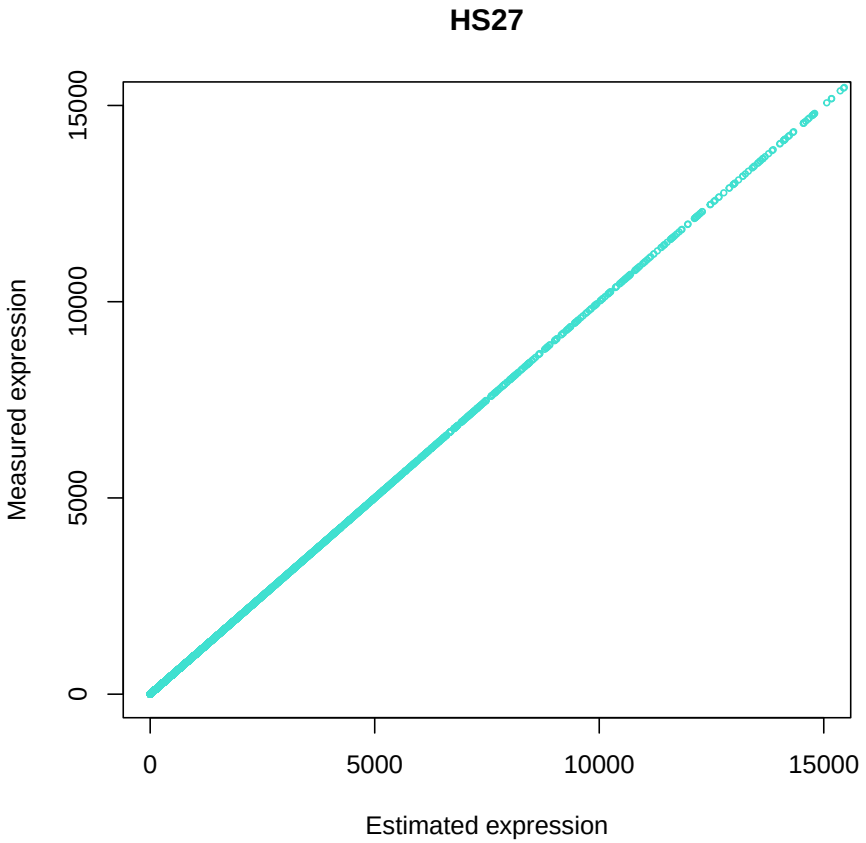
```
> #load pure tumor stroma expressions
> data(PureMCF7HS27)
> S <- exprs(PureMCF7HS27)
> two_source_deconv(X,lowper=0.4,highper=0.1,epsilon1=0.01,
+ epsilon2=0.01,A,S[,1],S[,2],return=0)

$Estimated_Mixing_Matrix
       [,1]      [,2]
1 0.7745182 0.2254818
2 0.1503423 0.8496577

$E1
[1] 0.001434569

$S1_correlation
[1] 1

$S2_correlation
[1] 1

>
```

Since we have ground truth in this case, we can compare the estimated expression of MCF7 and HS27 with pure expression level.

```
> # compute the estimated pure source expressions
> result <- two_source_deconv(X,lowper=0.4,highper=0.1,epsilon1=0.01,
+ epsilon2=0.01,A,S[,1],S[,2],return=1)
> Sest <- result[[5]]
> #draw the scatter plots between pure and estimated expressions of
> #MCF7 and HS27
> plot(S[,1],Sest[,1],main="MCF7" ,xlab="Estimated expression",
+ ylab="Measured expression", xlim=c(0,15000), ylim=c(0,15000),
+ pch=1, col="turquoise", cex=0.5)
```




```
> plot(S[,2],Sest[,2],main="HS27" ,xlab="Estimated expression",
+ ylab="Measured expression", xlim=c(0,15000), ylim=c(0,15000),
+ pch=1, col="turquoise", cex=0.5)
>
```



From the above results, E1 is very close to 0, representing the highly similarity between estimated mixing matrix and real mixing matrix. By observing the scatter plots from two cell lines, we find that the correlations between estimated and real expression profiles are as high as 1. Thus, we can conclude that UNDO successfully deconvolute the mixed samples from MCF7 and HS27.

5 Future Work

In the next version of UNDO package, we will add more dimension reduction functions so that users can select and compare different dimension reduction methods.

References

[1] Junttila, M.R., et al. Influence of tumour micro-environment heterogeneity on therapeutic response *Nature* 501, 346-354 (2013)

[2] Kuhn, A., et al. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain *Nat Methods* 8, 945-947 (2011)

[3] Shen-Orr, S.S., et al. Cell type-specific gene expression differences in complex tissues *Nat Methods* 7, 287-289 (2010)

[4] Ahn, J., et al. DeMix: deconvolution for mixed cancer transcriptomes using raw measured data *Bioinformatics* 29, 1865-1871 (2013)