Guannan Shen

Single-cell RNA-seq (scRNA-seq) experiments can identify rare cell populations that have distinct gene expression profiles. To get hands-on experience on scRNA-seq data analysis, I will understand and follow the methods in the “Recovery and analysis of transcriptome subsets from pooled single-cell RNA-seq libraries” paper. I will also use the dataset published along with this paper, GEO Series GSE119428. This dataset contains 13 samples. Unlike usual scRNA-seq approach, they developed a strategy to physically recover (“transcriptome resampling”) the DNA molecules comprising transcriptome subsets. As a result, this method yielded either greater sequencing depth per cell or higher coverage of targeted gene expression information. In this paper, they used “snakemake” pipeline in Python 3.5 to process scRNA-seq fastqs to generate count matrices, assembling TCR seqs, and producing UMI summary flat files. To generate results, CRAN, Bioconductor and Github packages were used, such as Seurat, rjson, ggrepel, doParallel, GenomicFeatures, GenomicRanges, GenomicAlignments, Gviz, rtracklayer and ComplexHeatmap. They developed their own package “kentr” (devtools::install\_github("kriemo/kentr")), which can extract sequence from fasta, compute Hamming distances, perform Smith-Waterman alignment etc. I will also understand and perform cell barcode analysis. Ideally, my analysis will prove that the resampled library can enrich and isolate certain Unique molecular identifiers (UMIs) compared with the original library. In addition, the pipeline and downstream analysis methods I learnt here can be applied to test other hypotheses on scRNA-seq data.