Summer Institute in Statistical Genetics 2019

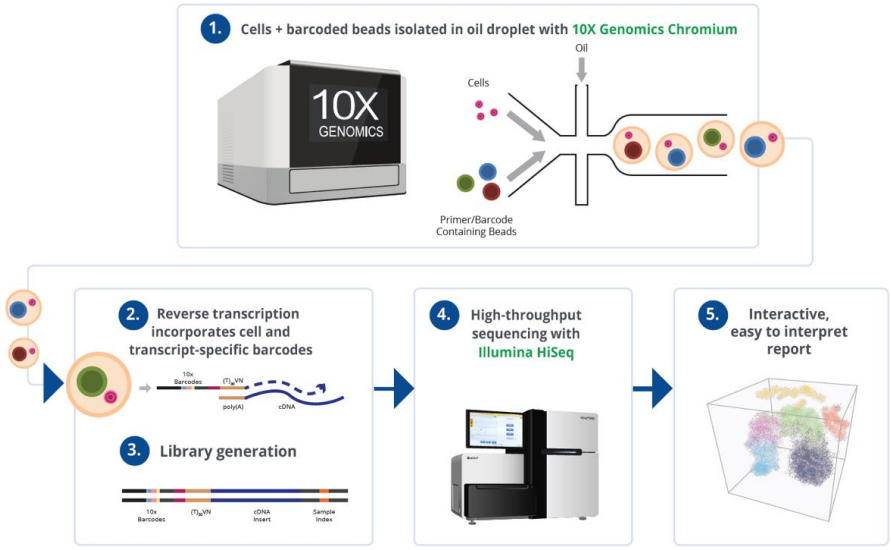
3. Single Cell RNAseq Basics

ggibson.gt@gmail.com
http://www.cig.gatech.edu

Georgia Tech

Center for Integrative Genomics

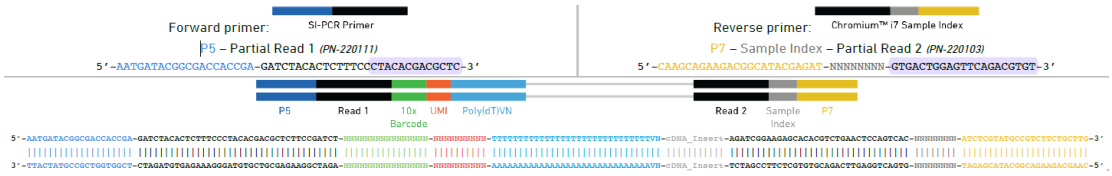
Single Cell RNA-seq: Easy as 1,2, 3, … 5



Types of Single Cell RNA-seq

1. SmartSeq2
   - Essentially full-length RNA-seq applied to libraries generated from single cells
   - Low throughput and relatively expensive, but comprehensive
   - Commercial option is Becton-Dickinson Rhapsody™
2. Droplet Sequencing
   - Each cell is encapsulated in a droplet with enzymes and reagents for sequencing
   - High throughput, dollars per cell, but only detects tags for each transcript
   - Commercial options are 10X Genomics Chromium™, BioRad, and OneCellBio
3. sci-Seq
   - Single cell Combinatorial Indexing in microtiter plates
   - High throughput, very inexpensive, amenable to dual profiling with other assays
   - Implemented in academic labs

Chromium Droplet Barcodes



Sample Index is a barcode specific for the sample (individual, tissue, treatment, etc)

10X Barcode is for the cell, it tags all molecules derived from the same cell

UMI is a Unique Molecular Identifier for each actual mRNA molecule, basically controls amplification biases

Since library costs start at $1300, multiple samples can be combined in one reaction by adding a 4th type of barcode such as a BioLegend cell surface antibody, or using the person's genotypes

In a typical cell: 50,000 reads may correspond to 10,000 UMI and 3,000 expressed genes
most transcripts may have from 1 to 5 UMI each represented by multiple reads

Read Depth, Cell number, and Expense

Sequencing is done on either a NextSeq or NovaSeq Illumina sequencer. Typical current options might be:

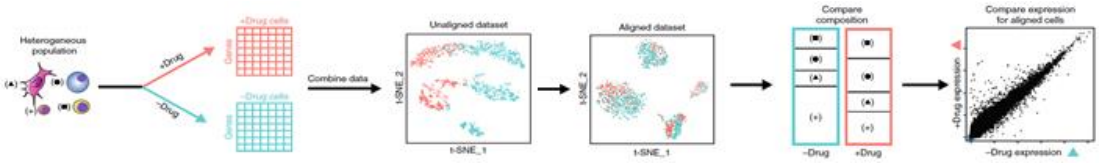
NetSeq lane = 400 Million 28x96 bp = 50,000 reads per cell for 8,000 cells, at a cost of ~$3,000 {40 c/cell}
S1 flow cell = 3 Billion 28x96 bp reads = 100,000 reads per cell for 30,000 cells, at a cost of ~$10,000 {30 c/cell}
S4 flow cell = 18 Billion PE reads = 50,000 reads per cell for 360,000 cells, at a cost of ~ $30,000 { 8 c/cell}

What read depth is required?

It depends on the cell-type: 50K is sufficient for many, but some require 110K

It depends on the application: if low abundance transcripts are key, you need more
if differential expression is key, you may need more
if defining novel cell types and states is key, you may need more

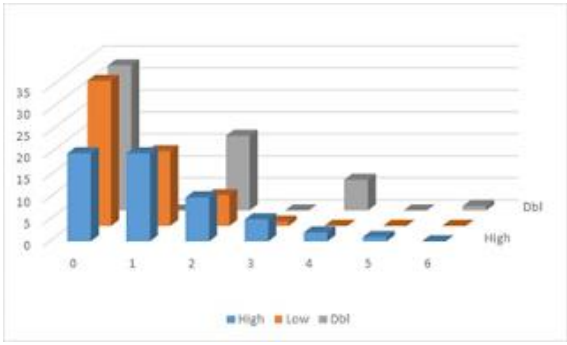
Typical Input and Output



Adjust for Batch Effects while preserving the Cell cluster identities

Compare the effects of the treatment on
a) The relative abundance of different types of cell
b) Gene expression within cell-types

Scaling to counts per 100,000 reads



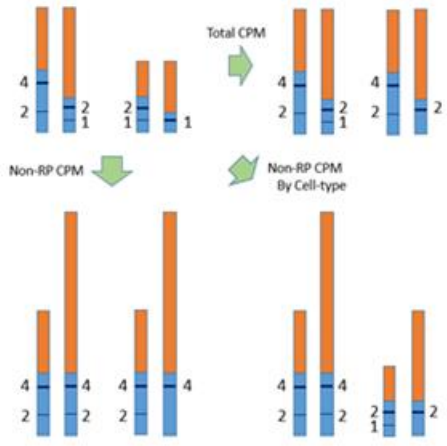
Scaling zero by any number still results in zero.

Here, the orange counts were derived by down-sampling the blue counts to half. The proportion of drop-outs increases.

Just multiplying all counts by 2 to scale back up does not restore the initial distribution – in fact, it results in a very skewed profile with no odd-numbered counts.

Some authors suggest getting around this by **imputation** to fill in the missing data. We think that is a dangerous strategy in general.

The normalization problem



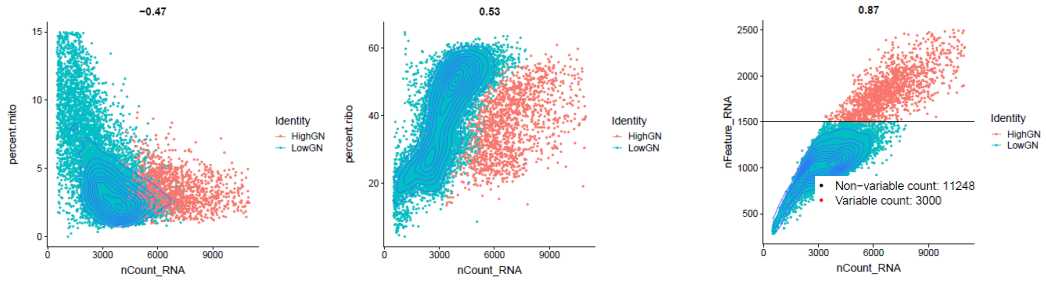
Suppose we represent scRNA abundances for two cells of each of two cell types by these bars, with ribosomal proteins in orange and common transcripts in blue. Now focus on two genes represented by the horizontal bars, with counts shown next to them.

Normalizing by total cpm leads to the conclusion that there is little difference between the left and right cell types, except for the drop-out transcript, but there is high within-sample variability.

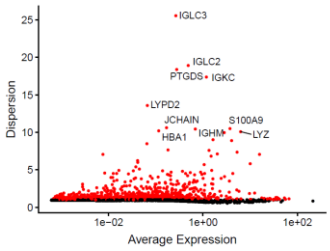
Normalizing by non-ribosomal CPM alone leads to the conclusion that all four cells are very similar.

Normalizing by cell-type and non-ribosomal CPM recovers the cell-type difference in absolute abundance.

Quality Control

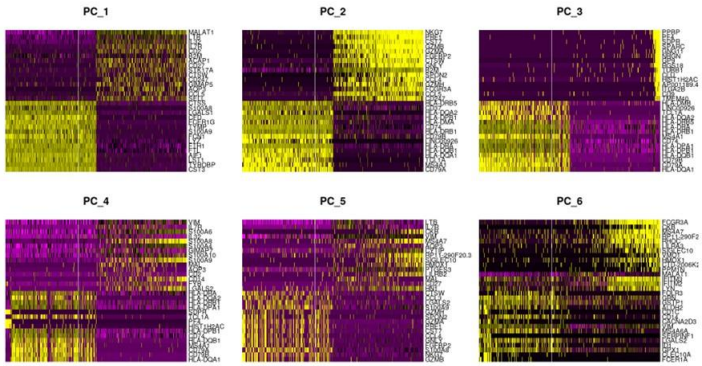


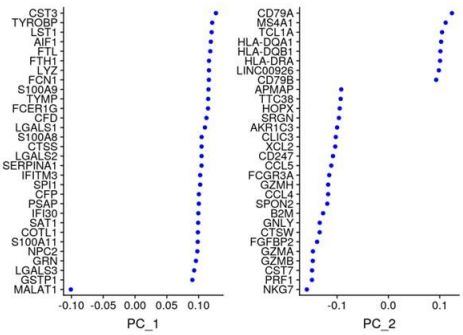
Detecting Highly Variable Genes

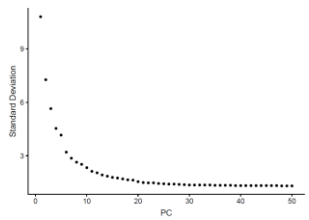


Deciding how many PC to include in CCA

https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html

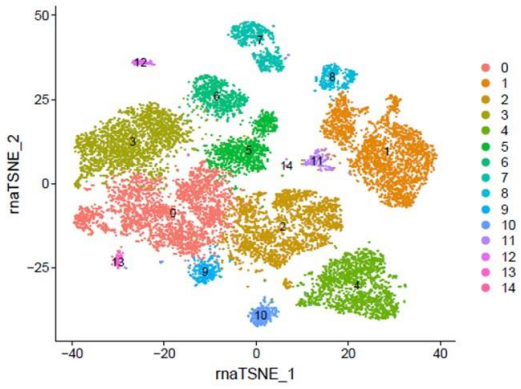




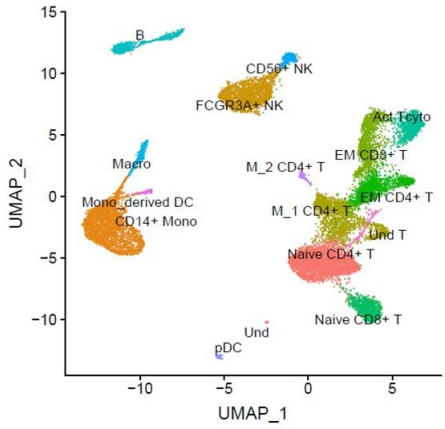


Cluster Visualization

tSNE

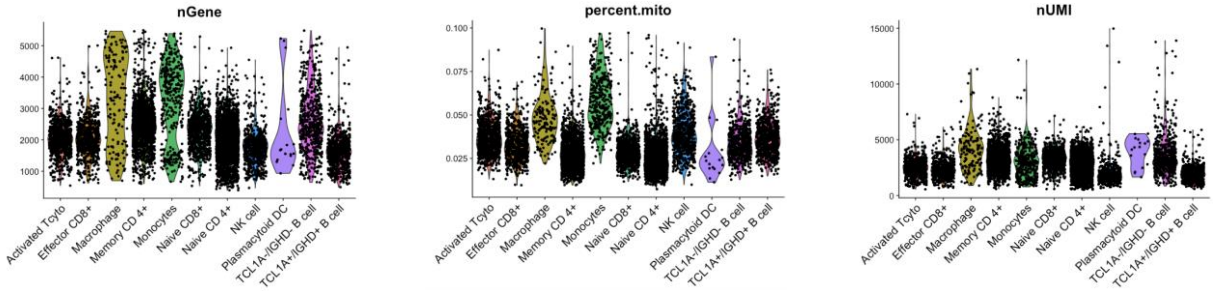


UMAP

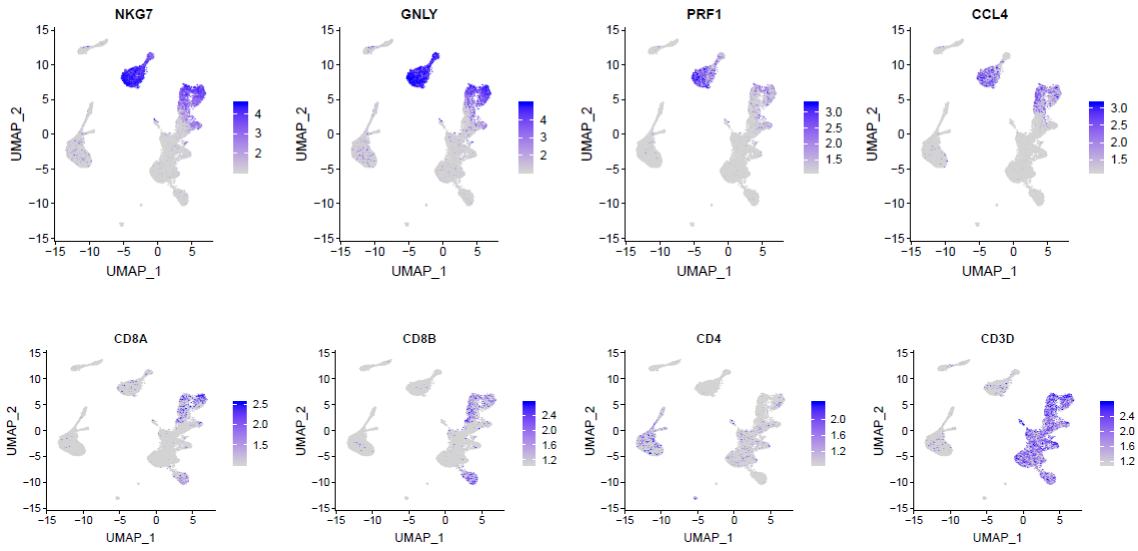


https://distill.pub/2016/misread-tsne/

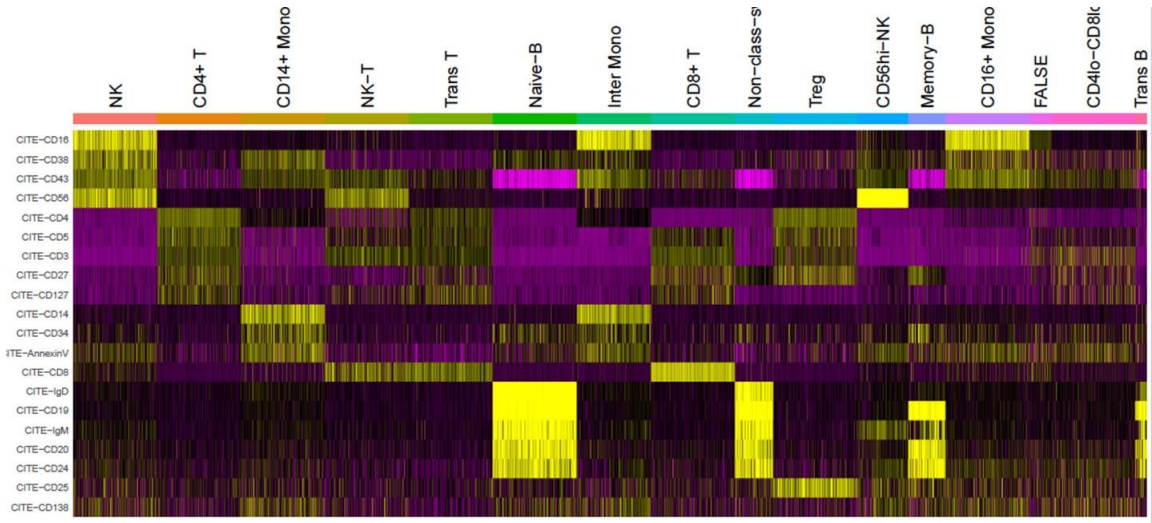
Properties of Immune Cell Types



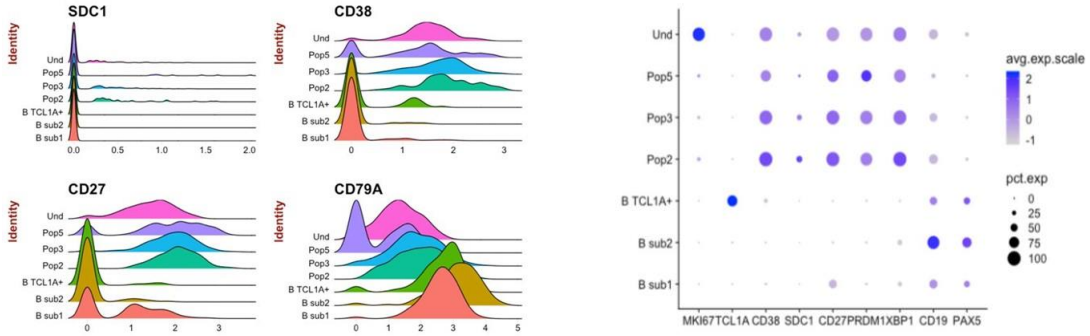
Anchoring Cell-types by Marker genes



Hierarchical Clustering of Marker Genes



Differential Expression Analysis



Some extensions of Single Cell Genomics

1. Crop-Seq / Perturb-Seq
   - Microdeletion of SNPs in single cells followed by RNA-Seq
   - Requires co-transfection with Cas9 and lentivirus or plasmid expressing guide RNAs
   - Generally useful to monitor alterations of gene networks
2. CITE-Seq
   - Addition of oligonucleotide-conjugated Antibodies that bind cell surface receptors
   - Receive the same cell barcodes as the cell contents, but sequenced separately
   - Supports gating to homologize flow cytometry with scRNAseq
3. ATAC-Seq
   - Assay for Transposon-Accessible Chromatin (basically, identifying enhancers)
   - 10X Genomics now provides kits; reports of joint scRNA and scATAC appearing
   - https://www.10xgenomics.com/solutions/single-cell-atac/
4. Repertoire-Seq
   - Sequencing of the TCR (T-cell receptors) or BCR (immunoglobulins) from single cells
   - Options available from 10X and OneCellBio
   - Data analysis requires specific expertise