Discussion outline

ZHAO Kai

Transfer learning for joint analysis of bulk and single RNA sequencing data

Assume Z^{blk}, Z^{sc} are $N_1 \times P, N_2 \times P$ matrices, respectively. Here the P features are common across the two datasets. Assume there are m cell populations in Z^{sc}, which are marked by a number of marker genes.

Transfer learning for joint dimension reduction

We learn latent embeddings for bulk samples and cell types from single-cell samples with a share gene representation in the reduction. The idea comes from my recent analysis with the GBM dataset. In the analysis of the GBM dataset, we found that there is a wide difference in cell population across samples. This fact can also be applied to other cancer data. This practice can help directly characterize the cell populations. The idea can be formulated as

$$z_{ik}^{blk} \approx b_i^T g_k$$
$$z_{tk}^{sc} \approx s_t^T g_k,$$

where b_i, s_t, g_k are latent representations of rank K for bulk sample i, cell type t, and gene k, respectively. Notably, b_i only contains cell type related expression information and thus is less noisy. Moreover, introducing batch strategy to increases its scability. This practice facilitates several downstream analysis.

Uncover contribution of cell types to phenotypes of bulk samples

The contribution of cell types to phenotyes of bulk expression can be revealed in the following ways:

- Approach 1 : since b_i is restricted to cell type related information, the relationship b_i and s_t can evaluated by correlation. Thus, for $b_i, \forall i \in (1, \cdots, m)$, we compute its relationship with cell type representations, donoted with c_i, a vector of length m.
- Approach 2 : For b_i, we deconvolute its cell type fractions and use the fractions as covariates to ucover the contribution of cell types to phenotyes. The procedure of cell deconvolution is discussed below.

Gene expression deconvolution

Denote $\hat{Z}^{blk} = BG$, where B is a matrix with b_i as its i-th row, and G is a matrix with g_k as its j-th column. Similarly, $\hat{T}^{celltypes} = SG$. Then, for the i-th row $\hat{Z}_i^{blk}$ of $\hat{Z}^{blk}$, we have

$$\text{argmin} \|\hat{Z}_i^{blk} - \hat{T}^T \beta\|_2^2 + \alpha \|\beta\|.$$

Here β is a vector of length m.

Easy cell deconvolution

If we restrict the above process on a number of marker genes, then cell deconvolution can be easiy done.

Study relationship between cell population by cell types latent embeddings

Analyses of ajdusted expression profiles for cell types can be done,

For example, BPs for cel types, changes in expression levels of genes for cell types, etc.