Deconvolution

Lou Shaoke

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

July 29, 2015

Yale

Clinical background

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

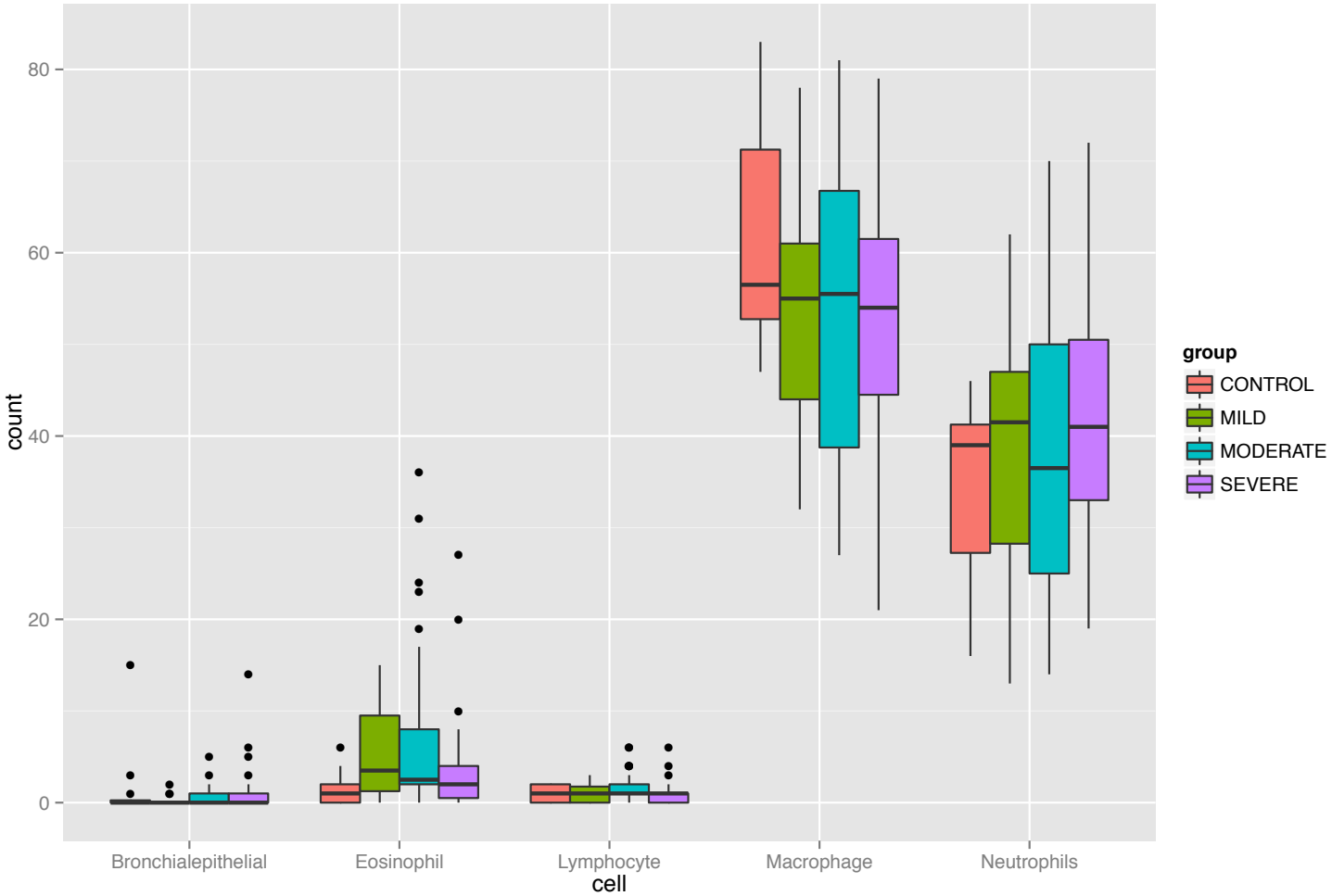
Neutrophils are prominent in airway secretions during acute severe asthma

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)

Yale

Cell types across samples



Yale

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.

Yale

Followup

- Expression value, log-based?

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

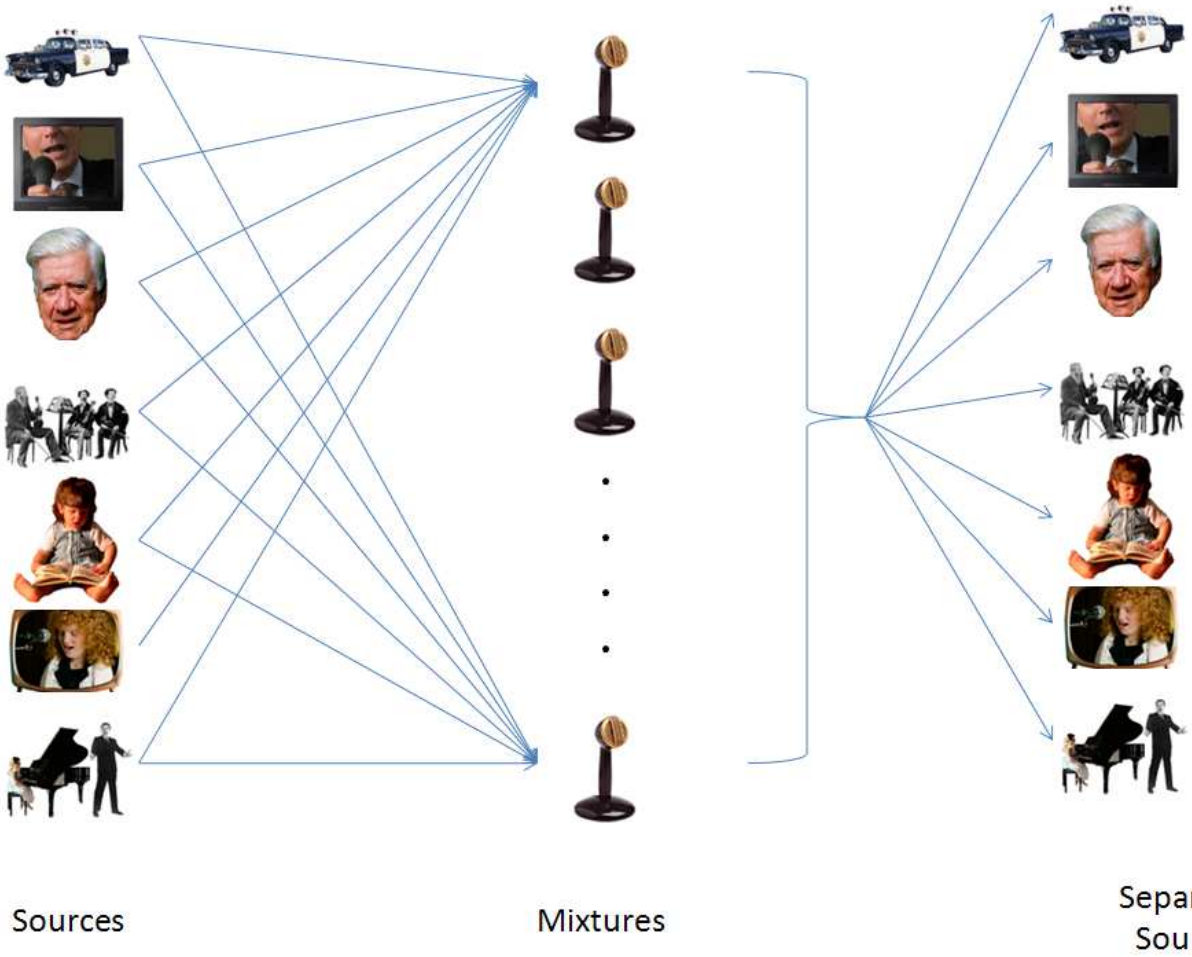
- other medical information: disease duration, medication etc

- 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

Yale

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 $w \in 1, 2, ..., W$.

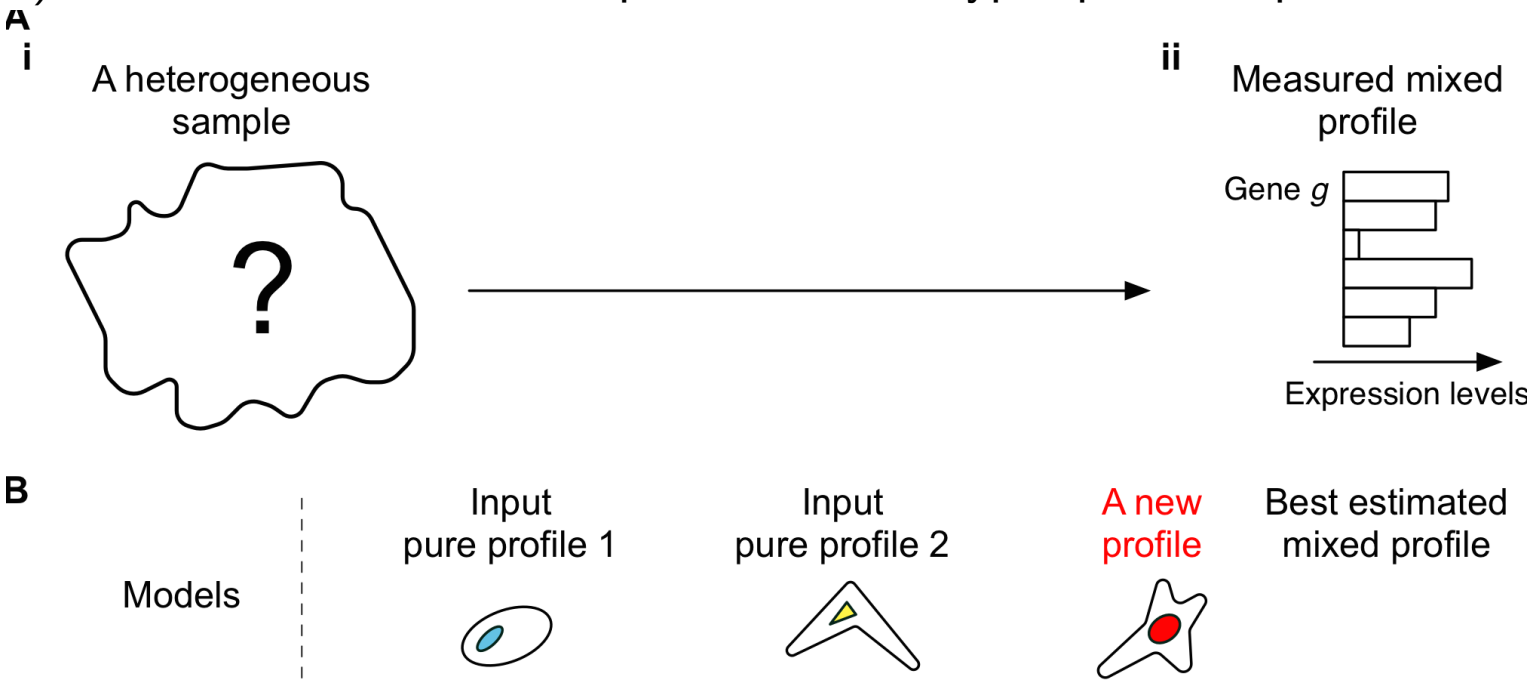
The motivation:

Yale

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)

Yale

Linear Deconvolution

$X = A \times S$

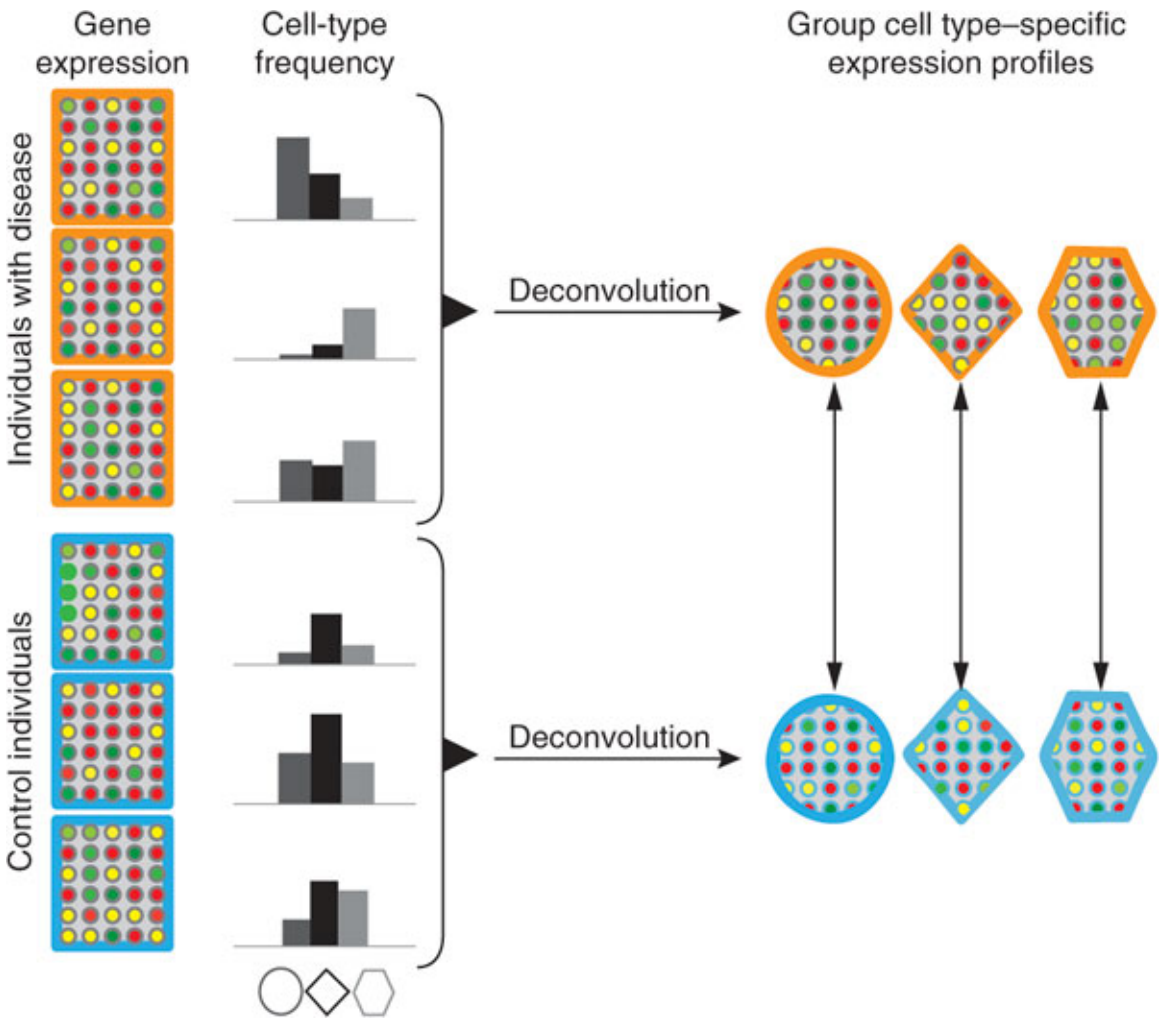
Yale

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



Yale

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

Yale

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.

Yale

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 $\rightarrow$ blood $\rightarrow$ Metastasis

Yale