Dissecting Differential Signals in High-throughput Data from Complex Tissues

Ziyi Li[1], Zhijin Wu[2], Peng Jin[3], and Hao Wu[1, *]

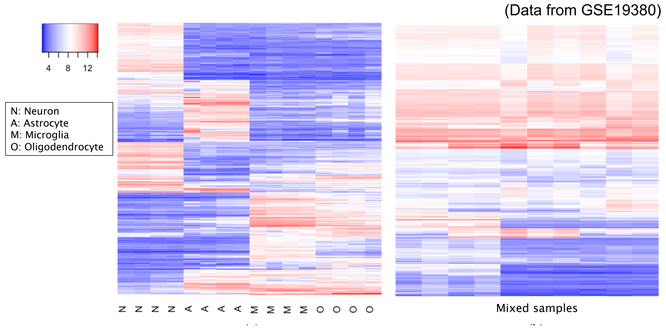
[1]Department of Biostatistics and Bioinformatics, Emory University; [2]Department of Biostatistics, Brown University; [3]Department of Human Genetics, Emory University.

* To whom the corresponding should be addressed: hao.wu@emory.edu

INTRODUCTION

- High-throughput technologies have been applied in larger-scale, population level clinical studies to identify diagnostic biomarkers and therapeutic targets (e.g. The Cancer Genome Atlas, The Rush Memory and Aging Project)
- These samples (blood, tumor, or brain) are **mixtures of many different cell types**
- Canonical differential expression (DE) and differential methylation (DM) analysis **fail to**
  - **adjust for cell compositions in complex tissue**
  - **reveal cell-type specific DE/DM (csDE/DM)**

(Data from GSE19380)



- Profile the purified cell types experimentally: **cell-sorting technology** - laborious and expensive.
- *In silico* identification of cell type specific effects:
  - **Estimation of mixture proportion**
    - reference-based deconvolution
    - reference-free deconvolution
  - **Identify csDE/DM**
    - cell-type specific significance analysis of microarrays (csSAM): two-step approach results in lower statistical efficiency
    - population-specific expression analysis (PSEA): relies heavily on cell-type specific marker genes

METHODS

Assume data generated from high-throughput experiments contain G features (genes or CpG sites) and N samples.

- Y_{gi}: measurement for g-th feature and i-th sample
- K: number of "pure" cell types
- $\boldsymbol{\theta_i} = (\theta_{i1}, \theta_{i2}, \cdots, \theta_{iK})$: mixing proportions for sample i (with constraint $\sum_k \theta_{ik} = 1$)
- X_{gik}: the underlying, unobserved expression in the k-th cell type for the g-th gene in the i-th sample
- $\boldsymbol{Z_i}$: subject-specific covariates ($Z_i = 0$ for controls and $Z_i = 1$ for cases)

$$E[X_{ik}] = \mu_k + \boldsymbol{Z_i^T \beta_k}$$

$$E[Y_i;\ \boldsymbol{\theta_i}] = \sum_k \theta_{ik} E[X_{ik}] = \sum_k (\theta_{ik}\mu_k + \theta_{ik} \cdot \boldsymbol{Z_i^T \beta_k})$$

Assume we have measurements $\boldsymbol{Y}$ from a total of N samples.

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix},\ \boldsymbol{W} = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1K} & \theta_{11} \cdot \boldsymbol{Z_1^T} & \theta_{12} \cdot \boldsymbol{Z_1^T} & \cdots & \theta_{1K} \cdot \boldsymbol{Z_1^T} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2K} & \theta_{21} \cdot \boldsymbol{Z_2^T} & \theta_{22} \cdot \boldsymbol{Z_2^T} & \cdots & \theta_{2K} \cdot \boldsymbol{Z_2^T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{NK} & \theta_{N1} \cdot \boldsymbol{Z_N^T} & \theta_{N2} \cdot \boldsymbol{Z_N^T} & \cdots & \theta_{NK} \cdot \boldsymbol{Z_N^T} \end{bmatrix}.$$

The observed data can be described as a linear model:

$$E[\boldsymbol{Y}] = \boldsymbol{W\beta}$$

Statistical inference for differential analysis:

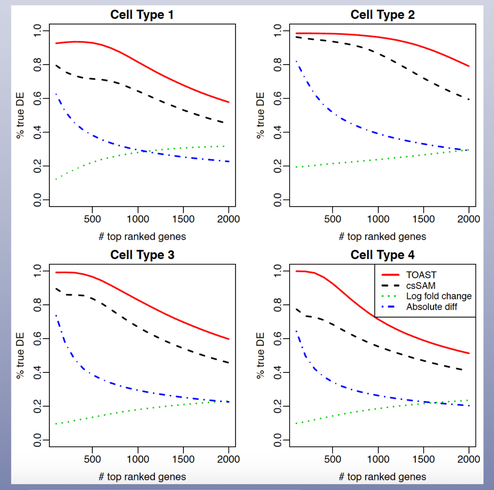
1. Testing the difference in cell type k **between two conditions** is H_0: $\beta_k = 0$;
2. Testing the difference between cell type p and q **in controls** is H_0: $\mu_p - \mu_q = 0$;
3. Testing the difference between cell type p and q **in cases** is H_0: $\mu_p + \beta_p - \mu_q - \beta_q = 0$;
4. Testing higher order changes, for example, the difference of the changes **between cell type p and q in two conditions**: H_0: $\beta_p - \beta_q = 0$.

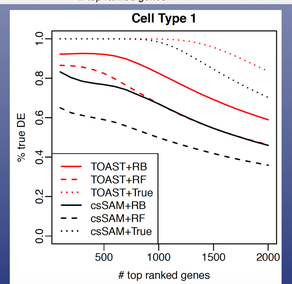
R package: TOAST (Tools for the Analysis of heterogeneouS Tissues)

SIMULATION STUDY

- A total of 100 simulation datasets are generated for each setting.
- Reference panel and measurement errors are simulated based on a true gene expression microarray dataset (GSE11058).
- Four cell types are simulated and 5% of genes are randomly selected to be differentially expressed genes.
- Proportions are drawn from Dirichlet distributions with parameters estimated based on a real proteomic dataset (Synapse.org with ID syn6098424).
- Reference-based algorithm, *lsfit*, and reference-free algorithm, *deconf*, are used to solve proportions.

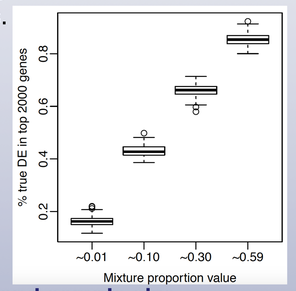
Comparison with existing methods



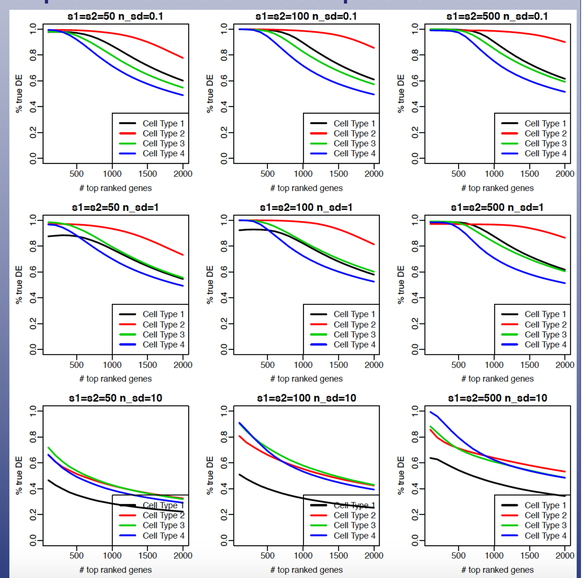


Impact of proportion estimation

Impact of proportion magnitude

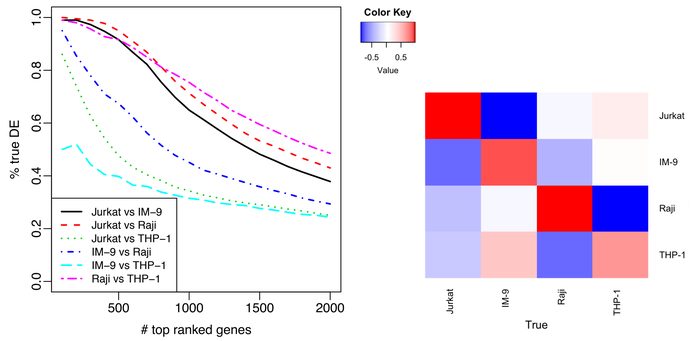


Impact of noise level and sample size



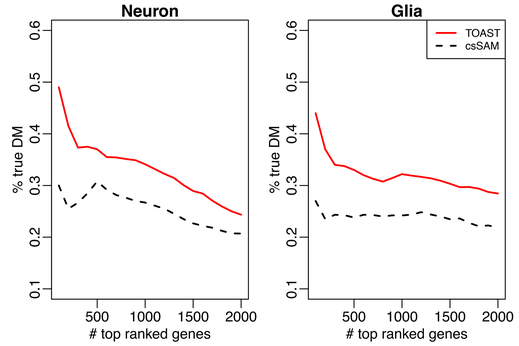
APPLICATION TO IMMUNE DATA

- GEO11058: gene expression microarray data of four immune cell lines (Jurkat, IM-9, Raji, THP-1) and their mixtures (four types of mixtures). Three replicates per cell line or mixture.
- **Goal of the analysis**: to detect DE genes for pair-wise comparisons of two different cell lines using the mixture data.
- **The "true" DE genes** are defined as the ones with the *limma* p-value smaller than 0.05 and the absolute log fold change greater or equal to 3.



APPLICATION TO HUMAN BRAIN METHYLATION DATA

- GSE41826: DNA methylation measurements for sorted neuron and glia from post mortem frontal cortex of 10 depression cases and 10 matched controls, and their unsorted, whole-tissue measurements.
- **Goal of the analysis**: to identify differentially methylated CpG (DMC) sites between depression and controls from DNA methylation data of whole tissue samples.
- **The "true" DMC sites** are defined as the *minfi* p-values smaller than 0.05 and the absolute methylation differences greater than 0.05.



References

1. Shen-Orr, Shai S., et al. "Cell type–specific gene expression differences in complex tissues." *Nature methods* 7.4 (2010): 287.
2. Abbas, Alexander R., et al. "Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus." *PloS one* 4.7 (2009): e6098.
3. Repsilber, Dirk, et al. "Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach." *BMC bioinformatics* 11.1 (2010): 27.

Software availability

TOAST package is freely available at https://github.com/ziyili20/TOAST.