Deconvolution

Lou Shaoke

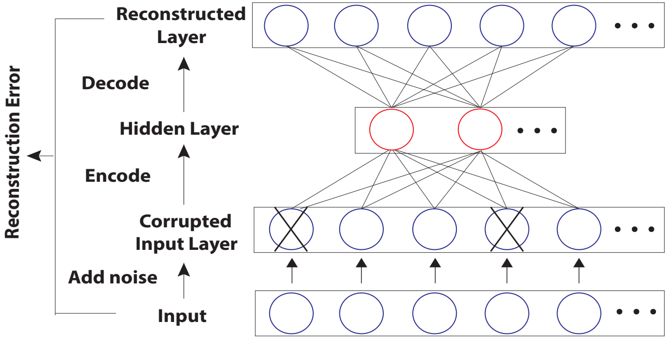
Department of Molecular Biophysics and Biochemistry *loushaoke@gmail.com*

August 12, 2015

Yale

Denosing autoencoder on sputum

Diagram



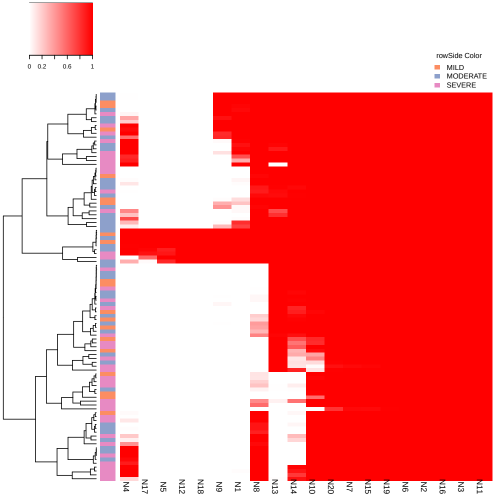
$y = sigmod(Wx + b)$
$z = sigmod(W^T y + b\prime)$
$L = \sum(-xlog(z) - (1 - x)log(1 - z))$
Then use stochastic gradient descent to find local minimum.
$\#visible(v) = \#gene, \#hidden(h) \in 20, 30, 50, \ldots,$ learning rate=0.01, corruption level=0.05; cycle = 100

Result

Weight matrix(W, $h \times v$), represent the contribution of each gene for each node. The hidden value(Y, hs), can be thought as the activity value of each node in each sample(s)

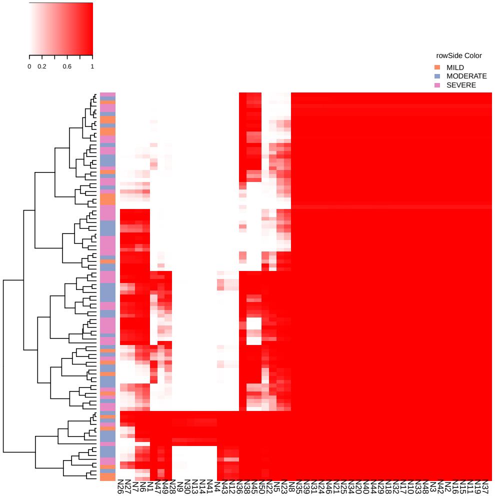
Clustering and Asthma severity

100 cycle, 20 hidden node,learning rate 0.2



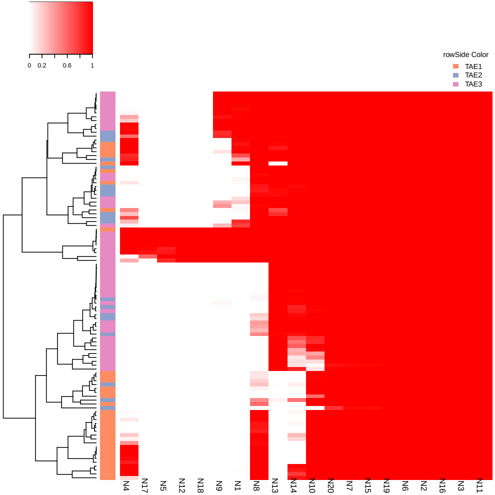
Clustering and Asthma severity

100 cycle, 50 hidden node,learning rate 0.1 and corruption level 0.01



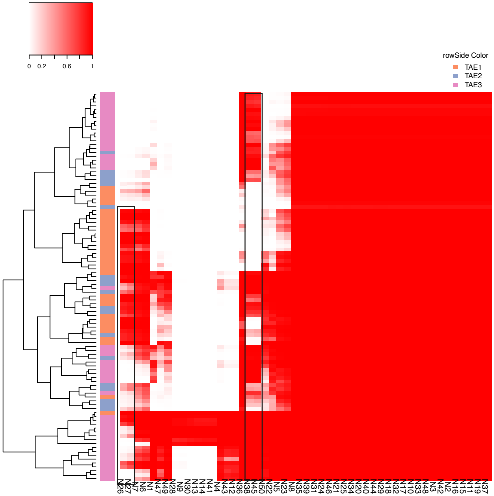
Clustering and TAE cluster

100 cycle, 20 hidden node,learning rate 0.2

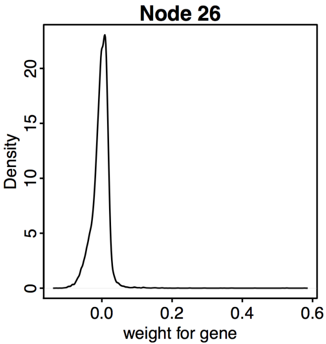


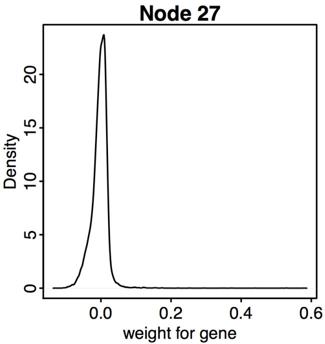
Clustering and TAE cluster

100 cycle, 50 hidden node,learning rate 0.1 and corruption level 0.01



High discriminantal node





We define genes contribution into three classes: extreme_high(H), normal, extreme_low(L) by $\mu \pm 2\sigma$

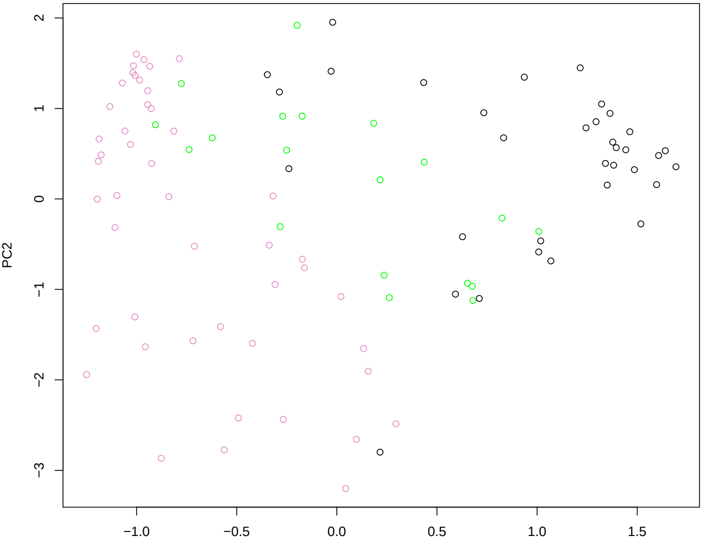
xx	node26	node27	both
H	232	246	232
L	713	681	676

The same way, we characterize node 38,45, 50.

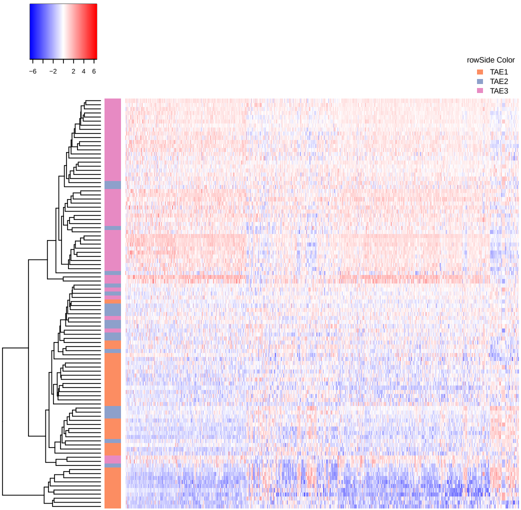
Function annotation

1. Highly Enrichment in keratinization for TAE1 specific high weighted gene. Asthma associate with keratinization.
2. Fatty Acid related process
3. Hormone response related process

Another validation

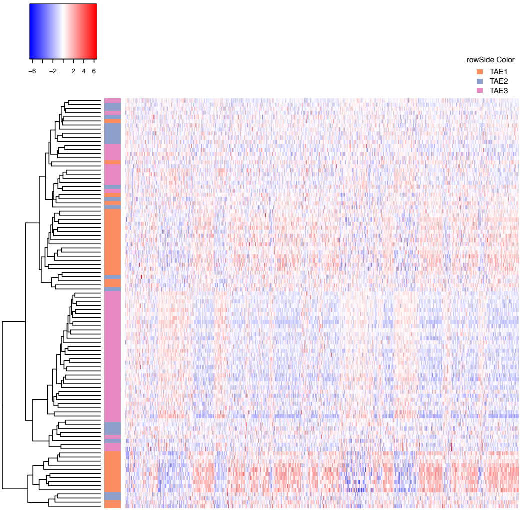


Another validation



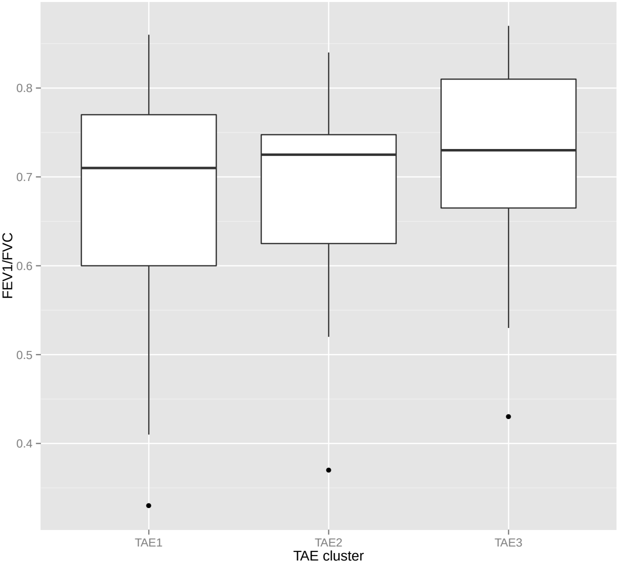
Clustering using genes expression from above nodes

Another validation



Clustering using random selected genes

FEV1/FVC



Measurement

1. purity
   $Purity(\Omega, C) = \frac{1}{N} \sum_k max_j |\omega_k \cap c_j|$
2. normalized mutual information
   $NMC(\Omega, C) = \frac{I(\Omega; C)}{(H(\Omega)+H(C))/2}$
3. Rand Index:
4. F-value:

Clinical background

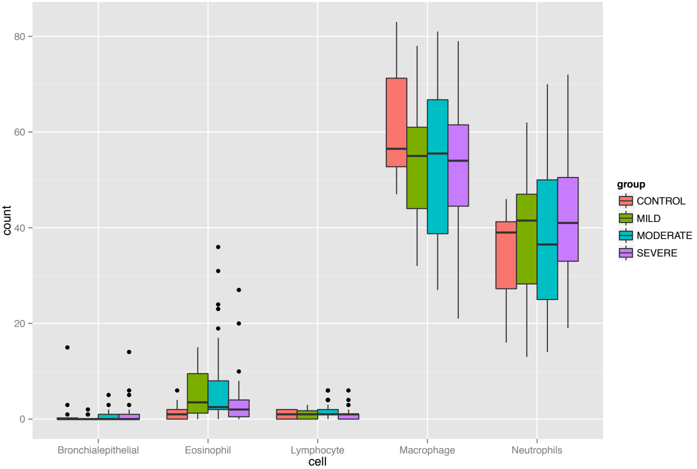
Eosinophilic and Neutrophilic Inflammation in Asthma. eosinophilic and neutrophilic inflammation in asthma, and they are not mutually exclusive subtype.

Neutrophils are prominent in airway secretions during acute severe asthma

Macrophages exert prominent effects in the defense of the respiratory tract from airborne pathogens

Distinct cellular subtypes of asthma based on the presence or absence of sputum granulocytesnamely, eosinophilic asthma (EA), neutrophilic asthma (NA), mixed eosinophilic and neutrophilic asthma (ME/NA) and paucigranulocytic asthma (PGA)

Cell types across samples



Decomposition

1. 5 cell types, all sample: control vs astham
2. 5 cell types, samples: control vs severe asthma
3. 3 cell types, all sample
4. 3 cell type, control versus severe asthma

Given FDR$<$0.05, all above comparison have no significant genes, if FDR ¡ 0.1, the last comparison finds :DEFA3, DEFA1 and RPS4Y1(robosomal protein) Microphase cell line.

Followup

- Expression value, log-based?

- Conventionally, use cell count to determine subtypes, know more about the experiment design

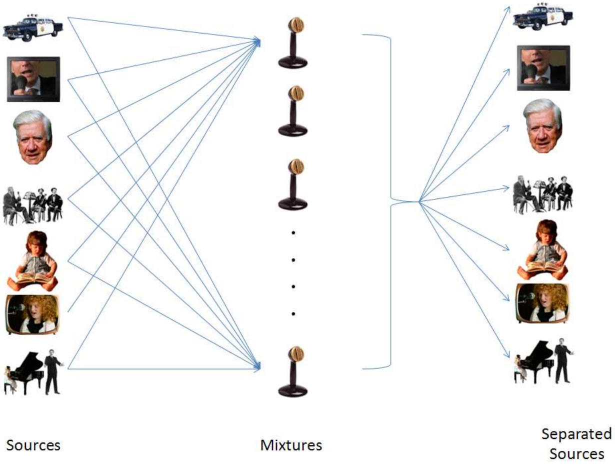
- other medical information: disease duration, medication etc

- cellular changes $<$ cell proportion changes, inflammatory response, proportion or count change is the best and easiest way to do it.

- meta analysis and supervised learning

Yale

Cocktail Party



http://research.ics.aalto.fi/ica/cocktail/cocktail_en.cgi

Problem Definition

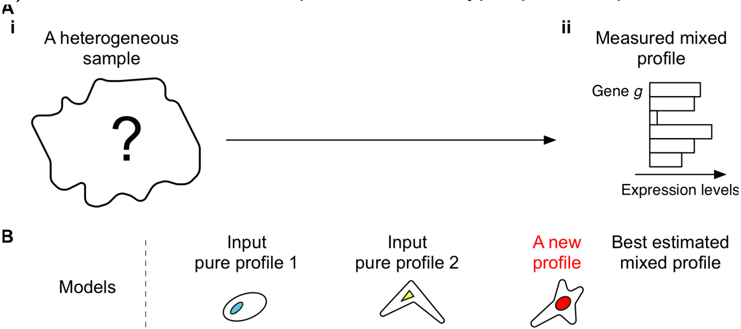
Given a set of mixed gene expression sets X_{gs} for gene $g \in 1, 2, ..., G$ and sample $s \in 1, 2, ..., m$. The samples can be from case-control tissue, different tissues type and blood sample etc. Due to the sample hetergeneous, the gene expression should be a mixture of expression of different cell type/condition $w \in 1, 2, ..., W$.

The motivation:

Problem Definition

Given a set of mixed gene expression sets X_{gs} for gene $g \in 1, 2, ..., G$ and sample $s \in 1, 2, ..., m$. The samples can be from case-control tissue, different tissues type and blood sample etc. Due to the sample hetergeneous, the gene expression should be a mixture of expression of different cell type/condition $w \in 1, 2, ..., W$.

The motivation:
1) Can we deconvolve the expression to cell-type specific expr? csSAM, PERT



Problem Definition

Given a set of mixed gene expression sets X_{gs} for gene $g \in 1, 2, ..., G$ and sample $s \in 1, 2, ..., m$. The samples can be from case-control tissue, different tissues type and blood sample etc. Due to the sample hetergeneous, the gene expression should be a mixture of expression of different cell type/condition $w \in 1, 2, ..., W$.

The motivation:
1) Can we deconvolve the expression to cell-type specific expr? csSAM, PERT
2) Can we deconvolve the expression to cell-type like value/latent? for example: cancer, control;(DeMix, ISOpure etc)

Linear Deconvolution

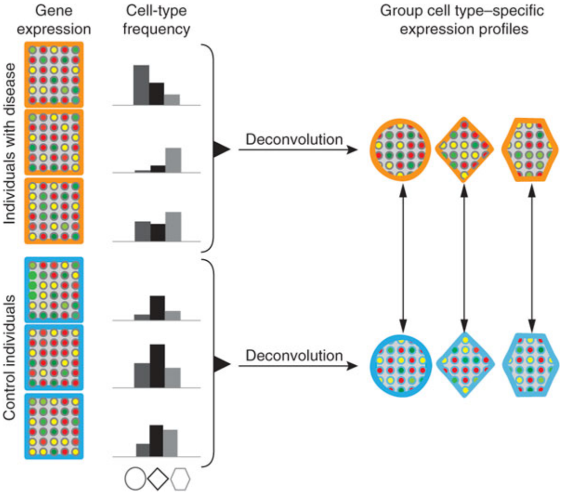
$X = A \times S$

Linear Deconvolution

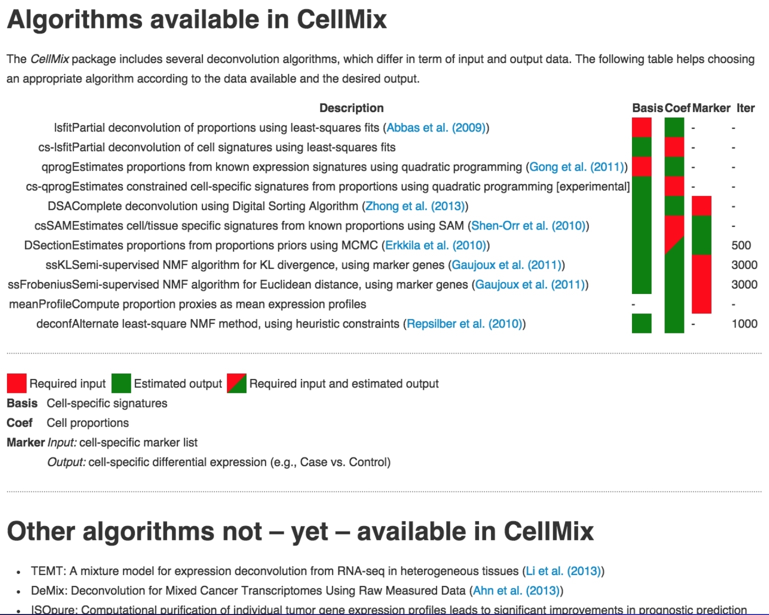
$X = A \times S$

SS Shen-Orr (Nature method 2010): Two-group Model

$X_{ij} = \sum_{k=1}^{K} w_{ik} h_{kj}^{(1)} + e_{ij}$ and $X_{ij} = \sum_{k=1}^{K} w_{ik} h_{kj}^{(2)} + e_{ij}$



Algorithms



Algorithms available in CellMix

The *CellMix* package includes several deconvolution algorithms, which differ in term of input and output data. The following table helps choosing an appropriate algorithm according to the data available and the desired output.

Description	Basis	Coef	Marker	Iter
lsfitPartial deconvolution of proportions using least-squares fits (Abbas et al. (2009))	🟥	🟩	-	-
cs-lsfitPartial deconvolution of cell signatures using least-squares fits	🟩	🟥	-	-
qprogEstimates proportions from known expression signatures using quadratic programming (Gong et al. (2011))	🟥	🟩	-	-
cs-qprogEstimates constrained cell-specific signatures from proportions using quadratic programming [experimental]	🟩	🟥	-	-
DSAComplete deconvolution using Digital Sorting Algorithm (Zhong et al. (2013))	🟩	🟩	🟥	-
csSAMEstimates cell/tissue specific signatures from known proportions using SAM (Shen-Orr et al. (2010))	🟩	🟥	🟩	-
DSectionEstimates proportions from proportions priors using MCMC (Erkkila et al. (2010))	🟩	◩	🟩	500
ssKLSemi-supervised NMF algorithm for KL divergence, using marker genes (Gaujoux et al. (2011))	🟩	🟩	🟥	3000
ssFrobeniusSemi-supervised NMF algorithm for Euclidean distance, using marker genes (Gaujoux et al. (2011))	🟩	🟩	🟥	3000
meanProfileCompute proportion proxies as mean expression profiles	-	🟩	🟥	-
deconfAlternate least-square NMF method, using heuristic constraints (Repsilber et al. (2010))	🟩	🟩	-	1000

🟥 Required input 🟩 Estimated output ◩ Required input and estimated output

Basis Cell-specific signatures

Coef Cell proportions

Marker *Input:* cell-specific marker list

Output: cell-specific differential expression (e.g., Case vs. Control)

Other algorithms not – yet – available in CellMix

- TEMT: A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues (Li et al. (2013))
- DeMix: Deconvolution for Mixed Cancer Transcriptomes Using Raw Measured Data (Ahn et al. (2013))
- ISOpure: Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction

Issue

- 1. technical reasons and data transformation. Yi Zhong et al 2012 response to SS Shen-Orr.
- 2. Theoretical and pratical. How to evalute the results? It is good if more DEGs were found?
  Celltype-wise seperation?
  blind expression seperation? especially for more complex situation. such as metatstasis tissue with adjacent and original tissue.

Our options?

- - use known algorithms to explore functional
- - From the practical view: diagnosis and prognosis
  blood test: marker and diagnosis (require clinical information)
- - Metastasis
  Seed and soil
- - Combination?
  Origin site → blood → Metastasis