Reading Part 2: Single Cell RNA-Seq

- Quantifies transcriptome expression for a single cell, one RNA-seq run can be used to examine expression levels of a human cell for all 20,000+ genes.
- High Throughput scRNA-seq: method can be applied to multiple cells so there is no information bottleneck, but need a way to distinguish individual cells = “Barcode” sequence.
- Drop-Seq: popular high-throughput scRNA-seq method
    - Droplet formed from water in oil environment that contains the same barcode as an individual cell
    - That cell is contained within the droplet, and then ruptures, keeping mRNA’s within the droplet by binding to poly-A arms on the bead.
    - Bead contains an unique molecular identifier (UMI) [was 8bp, now modern uses have 16bp]
    - Typical scRNA-seq with paired-end seq. using Illumina:
        - Read 1 contains cell barcode and UMI
        - Read 2 reads the cDNA
    - Reads then sorted based on cell barcode and then read 2’s are mapped and the # of UMI’s per gene is calculated as the count matrix (in an .mtx format)
- Bioinformatics: after importing the count matrix
    - QC used to remove data noise(using kneeplots) and multilets
    - Normalize the # of UMI’s by dividing by cell number, x10,000, and then natural log/log2

Reading Part 3: Machine Learning and Clustering

- Machine Learning: a way in which a program can improve performance with data without the prior static instructions from a human programmer, being more complex than a simple workflow redirection as well.
- Categorize by task
    - Classification
    - Regression
    - Clustering
    - Dimensionality Reduction
- Categorize by human supervision
    - Supervised
    - Unsupervised
    - Semisupervised
    - Reinforced
- t-SNE: t-distributed stochastic neighbor imbedding, acts as a dim.red. technique, can be used in R.
- UMAP: Uniform Manifold Approximation and Projection
    - Cluster visualization for scRNA-seq data points. Creates fuzzy topological representation of real data
- PCA: principal component analysis.