Updated on 01/23/2019

RNA-seq (Single Cell) data analysis

Multiple techniques are available to generate Single Cell RNA-seq (scRNA-seq) data that measures the genome-wide expression profile of individual cells. This document aims to provide a workflow for analysis of 10x Genomics® Chromium™ scRNA-seq data. 10x Genomics protocols are droplet-based; supports the sequencing of 3' or 5' end of a transcript molecule. To label a transcript it uses an UMI (Unique Molecular Identifier) before amplification. Cell Ranger Single Cell Software Suite 3.0.2 are a set of pipelines to process 10x Genomics RNA-seq data.

First, we demultiplex raw base call (BCL) data into FASTQ files and a count matrix where each row corresponds to a gene and each column corresponds to a cell. Based on library size (total reads) and the number of expressed features (ex. genes), we identify and remove low quality or empty cells. Second, using filtered count data, we perform dimensionality reduction and clustering to identify new subpopulations and, differential expression analysis to detect highly variable genes. The R software package Seurat will be used for all downstream analyses. The count data will be filtered and log-transformed. A principal component analysis (PCA) of the most variable genes will be performed and an elbow plot will be used to select the principal components (PCs) capturing the most variance in the dataset. These PCs will be used as edge weights in an unsupervised graph-based clustering to identify cell clusters. T-distributed stochastic neighbor embedding (tSNE) will be used for visualization of the cell clusters. Expression levels of cell-type specific markers, if available, will be used to determine the putative identities of each cell cluster. Once the cell clusters are identified, differential expression will be done using one of the four tests implemented in Seurat. We define differentially expressed genes as those with an adjusted p-value <= 0.05, an average log2 fold change >= 1.

Note: For sequencing data acquisition please contact Emory Integrated Genomics Core (EIGC@emory.edu). EIGC offers three scRNA-seq technologies such as the Fluidigm C1, 1CellBio inDropand and 10x Genomics. 10x Genomics works by parallel processing of up to 8 samples at once – it easily fits an experiment with 2 conditions and each with 3 replicates. Input is a single cell suspension between 500-1500 cells/ul (min 50ul). While 10x is capable of processing up to 8 samples in parallel, it may still be necessary to run the experiment over multiple days to minimize cell exposure to suboptimal conditions (depending on how cumbersome it is to prepare single cell suspensions at one time). Good quality cell suspensions are critical. Debris and clumps should be absent. Viability should be high (>95%). Poor quality cell preps can compromise encapsulation (clumps/debris) and also lead to noisy sequence data (low viability). Resulting libraries will be sequenced using in-house NextSeq (up to 400M reads/run), or outsourced, if greater coverage is required.

Required for data analysis:

1. Raw data files (BCL or fastq)

2. A file that gives the mapping between sample name and sample index for library construction.
3. Organism details (ex. Human)

Deliverables of Single Cell RNA-seq data analysis service:

1. tSNE plots
2. Differential expression analysis with multiple hypothesis testing
3. Potential markers for a subset of cells
4. Enrichment analysis – pathways or Gene Ontology
5. Details of analysis workflow for your writing (manuscript)

References:

1. https://www.10xgenomics.com
2. https://satijalab.org/seurat/

Questions? Comments? If you have additional requirements or questions, please feel free to contact us at EICC@emory.edu