Discussion outline

ZHAO Kai

Application of INSIDER to analyzing psychiatry disorders across different brain regions.

- individual level gene expression data across brain regions.
- may identify disease subtypes
- reveal unique or special biological processes to specific psychiatry disorders.
- uncover common biological processes to two or more psychiatric disorders.

Idea for future direction: reveal the contribution of cells (cell types) to the phenotypes of bulk RNA-seq samples

Assume Z^{blk}, Z^{sc} are $N_1 \times P, N_2 \times P$ matrices. The P features are common across the two datasets.

- Transfer learning for dimension reduction. We learn latent representations for bulk and single-cell samples with a share gene representation to ensure same metagenes in the decomposition. Thus, we reduce dimension of both bulk and single-cell data and ensume the same meaning of latent features of the two data. The idea can be formulated as

$$z_{ik}^{blk} = b_i g_k$$
$$z_{jk}^{sc} = s_j g_k,$$

  where b_i, s_j, g_k are latent representations of rank K for bulk sample i, single-cell j, and gene k, respectively.

- Cell cluster analysis. Find cell clusters with s_j. For illustration, assume m cell populations is discovered in the processes.

- Cell contribution decomposition. Use group lasso or sparse group lasso regression to find the relation between cells s_i and b_i, which can be formulate as

$$\text{argmin} \sum_{i \in N_p} \|b_i - S^T \beta_p\|_2^2 + (1-\lambda) \sum_i \sqrt{\#(N_{c_i})} \|\beta_p^{c_i}\| + \alpha\lambda\|\beta_p\|.$$

  Here N_p is a set of indices of bulk samples with phenotype p, S is a matrix of $N_2 \times K$, and β_p is a vector of length N_2. c_i is the i-th ($i \leq m$) cell population, discovered in the previous processes.

- Handle different types (e.g., binary, ordinal, continous) of phenotyple . The two former ones can be esaily handled. For continous phenotypes (e.g., survival time), we can discrete continous phenotypes by the quantile technique.