

# ClusDec – clustering approach to solving complete deconvolution problem

Konstantin Zaytsev

Research advisor: Maxim Artyomov

ITMO University, WashU

November 18, 2015



- ▶ Gene expression analysis – widely used technique
- ▶ Samples are often mixed
- ▶ Heterogeneity confounds interpretation of data
- ▶ Deconvolution algorithms try to guess cell types and their proportion between samples

- ▶ Let  $S$  be an  $n \times k$  gene expression matrix that contains  $k$  cell types and  $n$  genes.
- ▶ Let  $W$  be a  $k \times m$  matrix where each column of  $W$  contains the frequencies of  $k$  cell types in a particular observation
- ▶ Let  $O$  be an  $n \times m$  expression matrix that contains the observed gene expression level, where  $n$  represents the number of genes and  $m$  is the number of observed tissue samples.

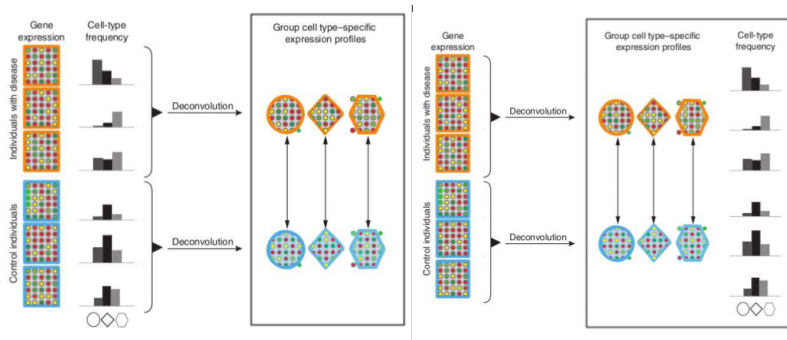
The mixing process can be modeled through a linear model:

$$O = S \times W$$

- ▶ Non-negativity: the resulting matrices  $S$  and  $W$  must be non-negative
- ▶ Sum-to-one constraint: the sum of every column of  $W$  must be equal to one

Deconvolution problem can be

- ▶ Partial (additional information)
- ▶ Complete



What kind of additional information can help us make deconvolution?

- ▶ S (gene expression in expected cell types)
- ▶ W (cell type proportions)
- ▶ Signature genes !

What is signature gene? We assume signature gene to be highly expressed in only one cell type.

$$S = \begin{pmatrix} \mathbf{g}_{1,1} & 0 & 0 & \dots & 0 & 0 \\ \mathbf{g}_{2,1} & \mathbf{g}_{2,2} & \mathbf{g}_{2,3} & \dots & \mathbf{g}_{2,k-1} & \mathbf{g}_{2,k} \\ 0 & \mathbf{g}_{3,2} & 0 & \dots & 0 & 0 \\ \mathbf{g}_{4,1} & \mathbf{g}_{4,2} & \mathbf{g}_{4,3} & \dots & \mathbf{g}_{4,k-1} & \mathbf{g}_{4,k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \mathbf{g}_{n-3,k-1} & 0 \\ 0 & 0 & 0 & \dots & \mathbf{g}_{n-2,k-1} & 0 \\ \mathbf{g}_{n-1,1} & \mathbf{g}_{n-1,2} & \mathbf{g}_{n-1,3} & \dots & \mathbf{g}_{n-1,k-1} & \mathbf{g}_{n-1,k} \\ 0 & 0 & 0 & \dots & 0 & \mathbf{g}_{n,k} \end{pmatrix}$$

In this example genes 1, 3, n-3, n-2, n are signatures

All known approaches to deconvolution solve partial deconvolution problem:

- ▶ DSA (Zhong et al. BMC Bioinformatics 2013)
- ▶ CIBERSORT (Newman AM, Liu CL et al. Nat Methods 2015)
- ▶ NMF approaches (Gaujoux 2012)
- ▶ and more

There's **no complete** deconvolution algorithm



The main idea:

- ▶ Somehow find signature genes in dataset
- ▶ Perform signature gene-based partial deconvolution algorithm



How do we get signature genes from dataset?  
"Clus" is for clustering ! Usual practice in gene expression analysis is to use correlation as a distance

- ▶ Kmeans
- ▶ WGCNA clustering

Signature genes will definitely fall into the same cluster

But just correlation is not enough

- ▶ Lets assume we know the proportions of one particular cell type  $c$  ( $c$ -th row in matrix  $W$ )
- ▶ How does signature to this cell type gene expression have to look like? Since we assume that signature gene  $i$  is only expressed in one cell type  $c$ :

$$O_i = W_c * O_{i,c}$$

- ▶ Thus any pair of signature to one cell type gene expression profiles must be linearly dependent
- ▶ But if we use correlation we will also put into the same cluster pair of genes which are just collinear:

$$O_i = A + B * O_j$$



What can we use as a distance so? Lets just remove means from correlation

$$d_{x,y} = \frac{\sum_{i=1}^n (x_i)(y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \sqrt{\sum_{i=1}^n (y_i)^2}} = \frac{\langle x, y \rangle}{\|x\| * \|y\|}$$

Lets assume we found the clusters. What's next?

- ▶ Take some of these clusters (ones that have enough genes and correlation inside the cluster is good enough)
- ▶ Use genes in these clusters as signatures in partial deconvolution algorithm for every combination of  $k$  clusters from good clusters
- ▶ For every combination measure the accuracy of deconvolution
- ▶ Choose the best describing combination of cluster

- ▶ Takes list of signature genes as input
- ▶ Every signature expression profile is then divided by its maximum as normalization
- ▶ For every cluster we take mean of gene profiles thus we will have  $k$  vectors corresponding to "raw" proportions of these cell types
- ▶ Fit sum-to-one constraint: find vector  $C$  such that

$$\tilde{W}_{m \times k} \times C_{k \times 1} = 1_{m \times 1}$$

- ▶ And the resulting  $W$  is

$$W_{k \times m} = \left( \tilde{W}_{m \times k} \times \begin{vmatrix} C_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & C_k \end{vmatrix} \right)^T$$

# The idea: deconvolution

When we found  $W$ , we run DSA with given  $W$  to find  $S$  matrix. These  $W$  and  $S$  (and signature genes as well) are deconvolution results

This question is next to "how to interpret deconvolution results?"

- ▶ Proportions of cell types and signature genes are the results of deconvolution
- ▶ No one can guarantee that we will find biology in these genes
- ▶

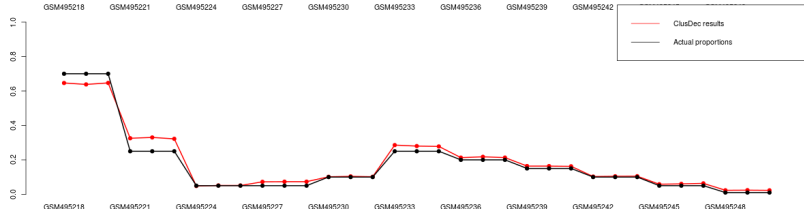
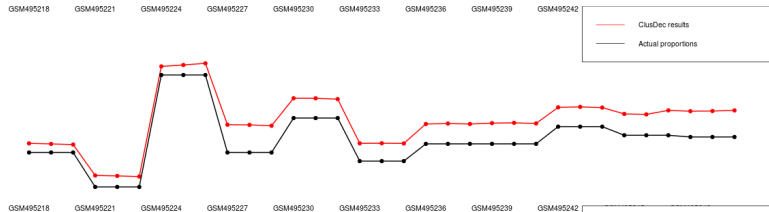
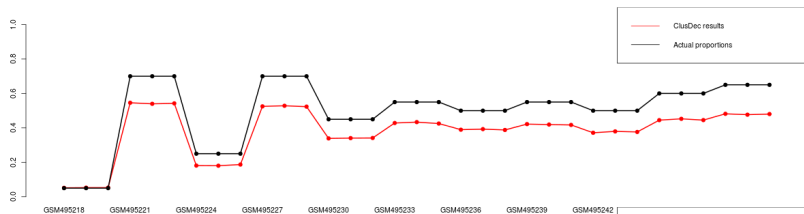
How to measure:

- ▶ How good we fit sum-to-one constraint
- ▶ Correlation between samples in  $O$  and  $S \times W$

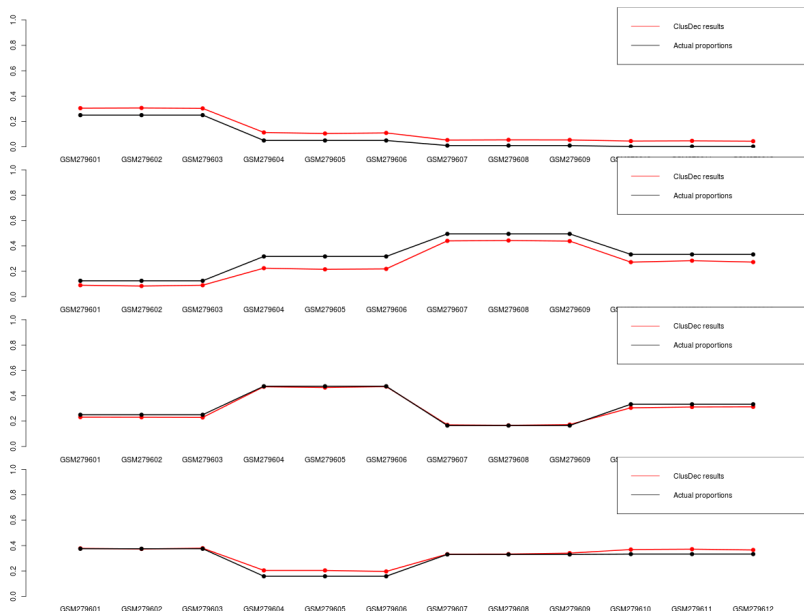


- ▶ Benchmark datasets: GSE19830 (the mixture of brain, lung and liver tissues), GSE11058 (the mixture of immune cell lines Jurkat, IM-9, Raji, THP-1). We know the actual proportions of cell types in these datasets
- ▶ GSE52245: human PBMC after vaccination
- ▶ GSE27563: mice PBCs with tumors
- ▶ GSE35710: adipose tissues

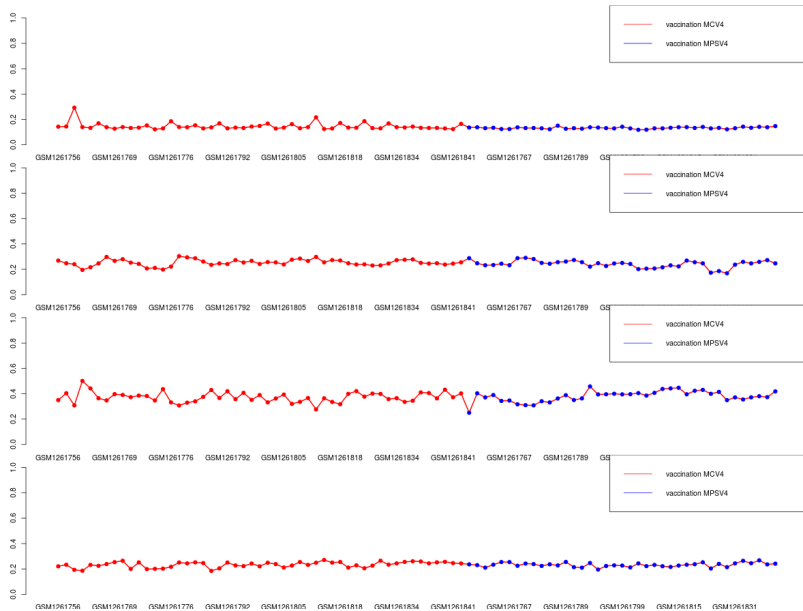
# Some results: GSE19830



# Some results: GSE11058



# Some results: GSE52245



- ▶ ClusDec works pretty well on benchmark dataset
- ▶ ClusDec can find some biology
- ▶ Interpretation of results is not easy at all

- ▶ Understanding what cluster is good is still vague
- ▶ How to check accuracy of deconvolution is still an open question
- ▶ Current cluster size on practice is about 20-60 genes. Can we enlarge it to make result easier to interpret

Thank you

Any questions?