Single Cell Sequencing Analysis

Part 1

Single Cell Sequencing Data Review

- Sequencing depth = (# of cells) x (required depth):
  - RNA - *50k paired end reads / cell* for cell type classification
  - RNA - *.25M-1M paired reads / cell* for transcriptome coverage
  - DNA - 30-100x per cell
- e.g. 1000 cell scRNA-Seq = 250M-1B reads **per sample!**
  - Bulk mRNA-Seq: 30M-80M per sample
- Sequences in one PE fastq file are entirely barcodes
- Read length > 50bp for annotated genome

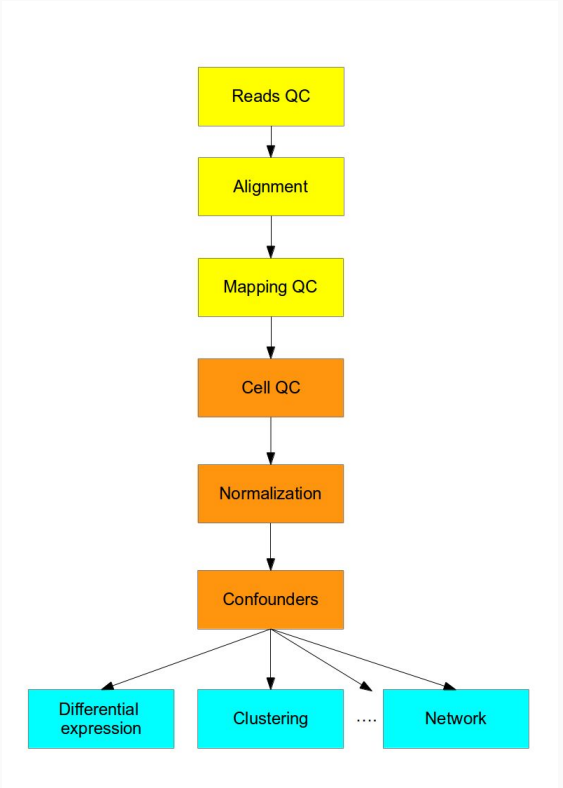
Rizzetto, et al. 2017. “Impact of Sequencing Depth and Read Length on Single Cell RNA Sequencing Data of T Cells.” *Scientific Reports* 7 (1): 12781.

The Trees: Cells

- What cell types are in a sample?
- What are their relative proportions?
- How do their transcriptomes differ?
- Which/how do cells respond to stimulus?
- How do cells change over time?
- What is the level of mosaicism in tissues?

Analysis Overview: scRNA-Seq

1. Sequence QC
   a. Demultiplex
   b. UMI Collapsing
2. Alignment+QC
3. Quantification
4. Normalization
5. DE, Clustering, etc



https://hemberg-lab.github.io/scRNA.seq.course/introduction-to-single-cell-rna-seq.html

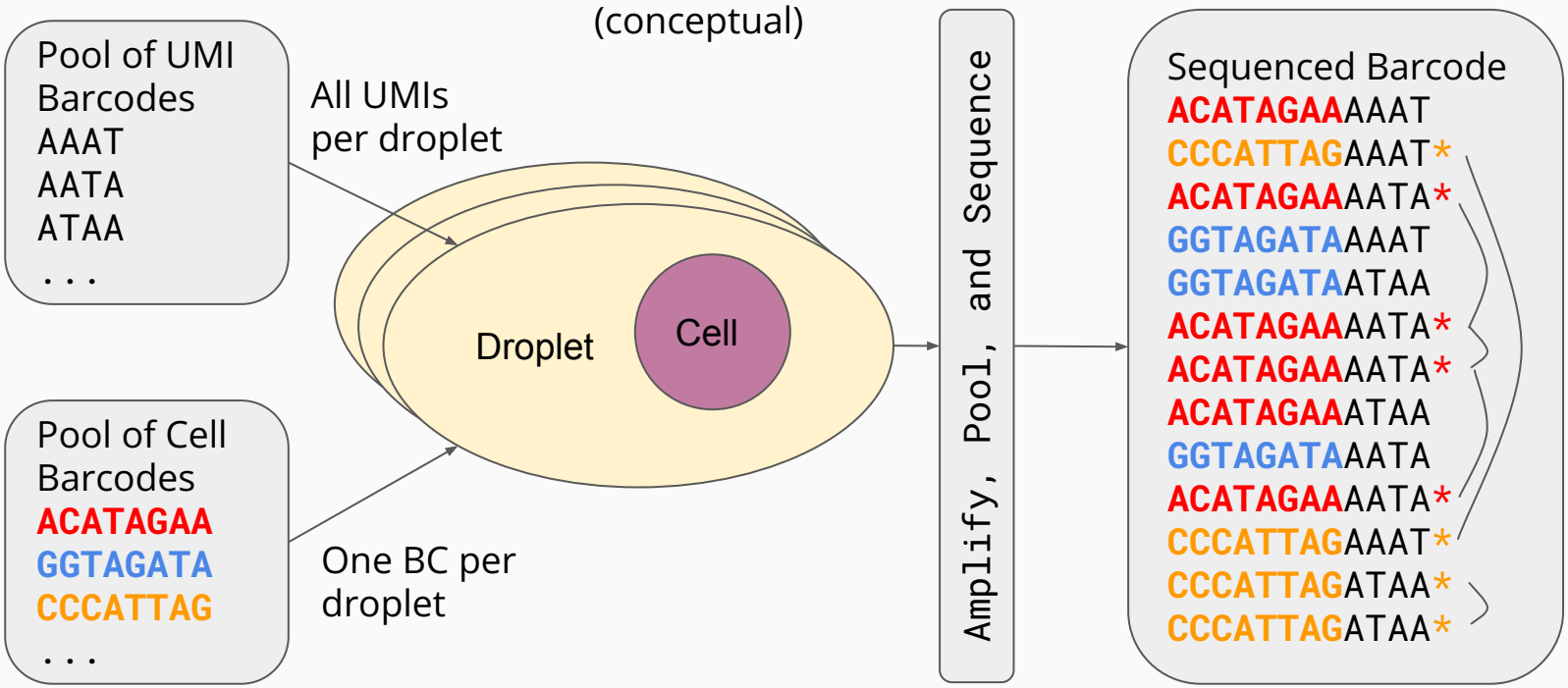
Sequence QC

- One sample is 100s-10,000s of cells
  - i.e. ~1,000 fastq files *per sample*
  - May or may not be already demultiplexed by core
- Paired end:
  - Read 1: molecule sequence
  - Read 2: barcode - used for demultiplexing and UMI collapsing
- Normal fastq processing and QC:
  - Adapter and quality trimming of both reads (barcode read can still have adapter)
  - fastqc, multiqc

Unique Molecular Identifiers

- Enable detection of qPCR amp. artefacts
- Not required, but often used
- Reads *deduplicated* or *collapsed* by cell barcode+UMI sequence prior to analysis
- Barcodes/UMIs designed to tolerate sequencing errors
  - i.e. >2 edit distance between any two sequences

Unique Molecular Identifiers



Red cell: 6 reads,
3 original fragments

Blue cell: 3 reads,
3 original fragments

Orange cell: 4 reads,
2 original fragments

Alignment

- Use either UMI collapsed or original reads
- UMI-tools: toolkit for working with UMIs
- Standard tools and QC, i.e.:
  - Alignment: STAR, bwa, bowtie, etc
  - QC: RSeQC, multiqc, etc
- **NB:** Some aligners have single cell mode
  - e.g. STARsolo - STAR aligner scRNA-Seq mode

QC: Mitochondria and spike-in controls

- High % reads mapping to mitochondrial genes = indicates low sample quality
- Spike-in (synthetic) RNA is sometimes used as an alternative control
- Idea: if mito/spike-in reads make up high proportion of reads, mRNA concentration was low

Quantification

- STAR+htseq-count, kallisto, salmon, etc
- Each sample has a different # of cells
- Each cell has the same number of measurements (e.g. genes)
  - = (# of samples) x (# of cells) x (# of genes)
  - Sparse: most will be zero!
- We consider only single sample case below

Count Matrix Normalization

- Normalization needed to make counts comparable between cells
- Two possible levels of normalization:
  - Within cell (e.g. divide by column sum, “library size”)
  - Within dataset (e.g. divide by total number of reads)
- All methods from bulk apply, i.e.
  - CPM, FPKM, DESeq2 etc...

The Counts Matrix

- Counts matrix contains either:
  - Read counts or
  - UMI counts if used
- Each cell has:
  - Total number of counts (col. sum, “library size”)
  - Number of non-zero genes
- Each gene has:
  - # of non-zero cells
  - Non-zero mean/variance
- Matrix is *sparse*: many zeros
- Zeros may be:
  - Cell lacks gene
  - A “drop-out”: gene present but was missed by qPCR

	cell1	cell2	cell3	cell4	cell5	cell6	…	cellM
gene1	93	25	0	0	3335	0		82
gene2	5	2	0	3	1252	0		12
gene3	0	0	0	0	0	0		0
gene4	98	21	1	1	5318	0		75
gene5	0	0	513	0	0	325		135
gene6	0	0	113	0	1	497		255
gene7	3	0	0	0	6	0		0
…								
geneN	68	52	0	2	4313			63

Examining the Counts Matrix

Each cell type has a signature, i.e. a pattern of gene expression

Consistent pattern of expression suggests same cell type:

- Cells 1, 2, and 5 (M?)
- Cells 3, 6 (M?)

	cell1	cell2	cell3	cell4	cell5	cell6	...	cellM
gene1	**93**	**25**	**0**	**0**	**3335**	**0**		**82**
gene2	**5**	**2**	**0**	**3**	**1252**	**0**		**12**
gene3	0	0	0	0	0	0		0
gene4	**98**	**21**	**1**	**1**	**5318**	**0**		**75**
gene5	**0**	**0**	**513**	**0**	**0**	**325**		**135**
gene6	**0**	**0**	**113**	**0**	**1**	**497**		**255**
gene7	3	0	0	0	6	0		0
...								
geneN	**68**	**52**	**0**	**2**	**4313**			**63**

Filtering Cells and Genes

- Many measurements
  - e.g. 30k genes x 1ks of cells
- Some cells are uninformative, e.g.:
  - Very few reads, few genes detected
  - Two cells sequenced together (i.e. doublets)
- Some genes are uninformative:
  - Low # reads, low variance across all cells
  - Too few cells express gene (e.g. < 10 of 10,000 cells nonzero)
- Must filter genes *and* cells to reduce noise

Filtering the Counts Matrix

Genes likely not expressed and should be **filtered**:

- Very few non-zero counts AND
- Low non-zero count mean

NB: Genes with few non-zero counts and HIGH non-zero count mean suggests rare cell type!

	cell1	cell2	cell3	cell4	cell5	cell6	...	cellM
gene1	93	25	0	0	3335	0		82
gene2	5	2	0	3	1252	0		12
gene3	**0**	**0**	**0**	**0**	**0**	**0**		**0**
gene4	98	21	1	1	5318	0		75
gene5	0	0	513	0	0	325		135
gene6	0	0	113	0	1	497		255
gene7	**3**	**0**	**0**	**0**	**6**	**0**		**0**
...								
geneN	68	52	0	2	4313			63

Filtering the Counts Matrix

Cells might also be filtered:

- **Very few or zero counts (cell4)**
- *Very many counts (cell5)*
  - Possible “doublet” of same cell type
- Inconsistent expression pattern (cellM)
  - Possible “doublet” of different cell types

Doublet: two cells with same cell barcode

	cell1	cell2	cell3	**cell4**	*cell5*	cell6	...	cellM
gene1	93	25	0	**0**	*3335*	0		82
gene2	5	2	0	**3**	*1252*	0		12
gene3	0	0	0	**0**	*0*	0		0
gene4	98	21	1	**1**	*5318*	0		75
gene5	0	0	513	**0**	*0*	325		135
gene6	0	0	113	**0**	*1*	497		255
gene7	3	0	0	**0**	*6*	0		0
...								
geneN	68	52	0	**2**	*4313*			63

Filtering the Counts Matrix: Quality

- Filtering thresholds are subjective!
- Must consider protocol, biological system, and study design
- Examples:
  - remove cells with median sum count < 3 median absolute deviations from median
  - Remove genes with more than 90% zeros AND non-zero mean < 10

Filtering the Counts Matrix: Variance

- Some genes are shared by all cells
- Normalization assumes most genes are not differentially expressed
- Genes with low variance across cells are uninformative
- Filtering threshold is subjective!

Typical Analysis Paths