Deconvolution of sputum gene expression and possible directions

Lou Shaoke

Department of Molecular Biophysics and Biochemistry

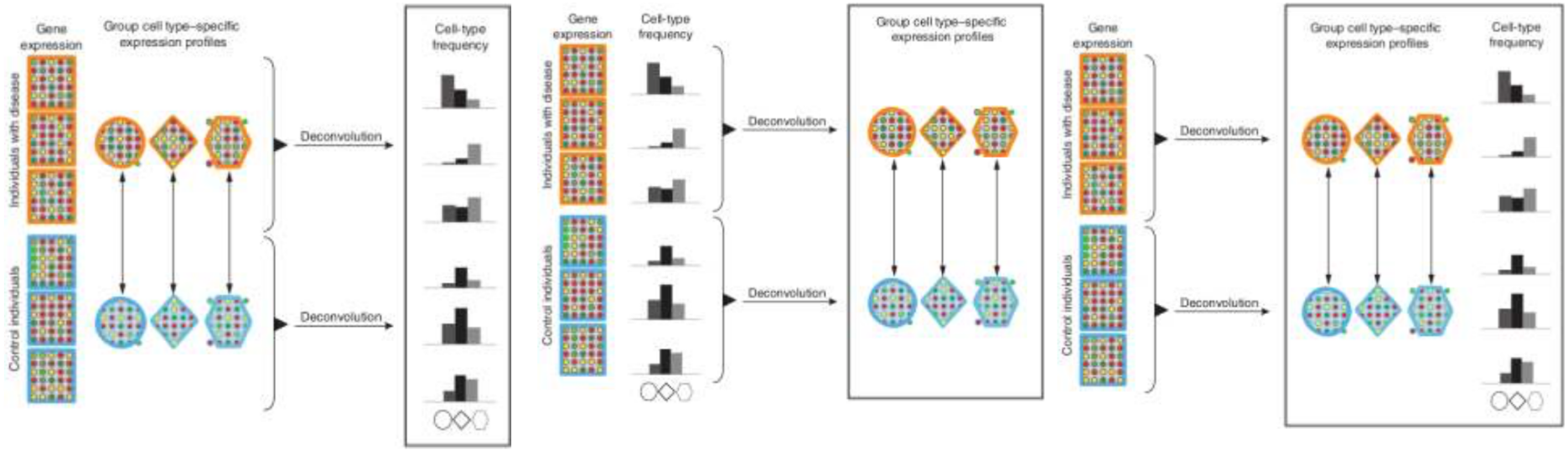
loushaoke@gmail.com

May 27, 2015

Yale

Background

(a) *Partial from available signatures* **(b)** *Partial from available proportions* **(c)** *Complete from global expression*



Yale

Deconvolution

- csSAM: based on standard least-squre regression
  -csSAM require the proportion of cell types; it can infer case-control significant gene
  - Step 1. Get the gene expression matrix (nxg, n sample, g Genes); proportion of cell type (nxc, n samples, c proportion of cell types); y group vector with length of n, 1 for case, 2 for control;
  - Step 2. first estimate the gene expr for case and control, using OLS
  - Step 3. calcualte the t-score for case and control (SAM)
  - Step 4. Permutation to estimate FDR and get the significant genes (Errors in this step, may because of colinearity after the permutation.
- Deconf: Non-negative matrix factorization -Only need to provide the number of cell types in the mixture

Yale

Datasets

- 112 samples, 12 control, and 100 cases
- propotion of sixe cell types

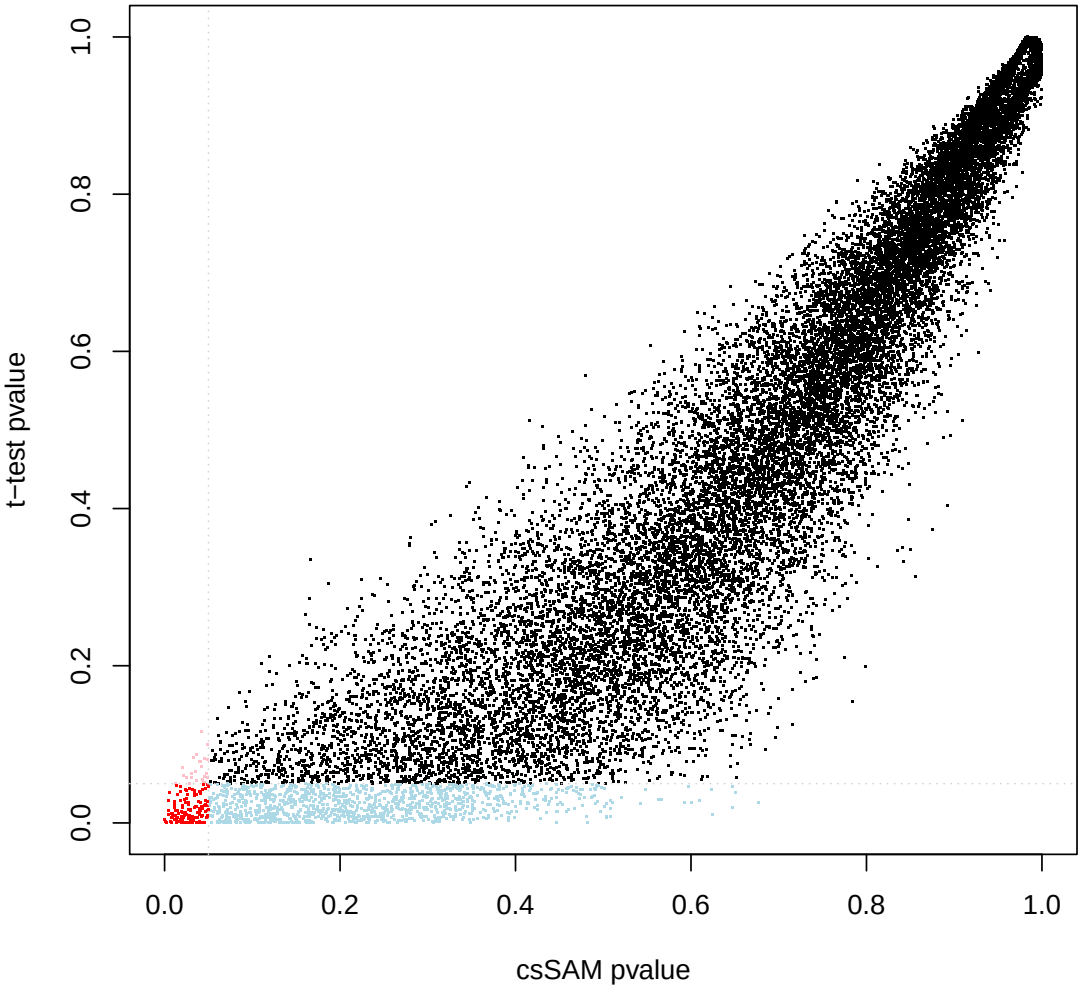
Table 1C: Sputum Characteristics of TEA Clusters in the YCAAD Cohort

	Controls(N=12)	Cluster 1 (N=34)	Cluster 2 (N=19)	Cluster 3 (N=47)	P Value
Mucus Cell Concentration	40.86±20.98[P]	83.02±105.75[P]	89.23±143.61[P]	73.72±62.48[P]	0.63
Squamous (%)	8.2±6.7	7.9±7.0	8.0±5.9	9.2±6.9	0.60
Viability (%)	58.1±9.6	56.5±16.1	64.4±11.9	61.7±17.8	0.14
Neutrophils (%)	34.6±10.0	41.5±13.0	41.9±15.2	37.8±14.6	0.34
Eosinophil (%)	1.5±1.8	5.8±6.7	4.7±5.9	5.2±7.7	0.91
Macrophage (%)	61.3±11.8	50.9±13.0	50.9±16.0	55.4±15.4	0.31
Lymphocyte (%)	1.0±0.9	1.3±1.5	1.2±1.0	1.3±1.4	0.90
Bronchial epithelial cell (%)	1.6±4.3	0.8±1.5	1.3±3.3	0.4±1.0	0.26
RIN (mean)	7.6±1.1	7.4±1.2	7.5±1.0	7.7±1.4	0.1

[P] Cells/Microliter x 10^4

Yale

csSAM results



#genecsSAM $p < 0.05$: 157; #genet-test $p < 0.05$:1347

Yale

Least-squares NMF

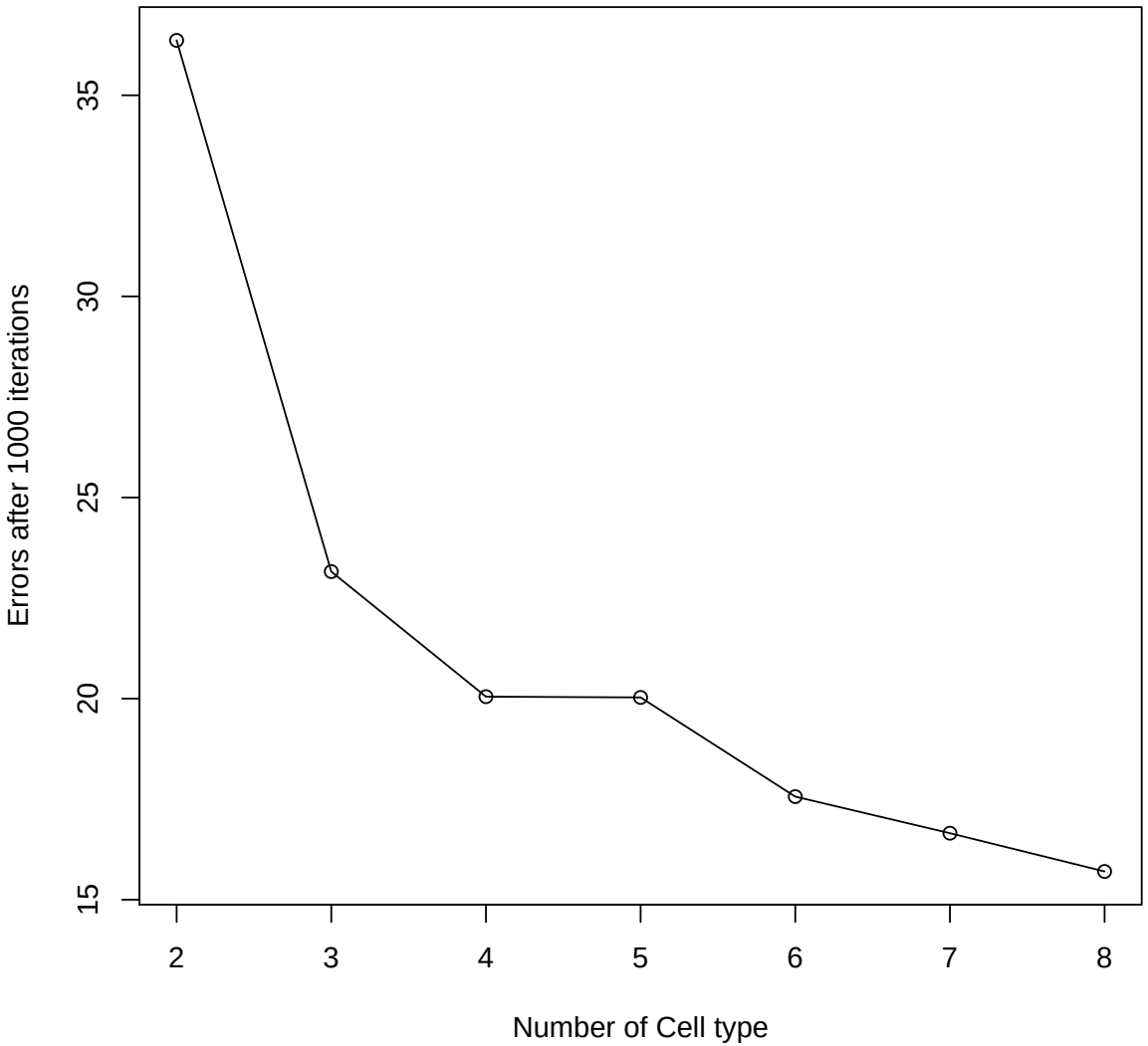
X=SC

input X and n
normalize columns of X (either centre, or by quantile normalization)
generate start values for S and C
apply constraints to S and C (see below)
(*) fix S, calculate C using lsqnonneg-algorithm
apply constraints for S
fix C, calculate S using lsqnonneg-algorithm
apply constraints for C
if $|X - SC| < a$ or number iterations $>$ b then EXIT and report S and C
else continue at (*)
Constrains:
1. S non-negative and normalized (either centered, or by quantile normalization)
2. $0 \leq cij \leq 1$ for all elements of C (cell type i, sample j)
3. $\sum_i c_{ij} = 1$ for all samples j (i.e. cell type proportions sum to 100%)

Yale

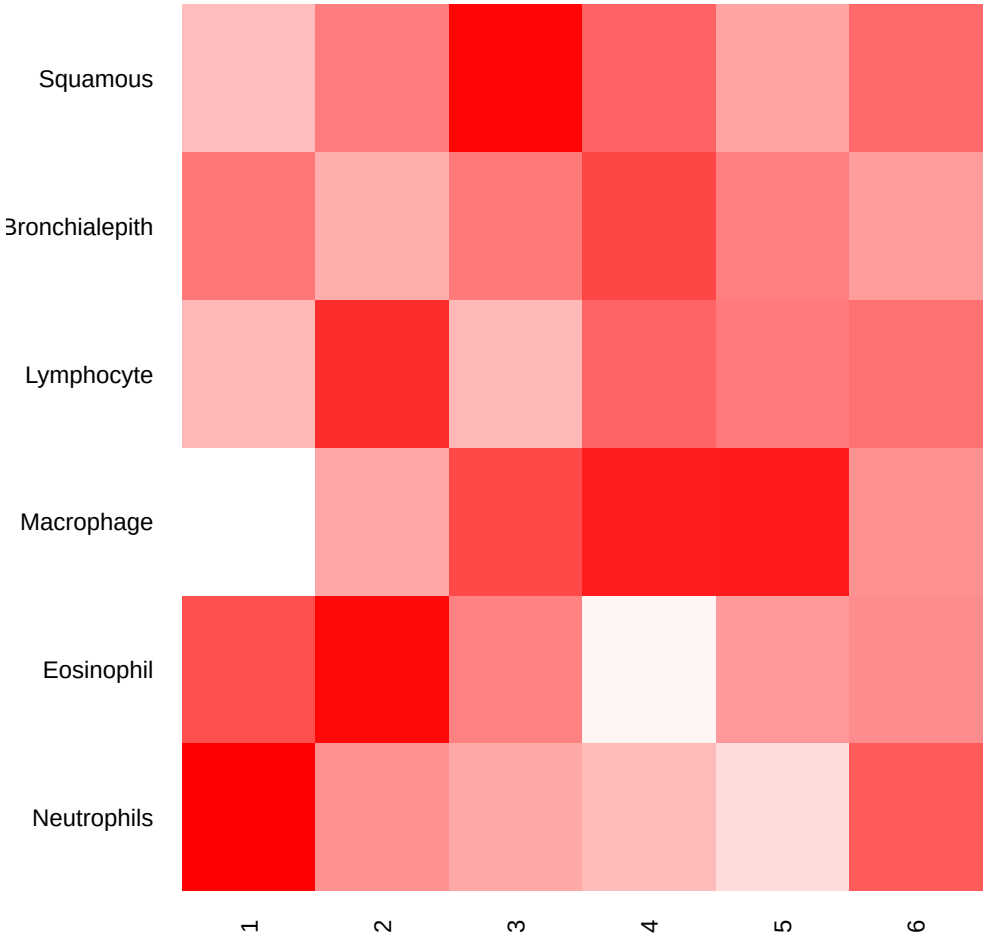
Proprotion from NMF

Evaluations:



Proprotion from NMF

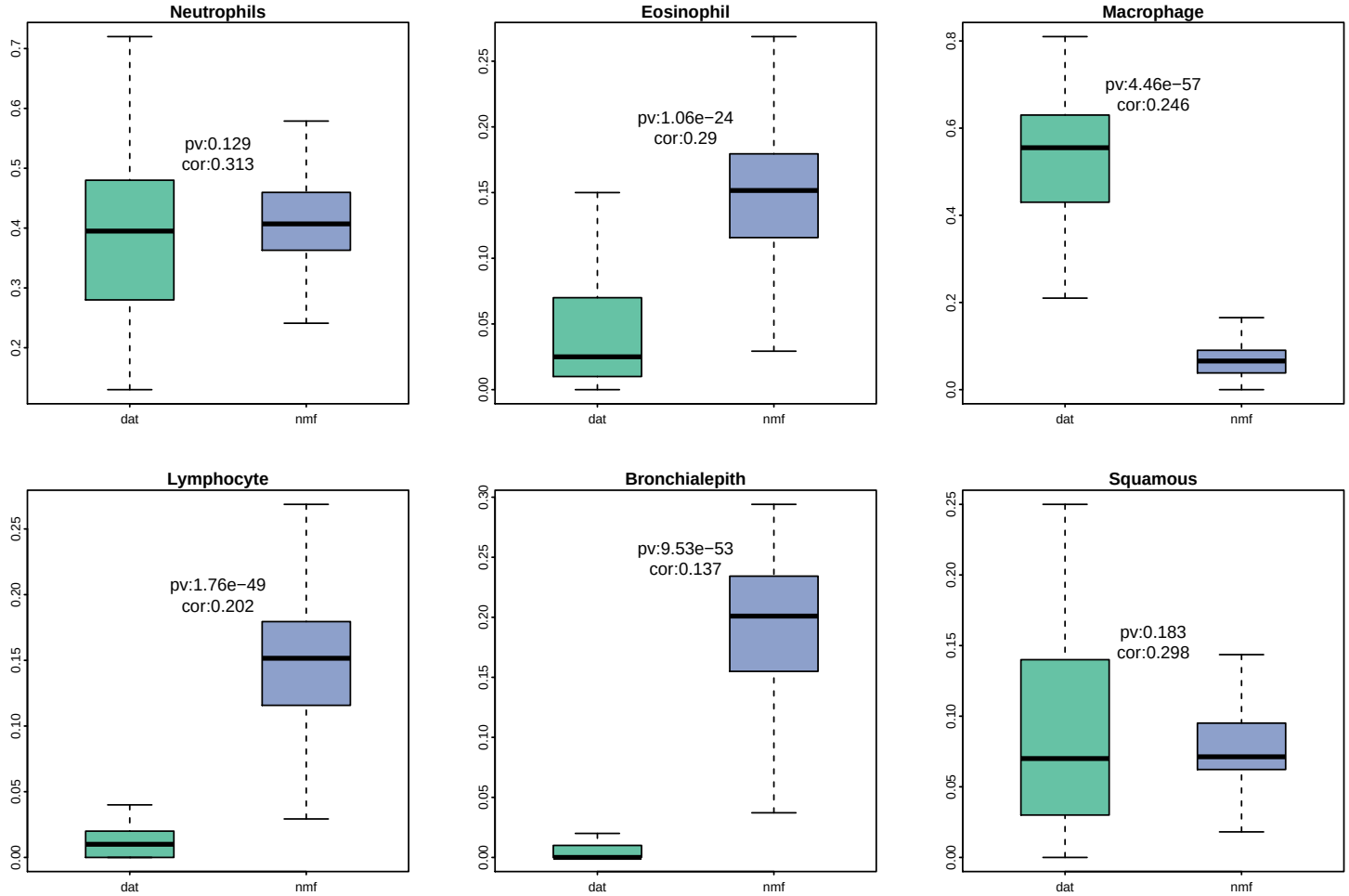
Evaluations:



Yale

Proprotion from NMF

Evaluations:



Possible Reasons

- algorithms
- not consistent with experiment
- Estimation based on error may not be reliable
- noises for the data in the microarray and RIN adjusted

Yale