Identification of cell-type-specific genetic regulation of gene expression for transcriptome-wide association studies

Duo Zhang*[1], Qiurui Ma*[2], Brandon Jew[3], Sriram Sankararaman[4]

[1]Shandong University; [2]Hong Kong University of Science and Technology; [3] Bioinformatics Interdepartmental Program, UCLA; [4]Department of Computer Science, Department of Human Genetics, UCLA

*** Equal contribution**

I. BACKGROUND

i. The Central Dogma

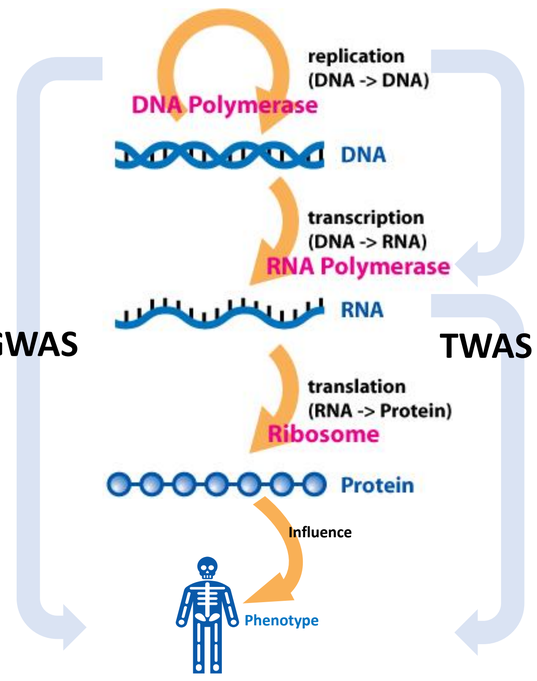
- Protein is generated through DNA transcription, RNA translation.
- Protein level dominantly affects phenotypes.
- Genotypes is a sequence of alleles along the **SNPs (X)**.
- Gene expression (**GE (Z)**) is the level of mRNA in one cell type. **Bulk level GE (G)** is the combined GE of all cell types in a tissue.

Encode DNA into SNPs data

Ind 1	Ind 2	Ind 3
AA	AG	AG
CT	CC	CC
⋮	⋮	⋮
AA	AG	AG

Geno1	Geno2	Geno3
0	1	1
1	1	2
⋮	⋮	⋮
0	1	1

The Central Dogma



ii. Current Studies

- Genome-wide association studies (**GWAS**)[3] linearly associate SNPs with phenotypes
- Transcriptome-wide association studies (**TWAS**)[2] linearly characterize the association with regulation of gene expression by SNPs.

iii. Challenges & Goals

Methodological

Unclear how SNPs affect phenotypes

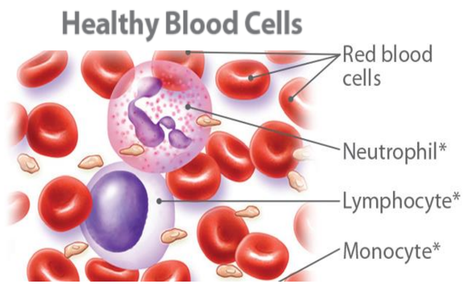
- Missing cell type information
- Fail to tell causality from correlation

Practical

Cell-type-specific biological data being resource intensive and expensive to acquire.

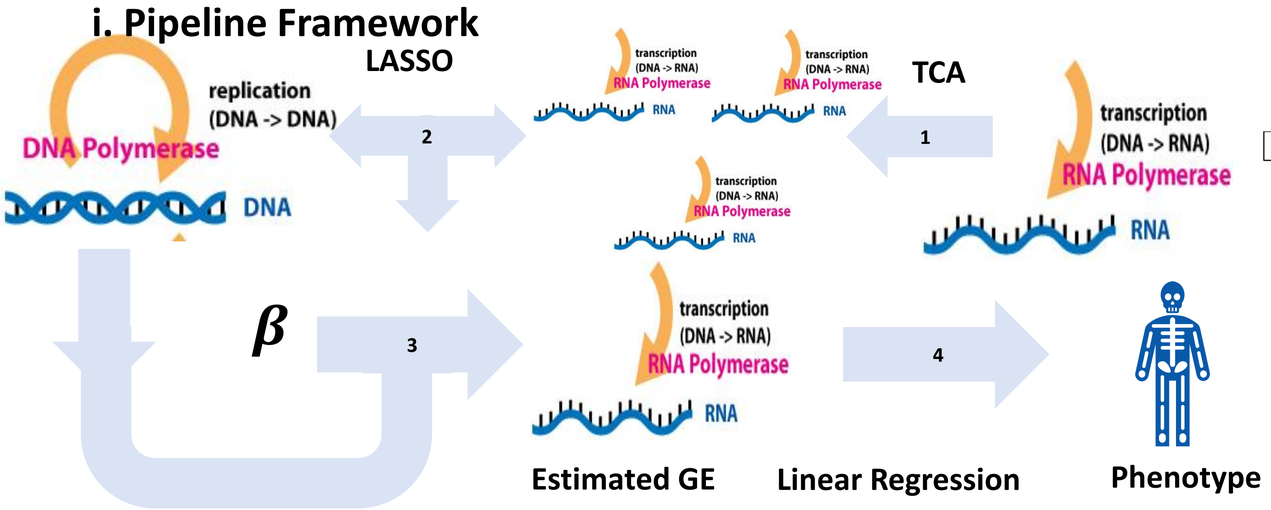
Our Goal

Deconvolute bulk level GE into cell-specific GE with SNPs and cell-type weights. Associate cell-type specific GE with phenotypes



II. METHODOLOGY

i. Pipeline Framework



1. TCA deconvolutes bulk level GE into cell-type-specific ones
2. Effect size of SNPs on cell-type-specific GE imputed by LASSO
3. Cell-type-specific gene expression imputed from effect size
4. Estimated GE is regressed into phenotype

Reference

[1] Rahmani, E., Schweiger, R., Rhead, B., Criswell, L. A., Barcellos, L. F., Eskin, E., ... & Halperin, E. (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *BioRxiv*, 437368.

[2] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., ... & Sullivan, P. F. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, *48*(3), 245.

[3] Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, *8*(12), e1002822. doi:10.1371/journal.pcbi.1002822

ii. TCA model

added SNPs effect: mdl1

TCA Model: mdl2

$$Z_h^i = \epsilon_z + \mu_h + \begin{bmatrix}1\\27\\0\end{bmatrix}^T \gamma_h + \begin{bmatrix}0\\2\\1\\\vdots\\0\\1\end{bmatrix}^T \beta_h \qquad \epsilon_z \in N(\mathbf{0}, \sigma_z)$$

Cell-Specific GE | Covariate: (Gender, Age, Smoking) | SNPs

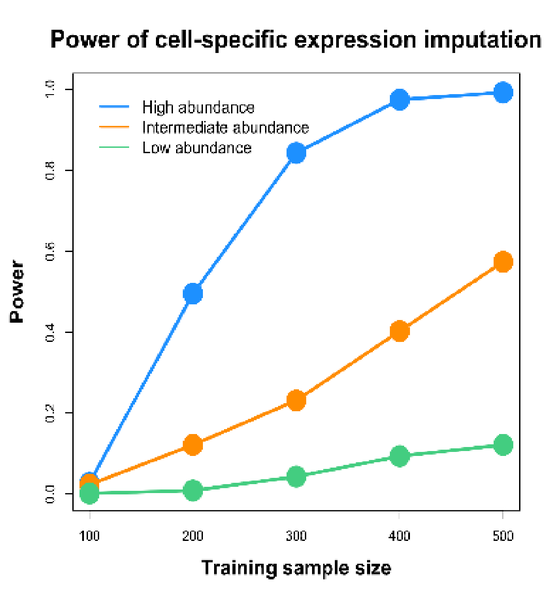
$$G_i = c_i^2\delta + \sum_{h=1}^{k} w_{hi} z_{hi} + \epsilon_g \qquad \epsilon_g \in N(\mathbf{0}, \sigma_g)$$

TCA assumes bulk level GE is a linear combination of GEs

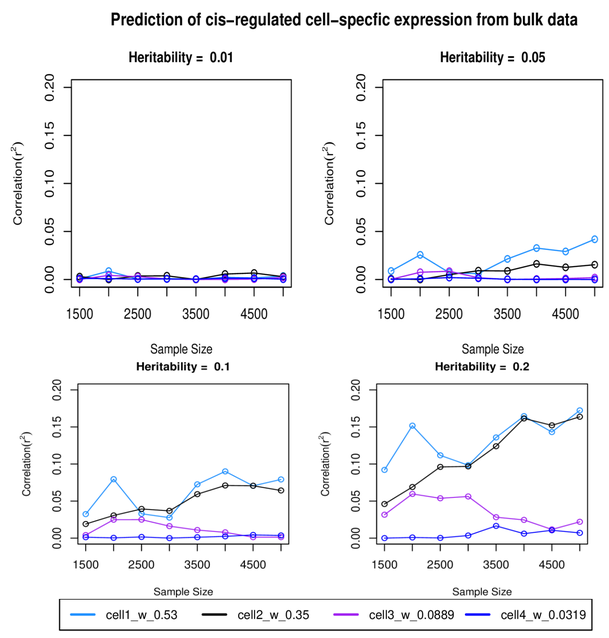
III. RESULT

i. Data Simulation

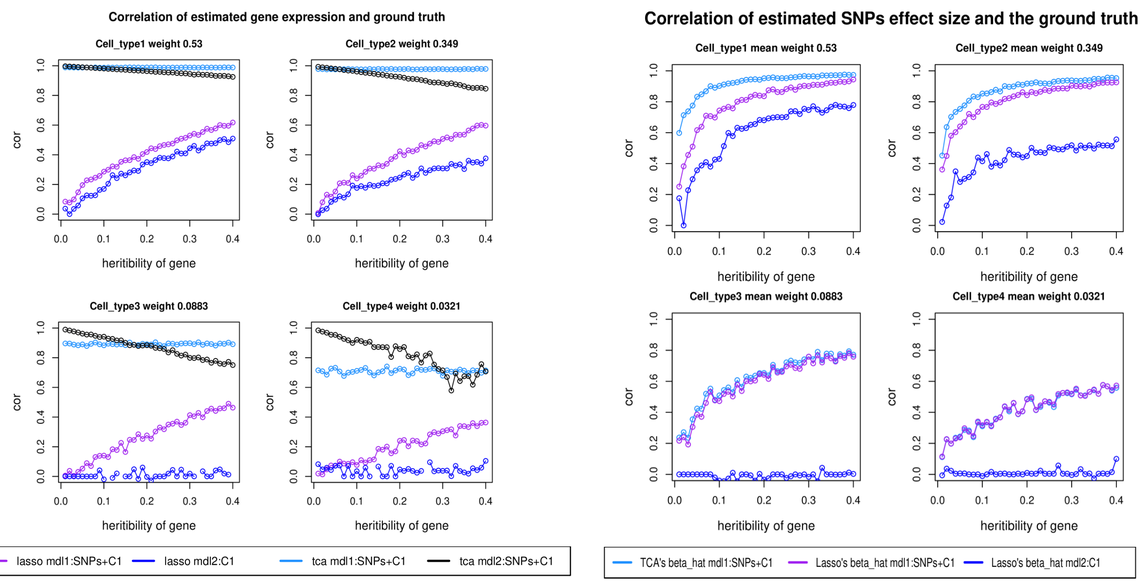
Is the prediction result of proposed model significant?



Is the model underfitting or overfitting?

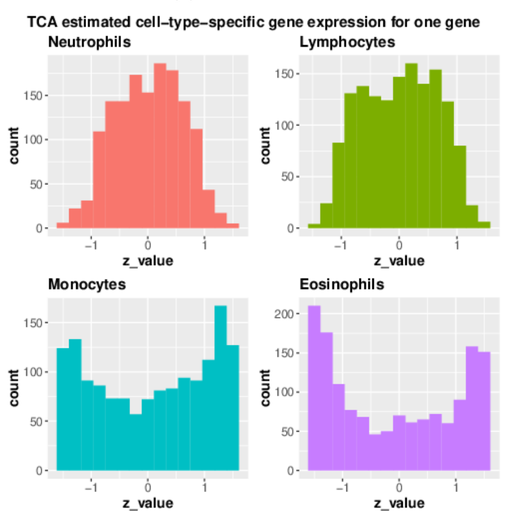


Compare our modified TCA and the original TCA, which is better?

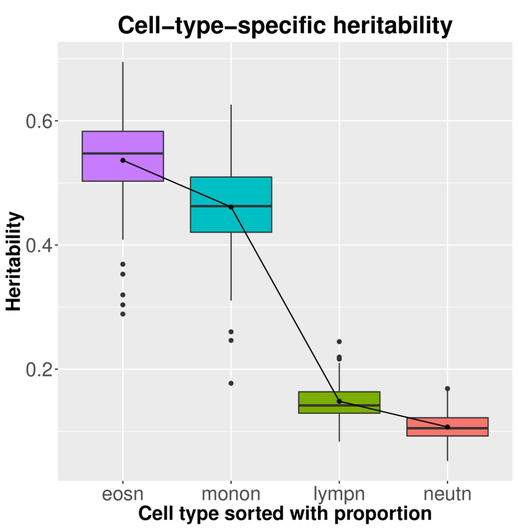


ii. Real Data

Is the model's performance consistent between cell types?



Is the model overfitting on real data?



Acknowledgement

- This work is supported by CSST program.
- The authors gratefully acknowledge the discussions with Elior Rahmani and Yue Wu.