Spatial Transcriptomics Deconvolution Methods

July 16, 2021

Spatial Transcriptomics

- Single cell RNA sequencing technologies (scRNA-seq) have allowed for low level cell-specific profiling of expression.
- However, this technique does not give any out information on the localized expression within a tissue sample - the expression rates are aggregates of each cell without spatial location context
- Spatial transcriptomics results in gene expression rates at each "spot" (a set sized group of cells) in a tissue sample, giving a spatial location on the tissue to each expression value
- One of the current challenges with spatial transcriptomics is its lack of specificity - each spot contains multiple cells of differing types in some proportion - which can can only be found through "deconvolution" of cell type proportions

MuSiC

- Bulk tissue RNA-seq data is similar to scRNA-seq data taken over various tissue states (ie progressions in a disease over time) and identifying cell type proportions within each state allows for exploration of its relationship with tissue state transitions
- MuSiC improves upon upon methods like CIBERSORT in that it does not require a pre-selection method to knock out insignificant genes and is sensitive to correlations/relationships between gene expression within a single cell across states

MuSiC

- Under some assumptions, the relative proportion of a particular gene across the tissues in a subject is proportion to a multiplicative combination of the proportion of that gene for a particular cell type, the molecular size (number of) that cell type, and the proportion of cells of that type - what we want to find
- The assumptions are: 1) the samples are all from the same population with the same distribution of gene proportions 2) the relative proportion of cell molecular sizes is the same across all states 3) the average read count is the same across varying cell types in all states
- The proportionality can be constrained since all cell type proportions must be non negative and the sum of proportions per each state must sum to 1

MuSiC

- Some similar cell types exist whose gene expression levels could be correlated, which needs to be accounted for
- Hierarchical clustering is used to split the possibly correlated cell types into trees of relative similarity, which are each split in proportions (since each branch is binary) by a non-negative least squares fit
- MuSiC can be fit to library sizes or UMI counts as a possible alternative to expression matrices in a different context of data

Stereoscope

- Method relies on the key assumption that both the spatial and single-cell data values follow a negative binomial distribution
- From the given data, the mean expression and rate parameter across both cell types gene types is possible to obtain
- For the cell type specific proportions, the values are obtained using maximum likelihood estimates by minimizing a negative log-likelihood loss function
- Similarly, proportions are found for the transcripts at each spatial location by summing all the transcripts from all the cell types at that particular spatial location
- This sum can be estimated as a fit of different cell type proportion distributions, which are permuted and fit using a derived posteriori fit relationship from the single cell proportions

SpatialDWLS

- This method attempts to fit a model using a modified least squares error minimization approach
- Certain genes are knocked out in pre-processing by omitting all that fail to meet a certain cutoff differential expression score
- The expressions posts are modelled as a function of the gene signature matrix, the sequencing data, and the cell type numbers - which is optimized in a least squares approach
- After removing cell types which were resolved at a low frequency across all spatial locations, the deconvolution via DWLS is run again
- The key advantage to the spatialDWLS method is that its repeated filtering of less relevant cell types increases its specificity (albeit at the cost of precision)

Integration

- Previous integration approaches have included a mutuals nearest neighbors method, which is not ideal for situations where only a small set of cell types are existent in the shared cell type lists
- Approaches using neural networks to 'scale' all the integrated data sets have been used before, however, they aren't designed to integrate genomic data in different forms
- The approach done in Seurat creates 'anchor cells' with a connection to other cells, allowing the distribution of those cell types to be based on the strength of the connections
- These neighbor graph connection weights allow for 'prediction scores' to be created for each cluster of cells, which are similar to likelihood estimates for the clusters

Integration

- The cell 'prediction scores' are created by first reducing the single-cell readings into a lower dimensional principal component space, then taking a transformation of the distance between each cell's representation and its k closest anchor cells
- To generalize this approach to spatial datasets successfully, the spatial data also needs to be properly labelled with different cell types occurring at different areas
- Unlike other deconvolution type methods, the prediction score approaches allows for viewing 'active' cell types in certain locations