Summary of computational methods

Kane Toh*

2022-03-10

10X scRNA-seq data pre-processing and analysis

10X data pre-processing

Raw Fastq files were downloaded from the 10X Genomics webpage. Paired-end reads were pseudoaligned onto a pre-built index and then quantified via `kallisto(v0.48.0)-bustools(v0.41.0)`Melsted et al.,2021.

Quality control

Droplets containing real cell transcriptomes were distinguished from those containing only ambient RNA using the`DropletUtils` function `emptyDrops` (default parameters). Cells were then imported into a `Seurat(V.4.0.6)`Hao et al., 2021 object for downstream analyses. To retain only high-quality cells, cells were retained only if they jointly pass the following four quality thresholds: 1) RNA count (nCount_RNA) $<$ 20000; 2) RNA count (nCount_RNA) $>$ 1000; 3) Number of unique features (nFeature_RNA) $>$ 1000; 4) Percentage of mitochondrial reads (percent.mt) $<$ 20.

Normalisation and dimensional reduction

Counts were log-normalised with the default parameters (`NormalizeData`) and the top 2000 highly variable genes were retained for analysis (`FindVariableFeatures`). The normalised count matrix was then scaled and centered (`ScaleData`). The top 10 principal components (PCs) were retained for analysis after examination of the corresponding elbow plot. Next, cells were clustered using the Louvain algorithm by first conostructing a shared nearest neighbor (SNN) graph (`FindNeighbors`) and then determining the number of clusters (`FindClusters`) with a resolution parameter of 0.5. A total of 8 clusters were obtained and visualised with a Uniform Manifold Approximation and Projection (UMAP) embedding, alongside the expression of several features of interest.

Identification of Differentially Expressed Genes

Cluster-defining transcripts were identified for all 8 clusters via the default Wilcoxon Rank Sum test (`FindAllMarkers`). P-values were adjusted for multiple hypothesis in `Seurat` using a Bonferroni correction.

Automated Annotation of Cell States

Automated, unbiased cell type recognition was carried out using the `SingleR` package, leveraging on the `MonacoImmuneData` reference index downloaded from the `celldex` R package. The reference dataset contains curated cell type labels from 114 bulk RNA-seq samples of sorted immune cell populations.

*Genomics and Data Analytics Core (GeDaC), Cancer Science Institute of Singapore, National University of Singapore; kanetoh@nus.edu.sg

References

1. kallisto|bustools workflow: Melsted et al.,2021: Pseudoalignment and transcript quantification
2. kallisto: Bray et al., 2016
3. Seurat 4.0: General single-cell RNA sequencing processing
4. singleR: Automated annotation of cell clusters
5. CellDex