Frederick National Laboratory for Cancer Research

scRNA-seq preprocessing and quality control

Nathan Wong (CCBR) and Vicky Chen (CCR-SF Bioinformatics Group)

Cell Ranger

Frederick National Laboratory for Cancer Research

- **Used to process FASTQ files for 10X samples**
- **Generates UMI expression matrices, basic sample statistics, and interactive analysis platform**

Summary | Analysis

3,974
Estimated Number of Cells

25,445
Mean Reads per Cell

2,256
Median Genes per Cell

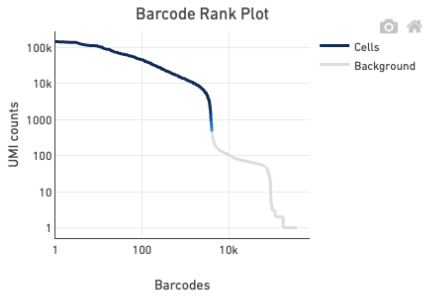
Sequencing

Number of Reads	101,117,641
Valid Barcodes	96.2%
Valid UMIs	100.0%
Sequencing Saturation	22.4%
Q30 Bases in Barcode	96.3%
Q30 Bases in RNA Read	82.7%
Q30 Bases in Sample Index	93.7%
Q30 Bases in UMI	95.7%

Mapping

Reads Mapped to Genome	91.1%
Reads Mapped Confidently to Genome	86.3%
Reads Mapped Confidently to Intergenic Regions	2.5%
Reads Mapped Confidently to Intronic Regions	10.8%
Reads Mapped Confidently to Exonic Regions	73.0%
Reads Mapped Confidently to Transcriptome	69.4%
Reads Mapped Antisense to Gene	1.0%

Cells



Estimated Number of Cells	3,974
Fraction Reads in Cells	88.4%
Mean Reads per Cell	25,445
Median Genes per Cell	2,256
Total Genes Detected	21,295
Median UMI Counts per Cell	8,614

Sample

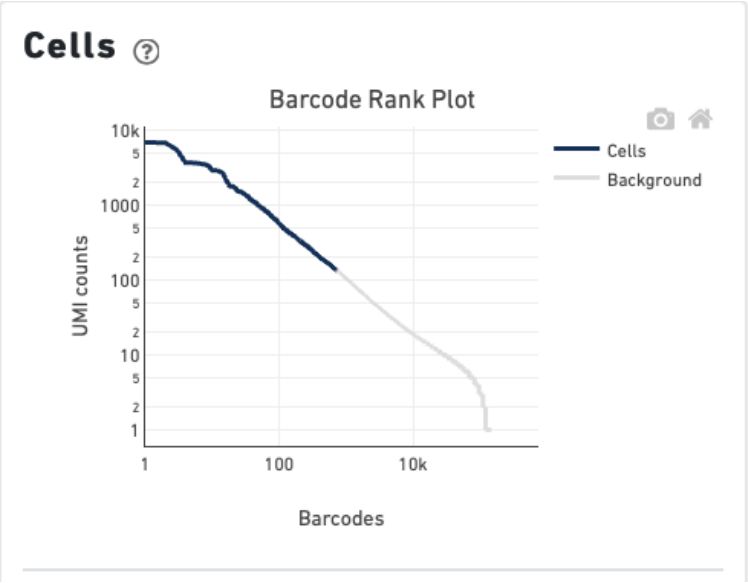
Sample ID	Fixed_M_HCC1395BL-95_HCC1395-5
Sample Description	
Chemistry	Single Cell 3' v2
Transcriptome	GRCh38-3.0.0
Pipeline Version	3.1.0

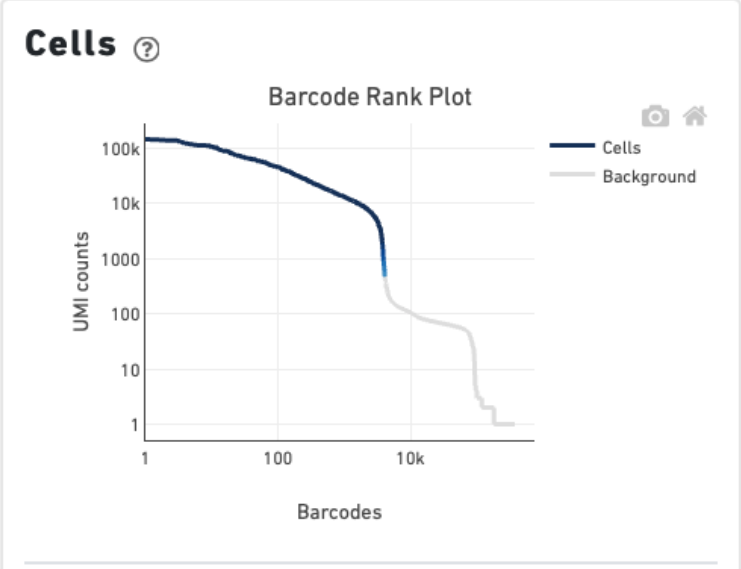
Cell Ranger

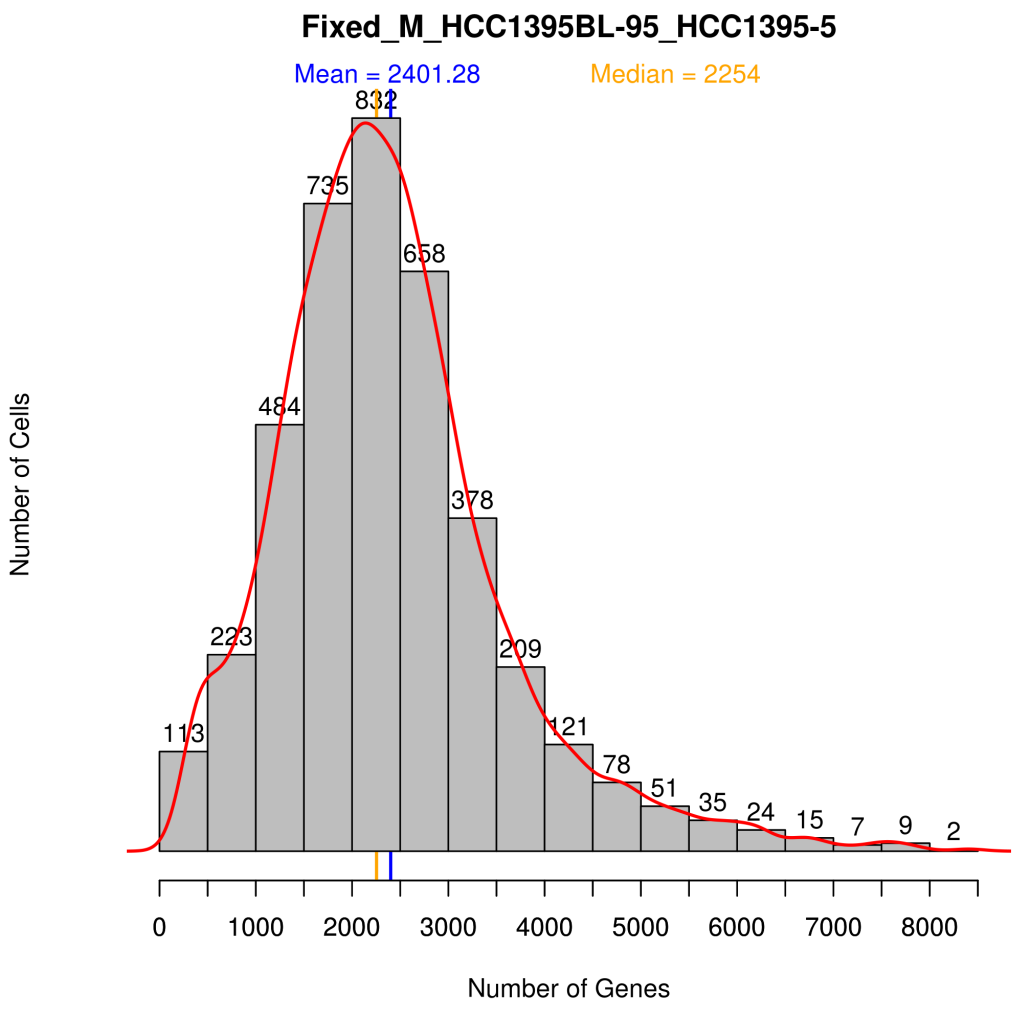
Frederick National Laboratory for Cancer Research

- **Barcode Rank Plot (Knee plot) can be used to determine sample quality**

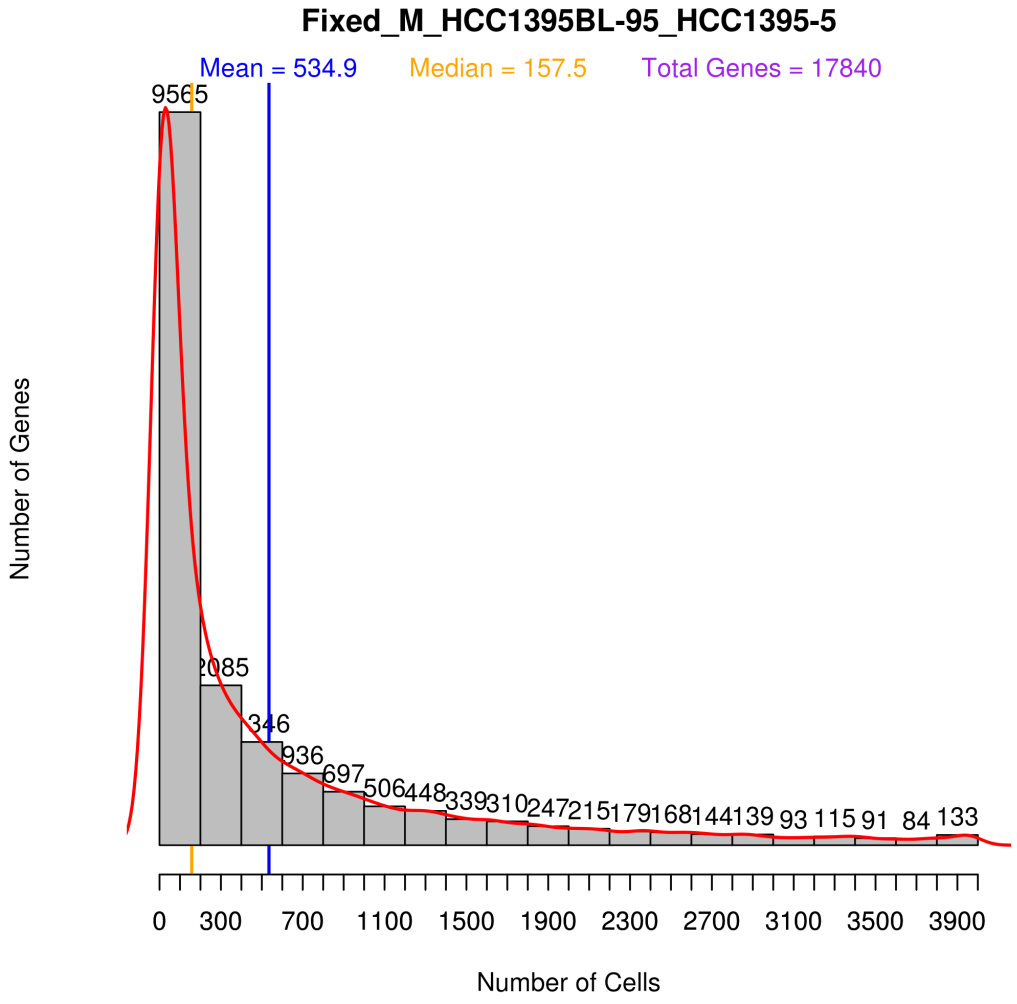
- **Cell Ranger 3 increased sensitivity for low UMI cell populations**







Genes per Cell



Cells per Gene

Cell Filtering

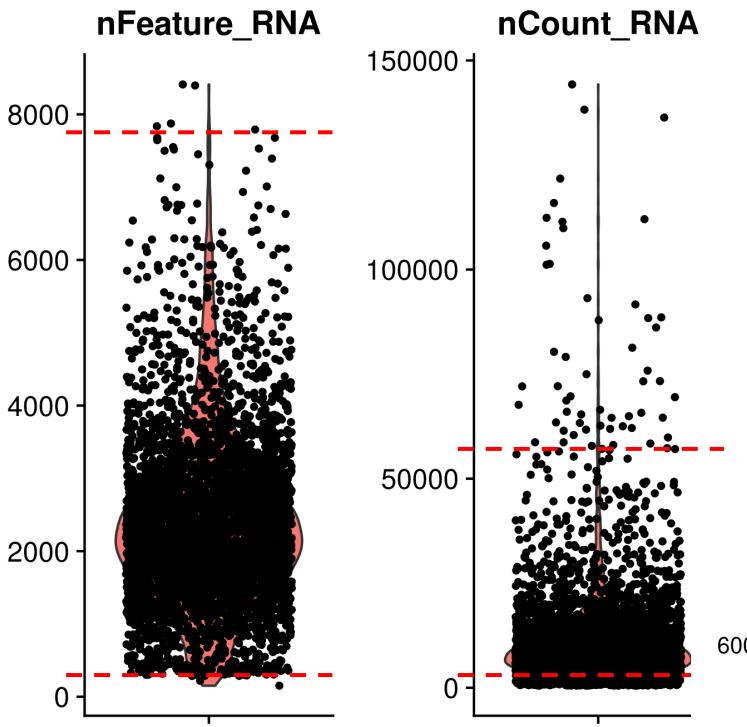
- **Useful because low quality cells or doublets/multiplets might be included in data**
- **Doublet/Multiplets are when more than one cell is captured and labeled with the same cell barcode**
- **Low quality cells include dying cells or cells with broken membranes**
  - Contains lower amounts of genes
  - Has a higher expression of mitochondrial genes

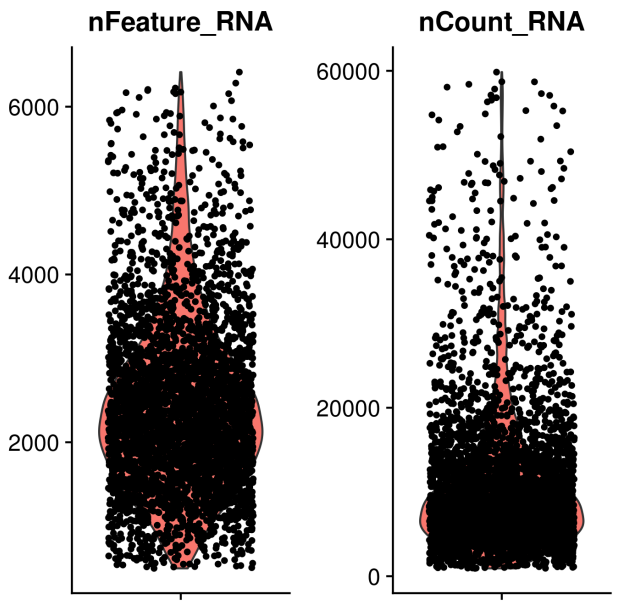
Cell Filtering

- Low quality cells or doublets/multiplets might be included in data
- Filtering is used to remove the excess noise to have a clean analysis
- Stringent filters risk losing useful data
- Loose filters risk leaving in noise

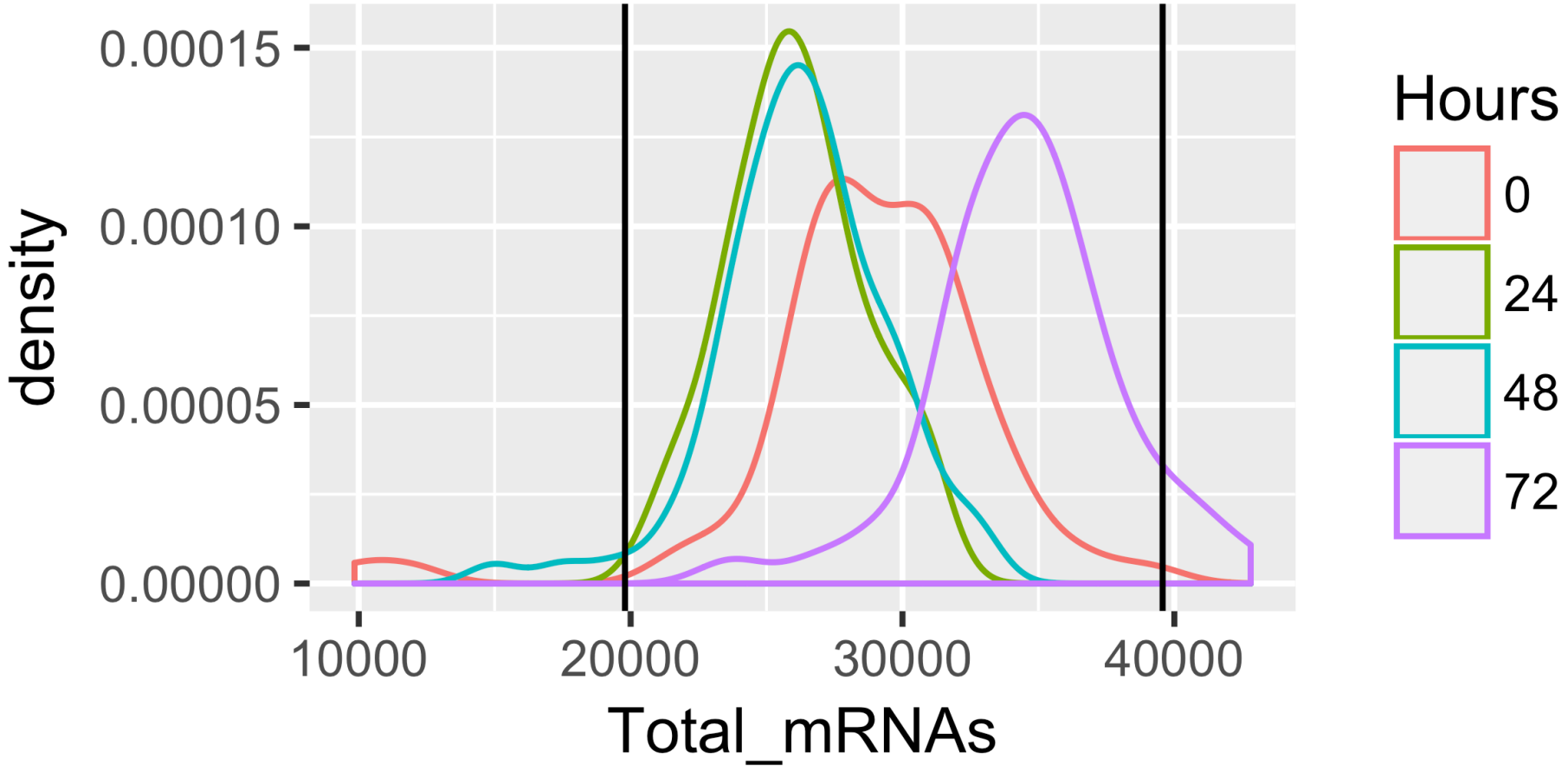
Cell Filtering

- Different cell types have different expression levels
- Filtering based on <u>UMI count, gene count</u>, and mitochondrial gene expression
- UMI count and gene count filters based on negative binomial distribution
- Other distribution and statistical methods can be used





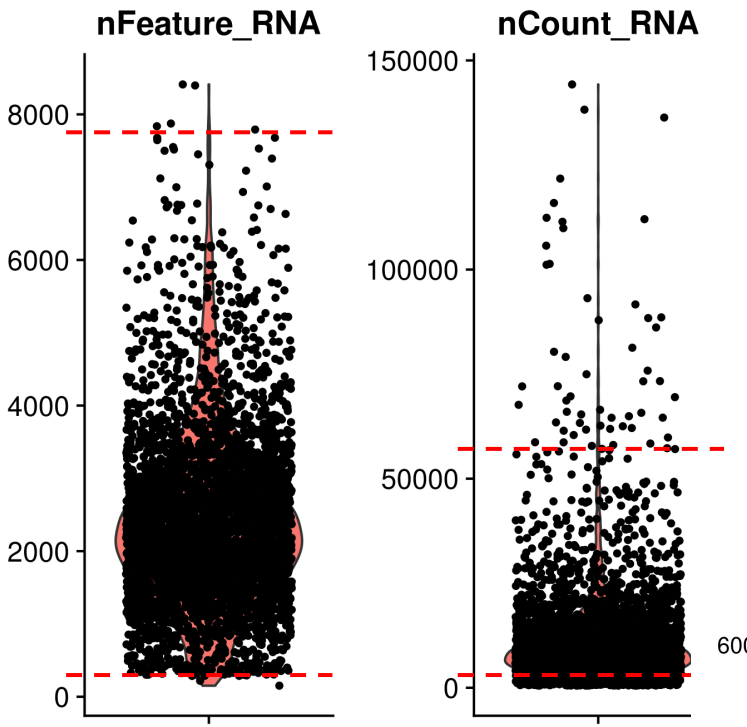
Cell Filtering

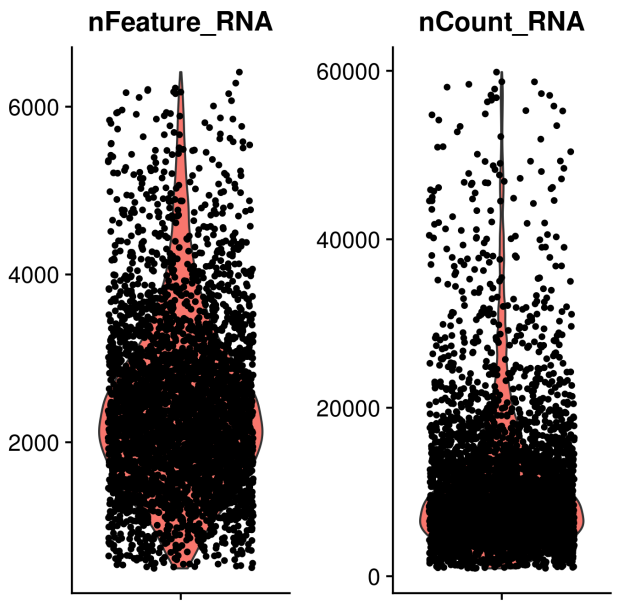


http://cole-trapnell-lab.github.io/monocle-release/docs/#getting-started-with-monocle

Cell Filtering

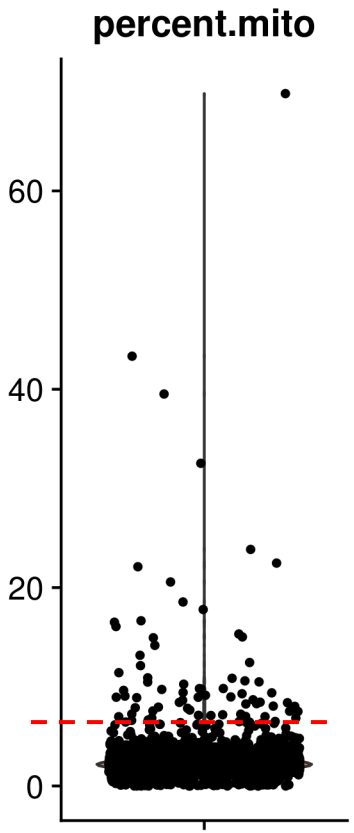
- **Different cell types have different expression levels**
- **Filtering based on <u>UMI count, gene count</u>, and mitochondrial gene expression**
- **UMI count and gene count filters based on negative binomial distribution**

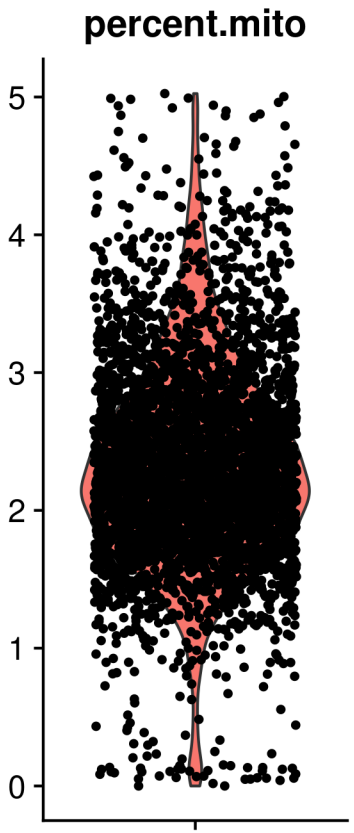




Cell Filtering

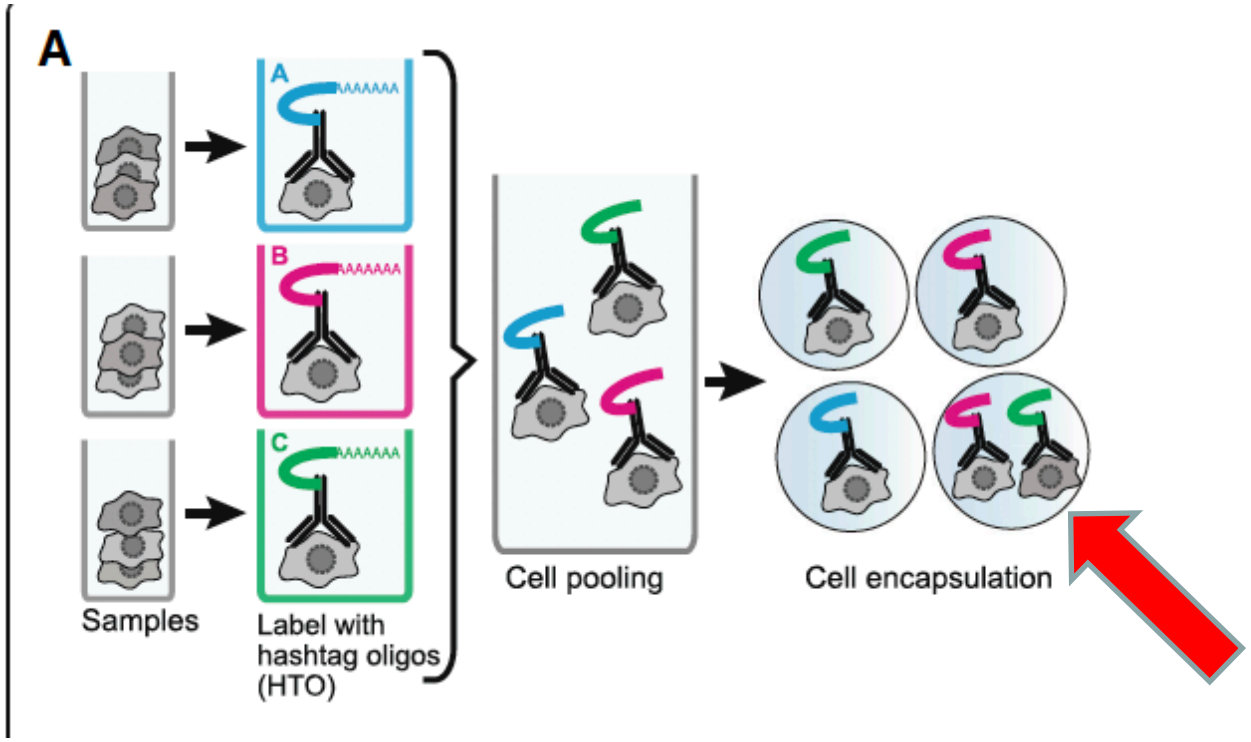
- Filtering based on UMI count, gene count, and <u>mitochondrial gene expression</u>
- Mitochondrial gene expression threshold is 4 median absolute deviation above median
- Mitochondrial fraction is linked to cell death, which may influence normalization
- Different cell types have different expression levels





Finding Doublets

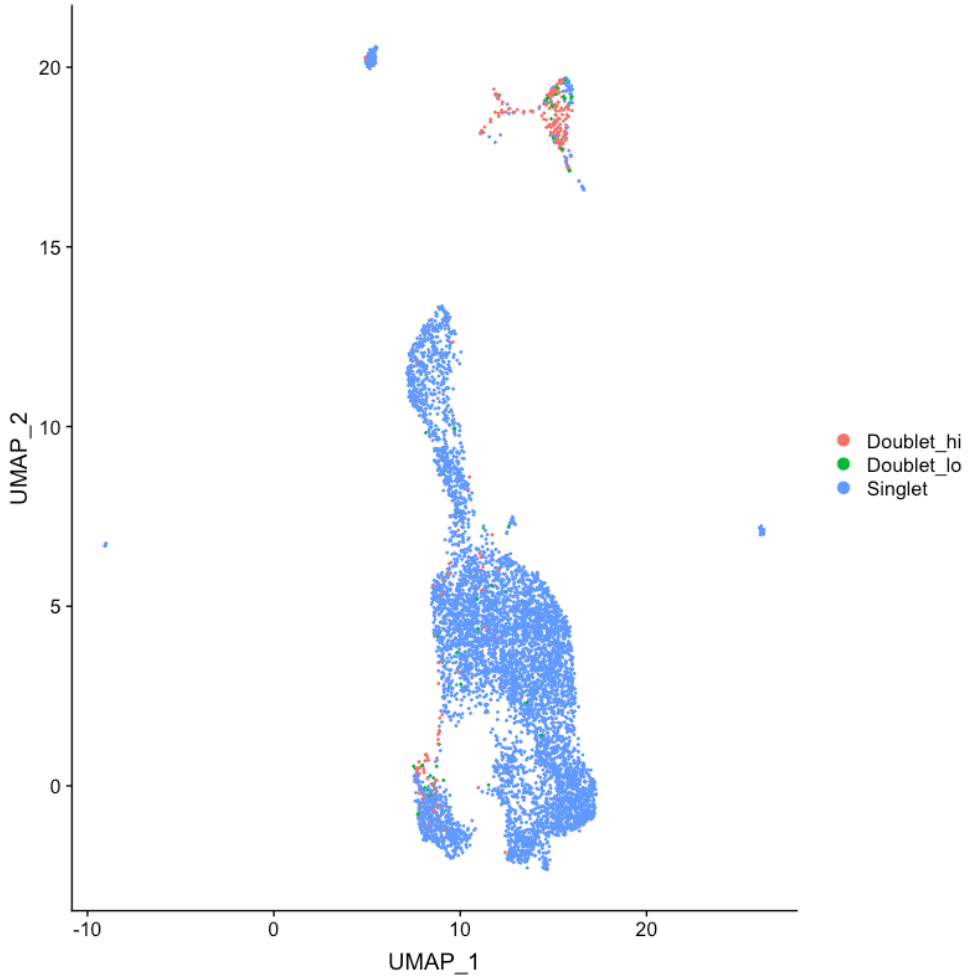
- Doublets (or multiplets) are a technical byproduct of single-cell droplet sequencing
- Doublets can interfere with downstream analysis by including high read counts per “cell” and changing cluster identities
- There is no current method to identify transcripts associated with the individual cells in doublets
- Doublets can be homotypic (same cell type) or heterotypic (different cell types)



Finding Doublets

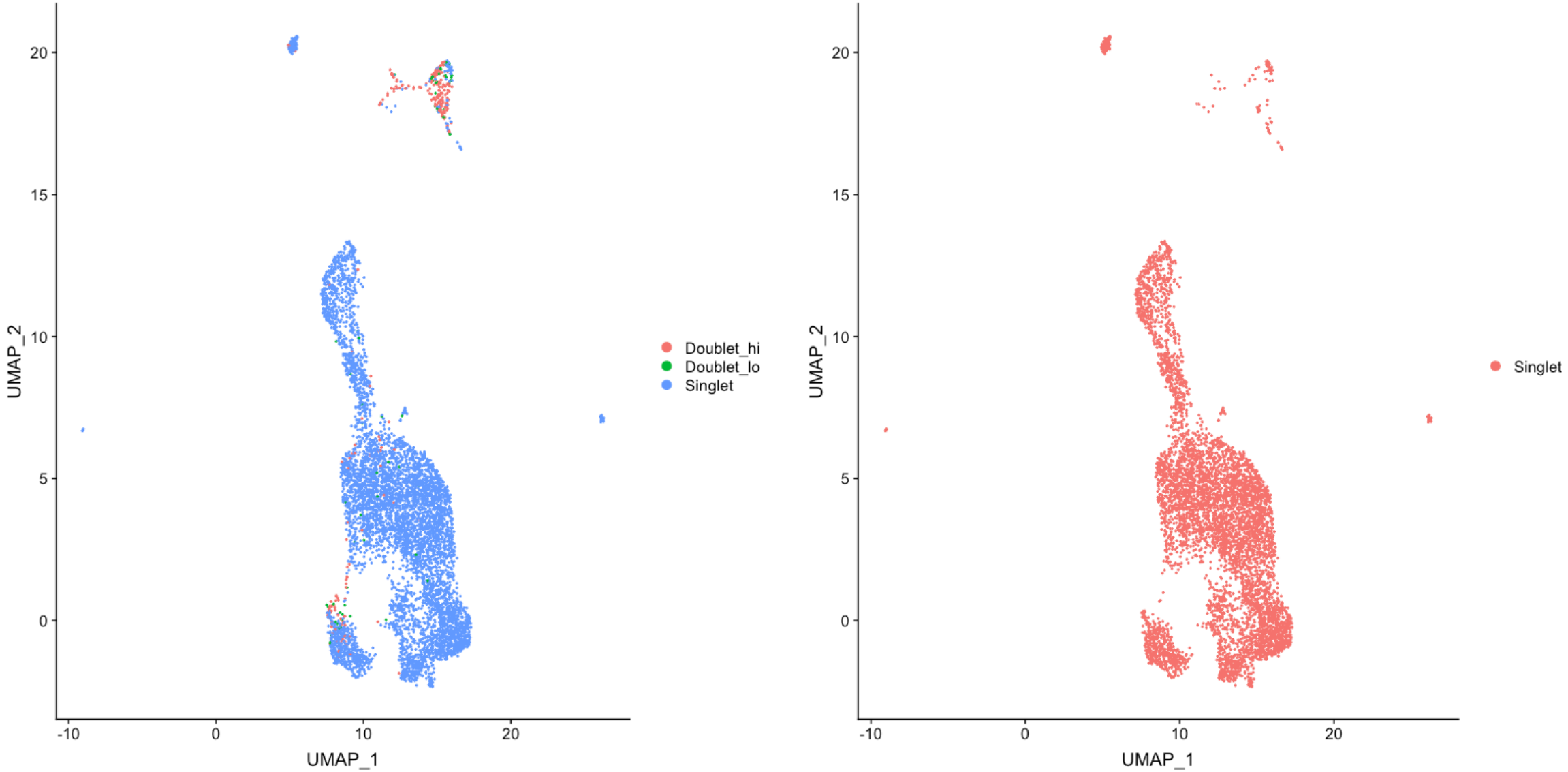
Frederick National Laboratory for Cancer Research

- **Statistical removal of doublets:**
  - UMI count and gene count based filters
- **Algorithmic removal of doublets:**
  - DoubletFinder (McGinnis, Murrow and Gartner 2019)
  - Scrublet (Wolock, Lopez, and Klein 2018)
- **The estimated doublet rate as provided by 10x Genomics is:**
  - $\boldsymbol{Doublet\ Fraction} = \mathbf{0.008} \times \frac{\boldsymbol{n_{Cells}}}{\mathbf{1000}}$



Removal of doublets allows for downstream re-clustering

Frederick National Laboratory for Cancer Research

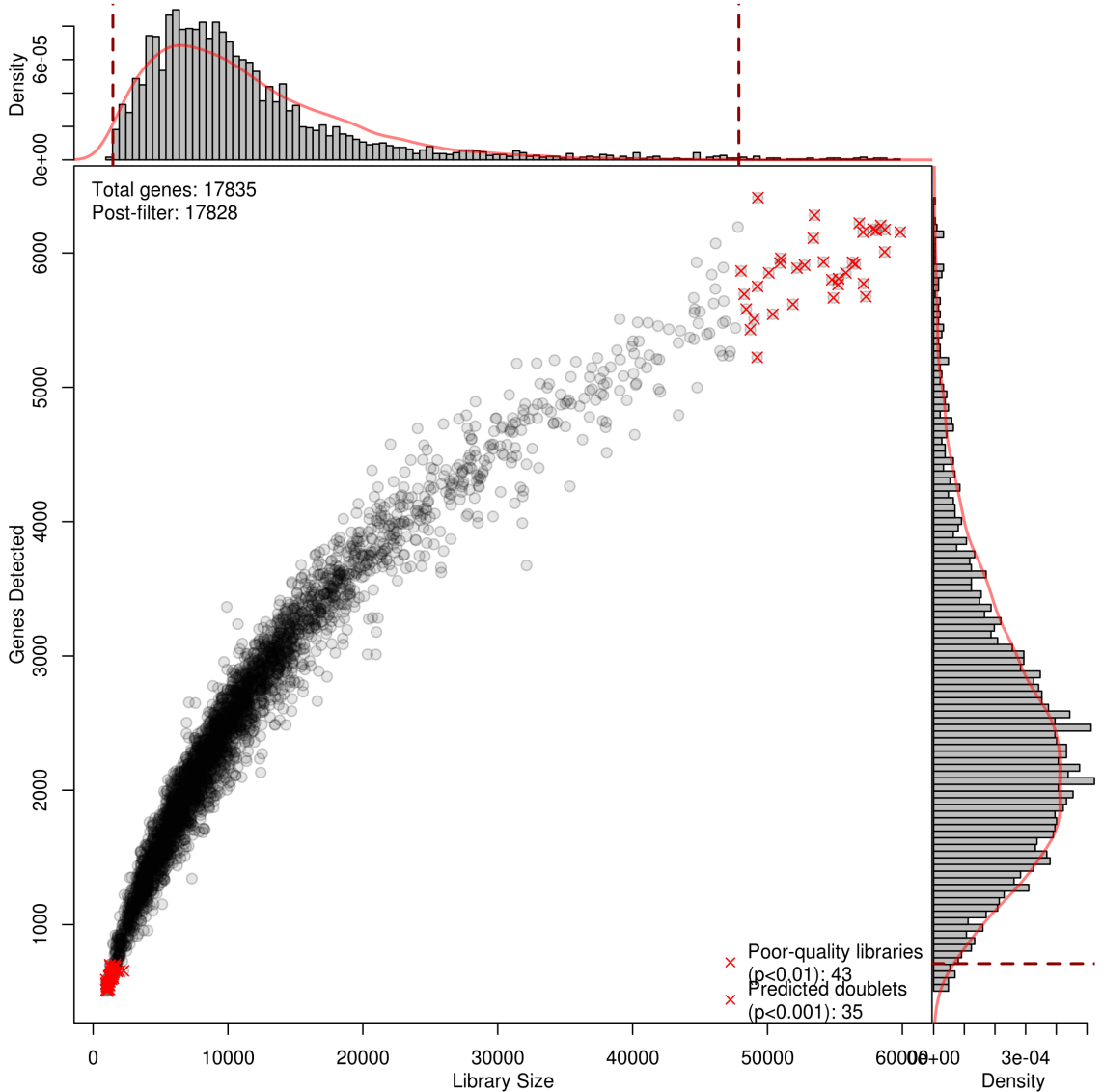


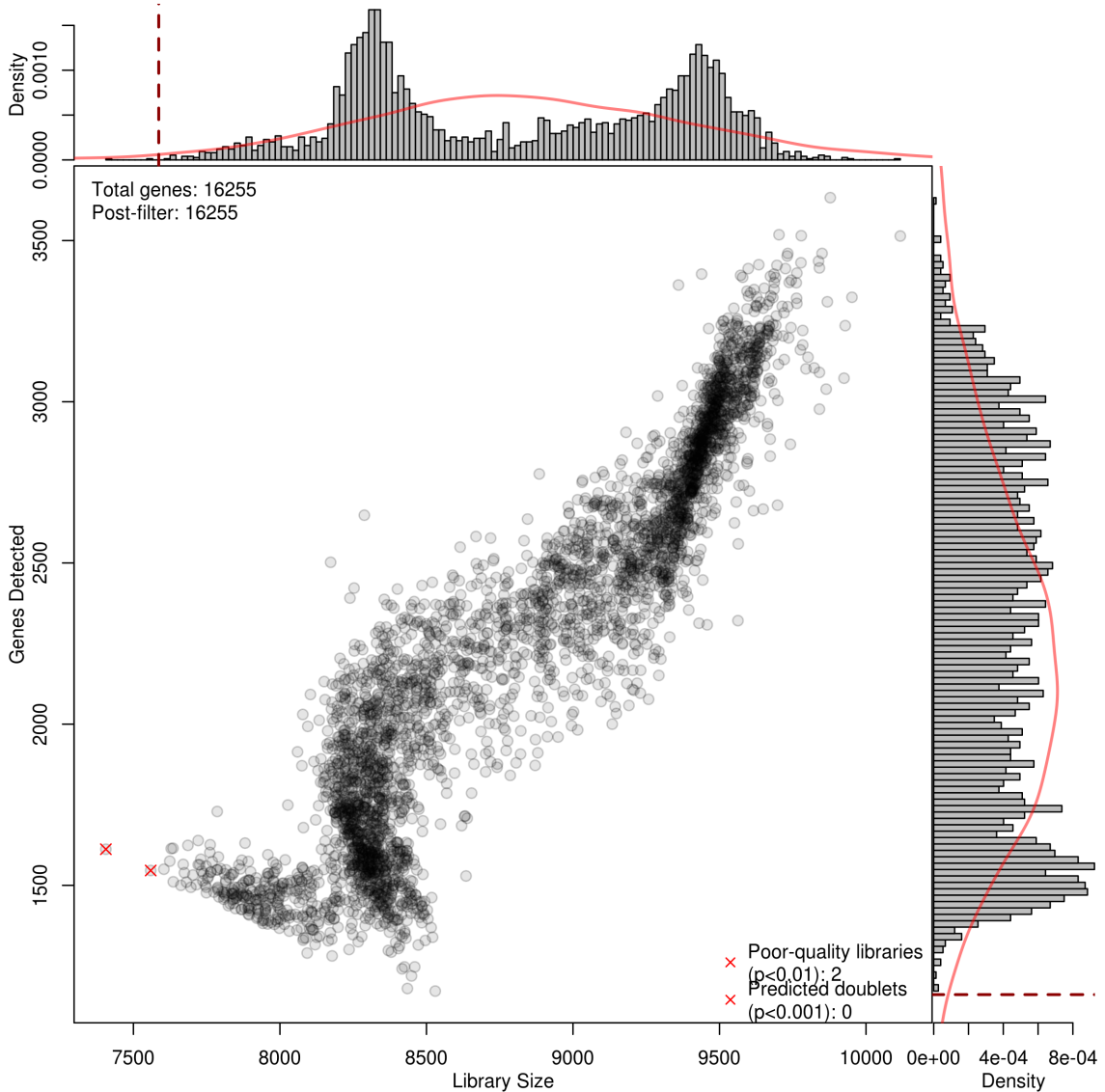
Normalization

- **Aim is to remove technical effects while retaining biological variation**
  - Differences in detected gene expression can be due to sequencing depth of cell
- **Many different normalization techniques available**
- **Seurat has different normalization algorithms available**
  - NormalizeData, ScaleData
    - NormalizeData - Default normalization is log normalize. Each cell divided by total counts, multiplied by scale factor, and natural log transformed
    - ScaleData - Scales and centers features in the data. Can optionally regress out effects of variables (i.e. mitochondrial expression, cell cycle)
  - scTransform - combined NormalizeData, FindVariableFeatures, ScaleData

Seurat log Normalize vs scTransform

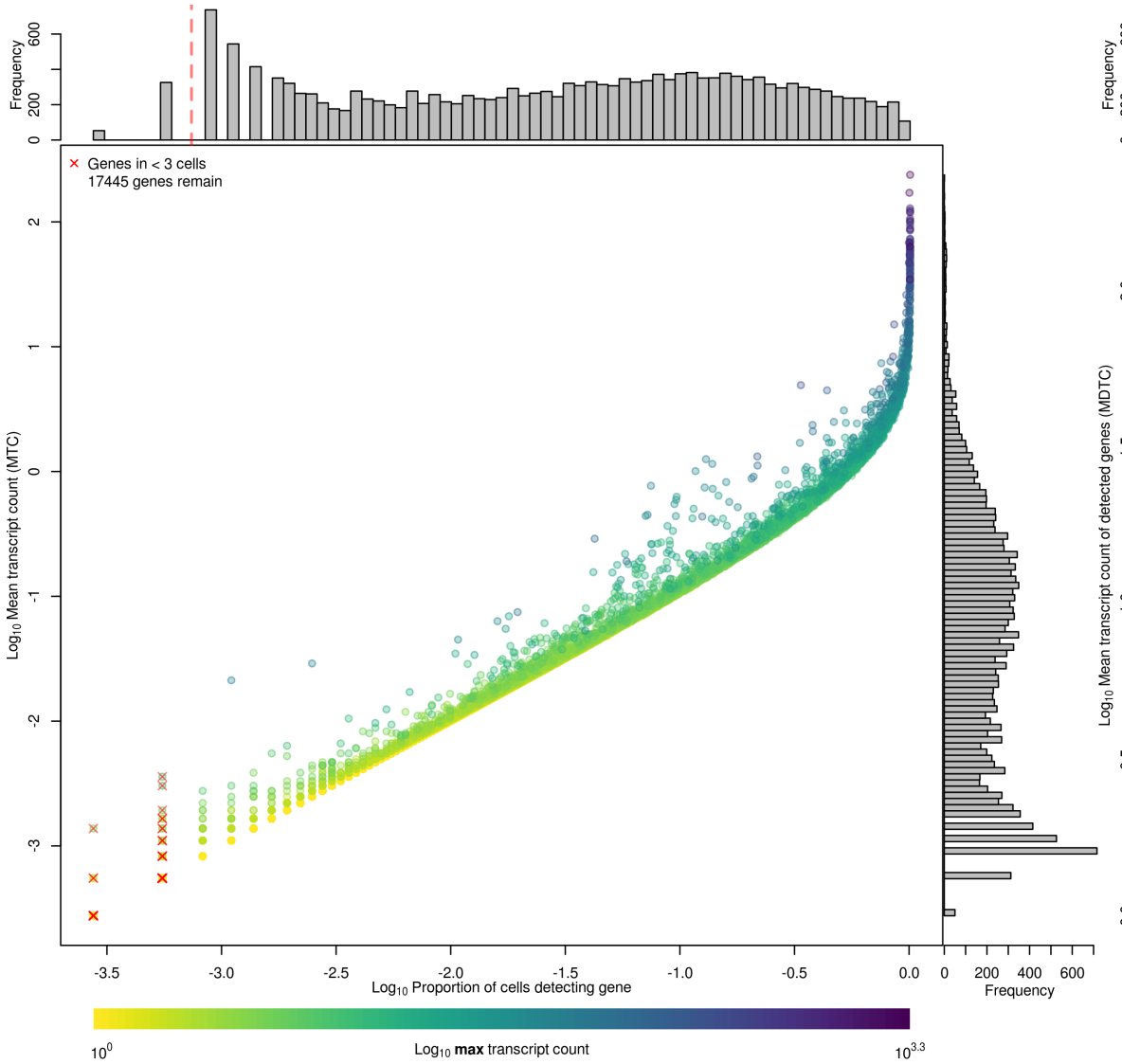
Frederick National Laboratory for Cancer Research

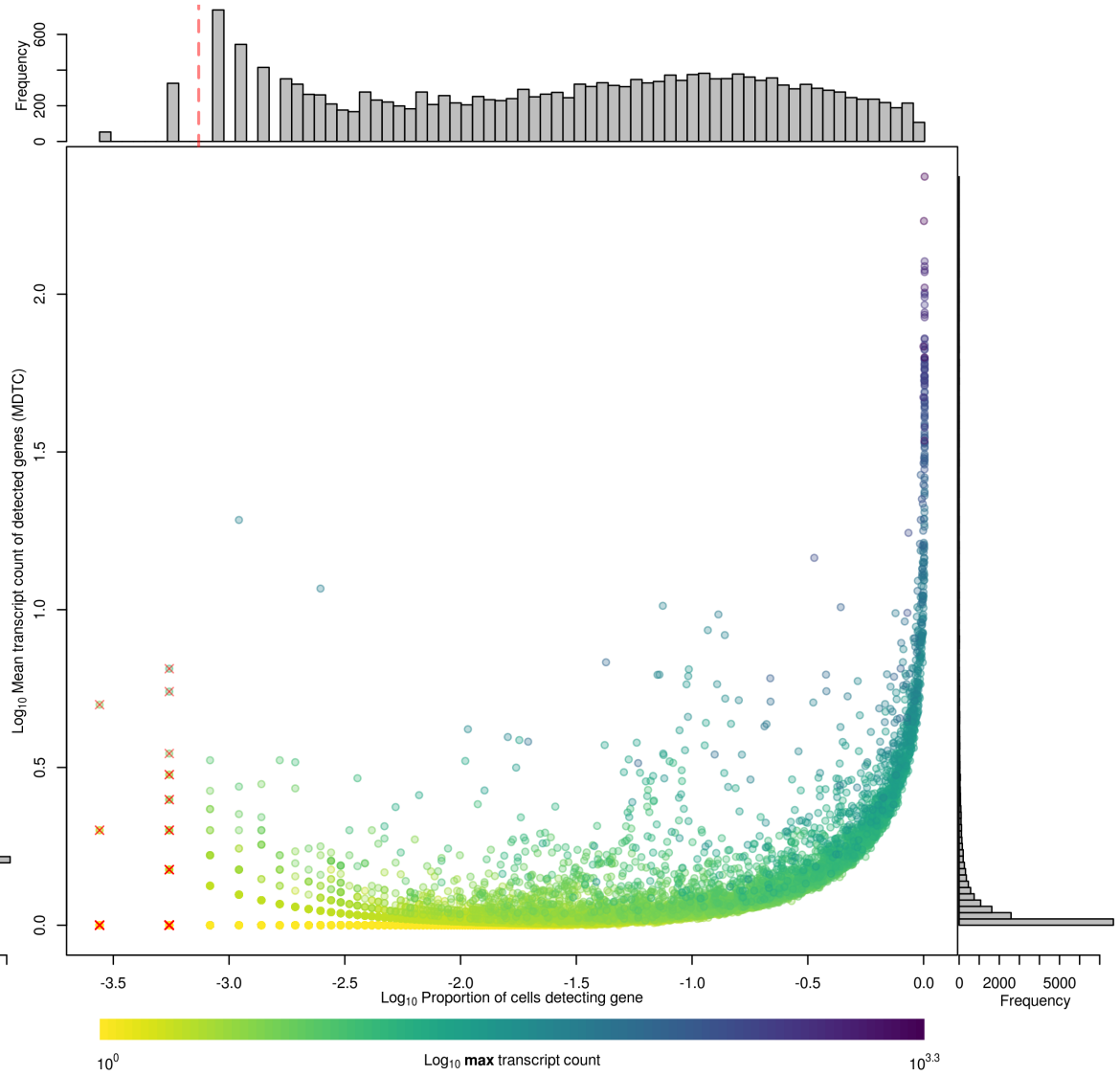




Expression Plot v2

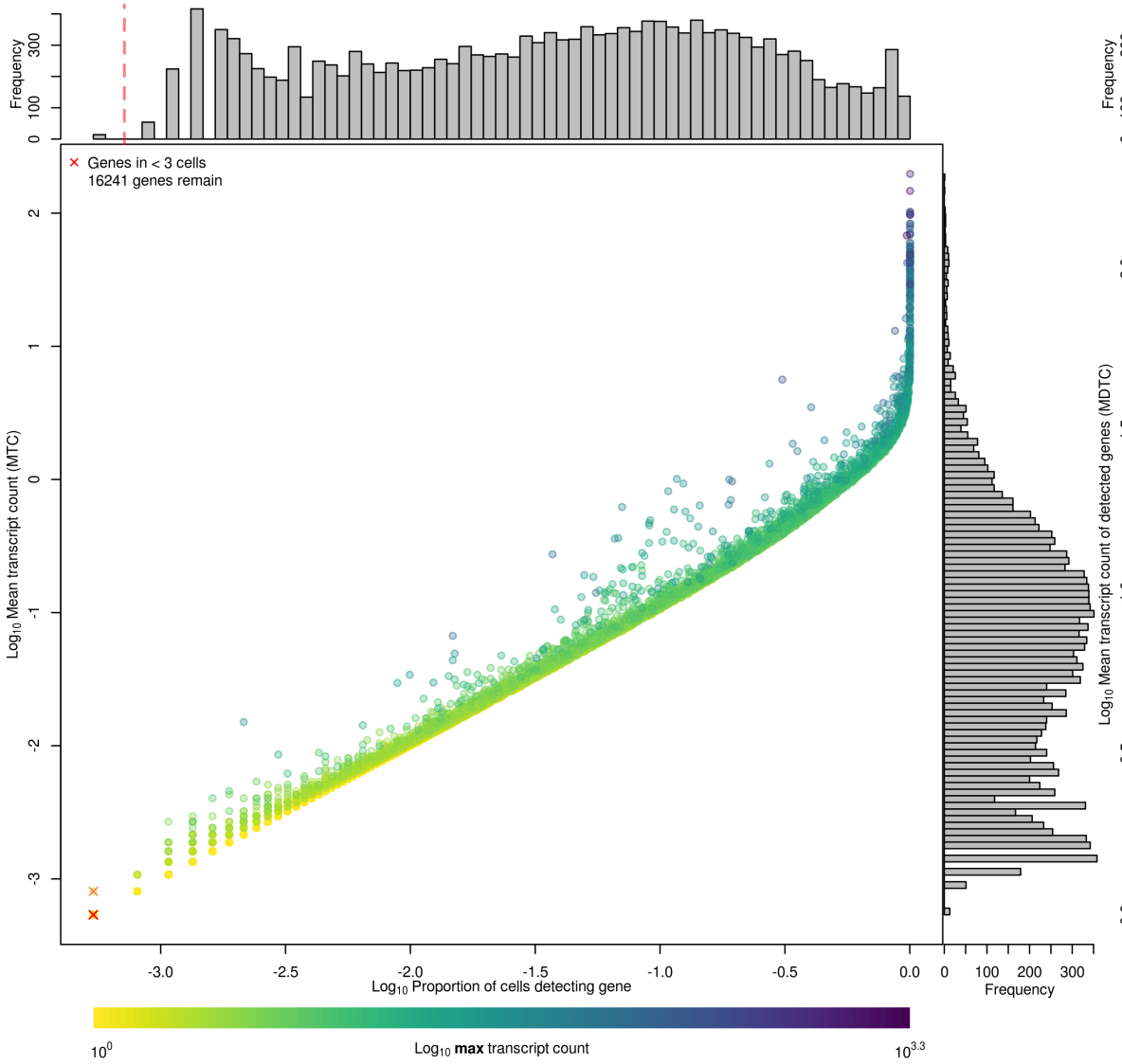
Frederick National Laboratory for Cancer Research

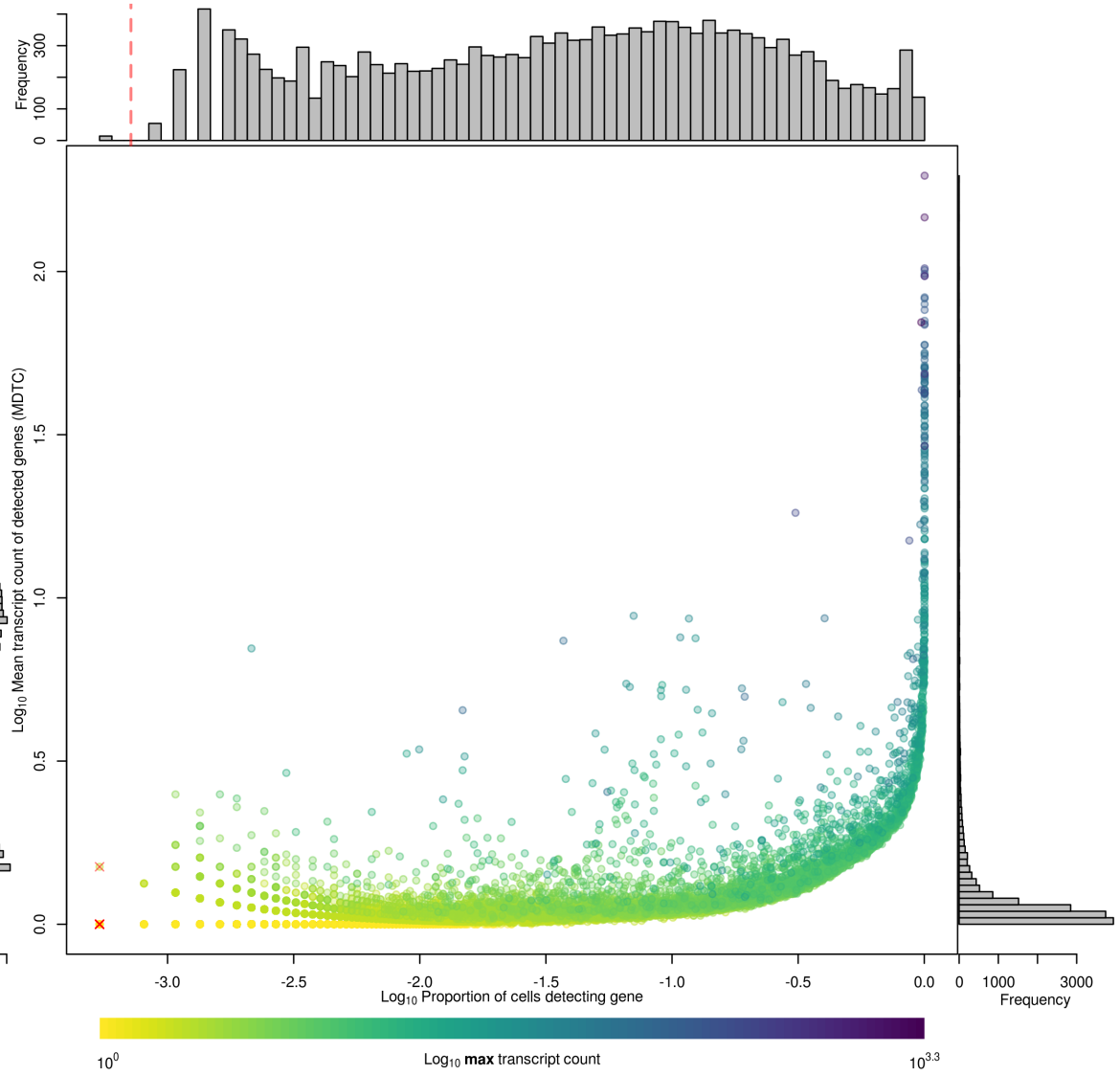




Expression Plot – v3 scTransform

Frederick National Laboratory for Cancer Research





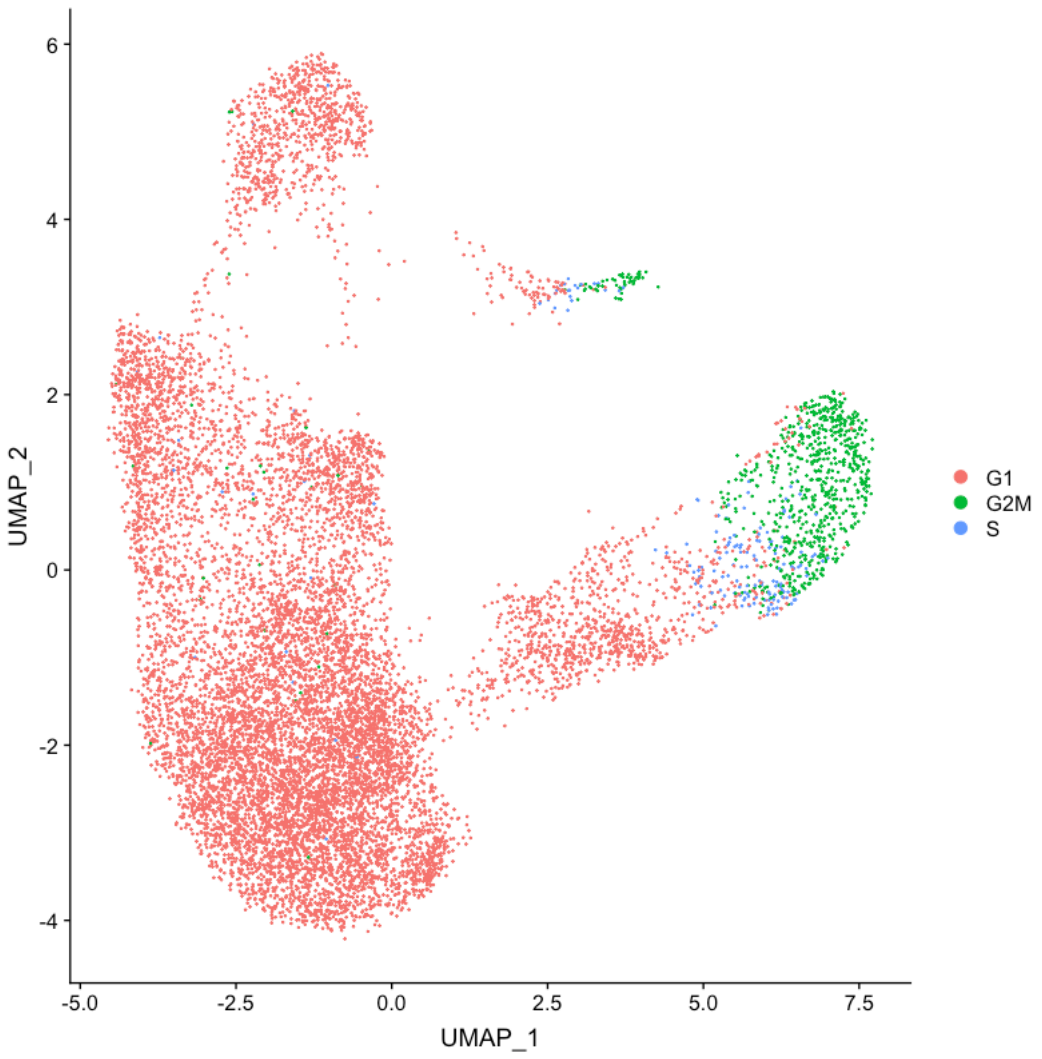
Cell Cycle

Frederick National Laboratory for Cancer Research

- **Cell cycle can introduce bias or obscure differences in expression by cell types**

- **Cell cycle can be identified using available tools, including:**
  - Seurat: CellCycleScoring
  - Scran: Cyclone

- **A variety of tools and techniques are available that can be used to remove effect**
  - ccRemover (Li and Barron 2017)
  - Seurat – ScaleData can be used to regress out effects after labelling cell cycle

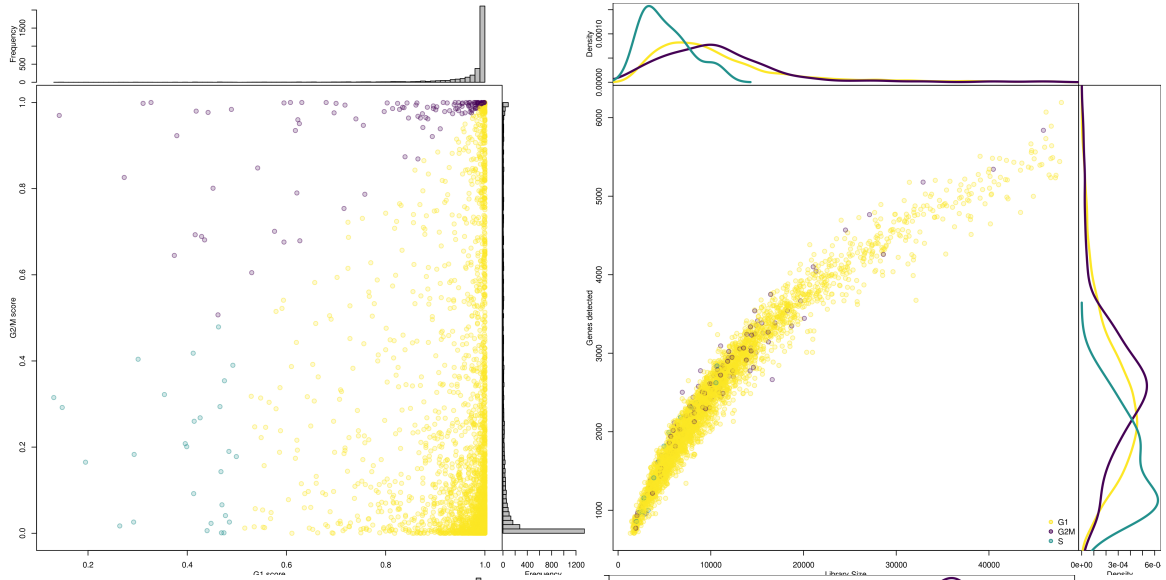


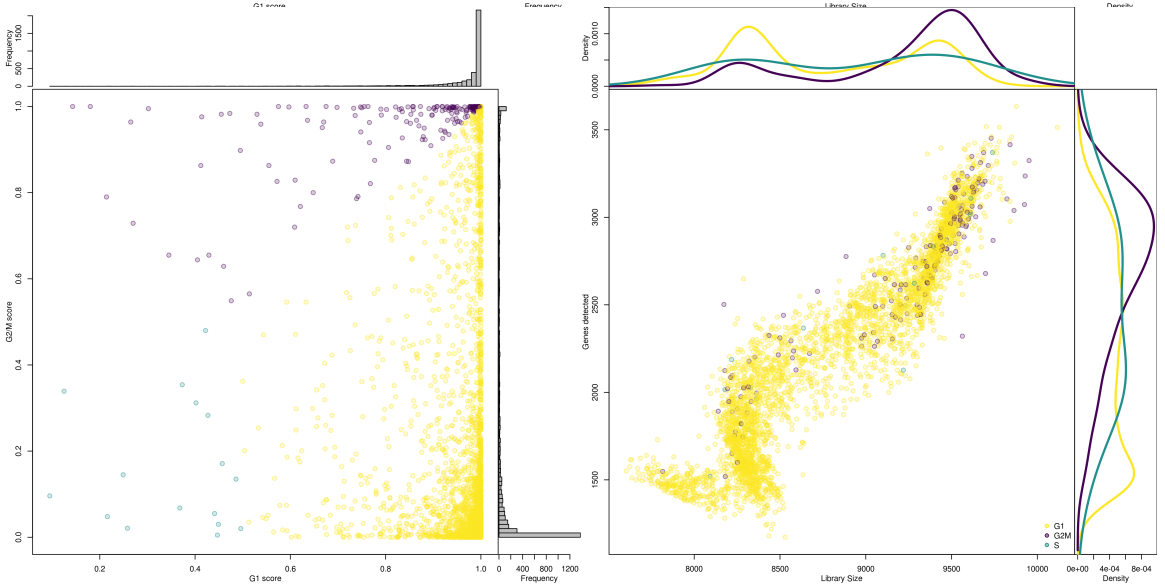
Cell Cycle

- **Cell cycle can introduce bias or obscure differences in expression by cell types**

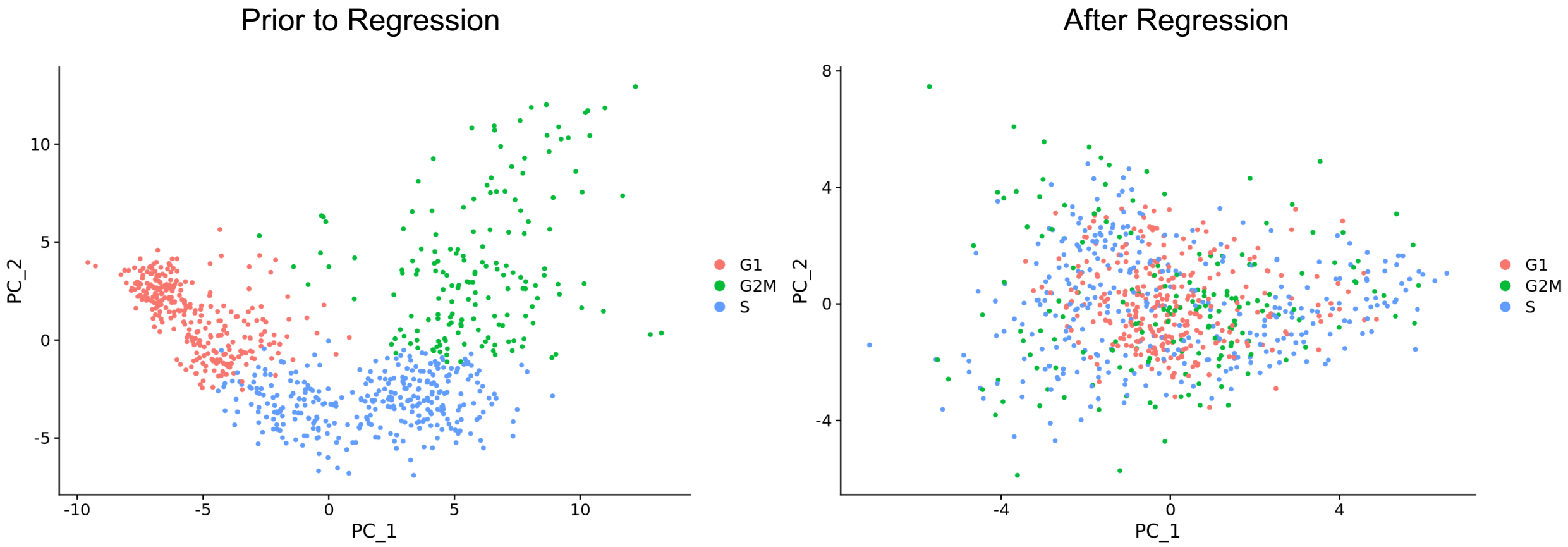
- **Cell cycle can be identified using available tools, including:**
  - Seurat: CellCycleScoring
  - Scran: Cyclone

- **A variety of tools and techniques are available that can be used to remove effect**
  - ccRemover (Li and Barron 2017)
  - Seurat – ScaleData can be used to regress out effects after labelling cell cycle





Regressing out cell cycle effects

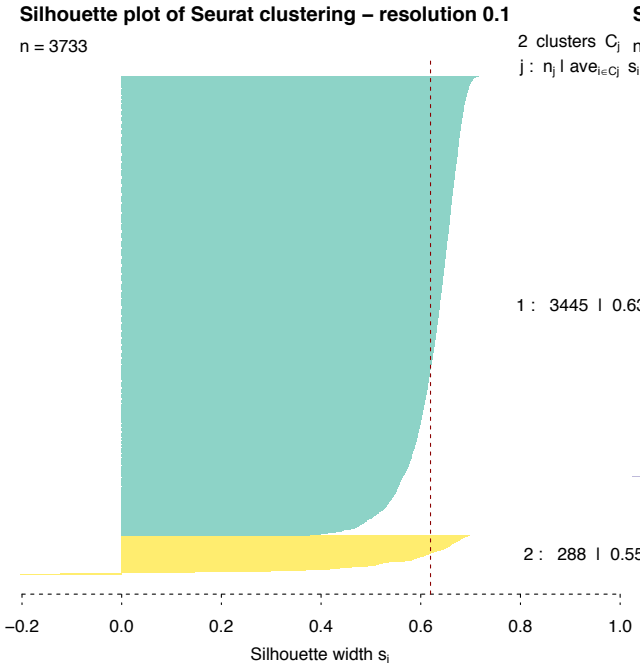


Measuring Cluster Quality

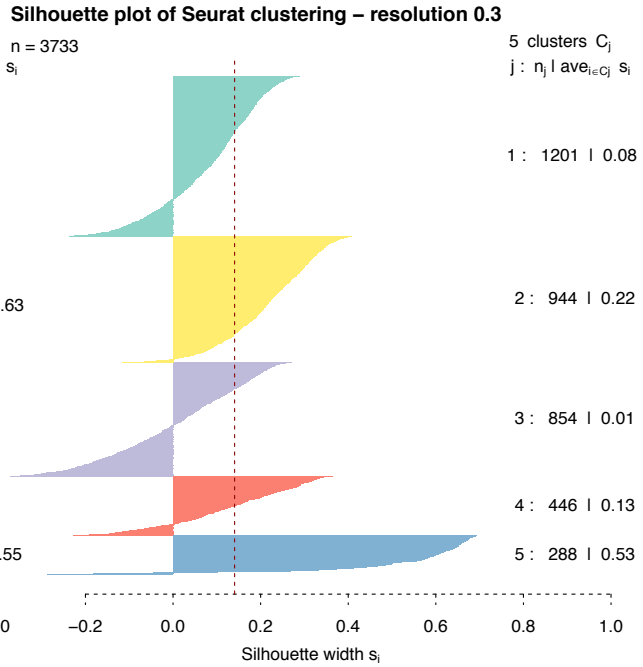
- **Different numbers of clusters can be used to group cells within a sample**
- **Can be difficult to determine appropriate number of clusters without prior knowledge**
- **Metrics can be used to measure the quality of the clusters**
  - Silhouette score, Rand index, Davies-Bouldin index
- **Cluster size that results in best score indicates an appropriate number of clusters**

Silhouette Plots – After Seurat Clustering

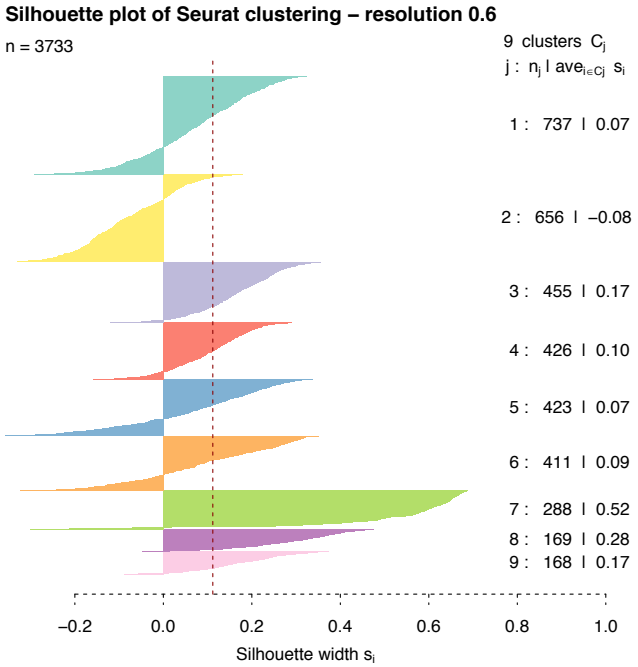
Frederick National Laboratory for Cancer Research



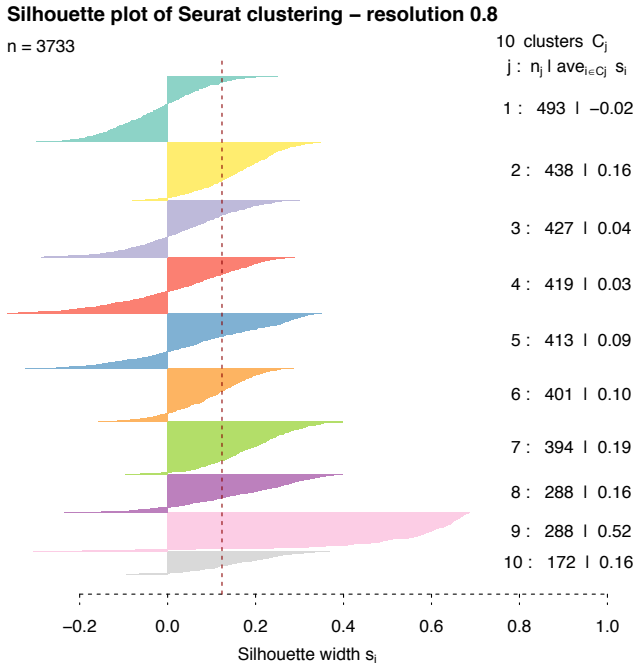
Average silhouette width : 0.62



Average silhouette width : 0.14



Average silhouette width : 0.11



Average silhouette width : 0.12

Imputation

Frederick National Laboratory for Cancer Research

- **Noise and signal dropout are (currently) unavoidable errors in single cell RNA-Seq**
- **Characterized by zero count genes in individual cells**
  - 10x Genomics v3 captures 30-32% of mRNA transcripts per cell
- **Imputation attempts to fill in those zeros based on:**
  - Count distribution
  - Overdispersion
  - Sparsity of the data
  - Noise modeling
  - Gene-gene dependencies

0	2	5.0	3.0	6.0	NaN
1	9	NaN	9.0	0.0	7.0
2	19	17.0	NaN	9.0	NaN
3	7	10.0	3.0	6.0	4.0
4	2	8.0	10.0	NaN	3.0

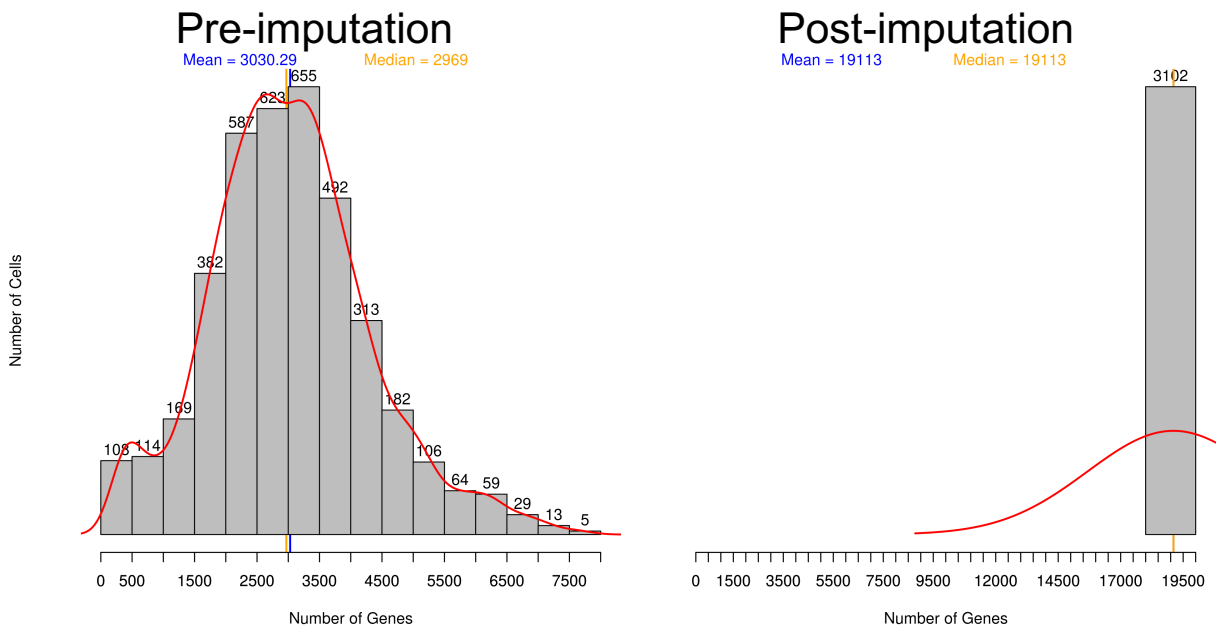
→

0	2.0	5.0	3.00	6.00	4.666667
1	9.0	10.0	9.00	0.00	7.000000
2	19.0	17.0	6.25	9.00	4.666667
3	7.0	10.0	3.00	6.00	4.000000
4	2.0	8.0	10.00	5.25	3.000000

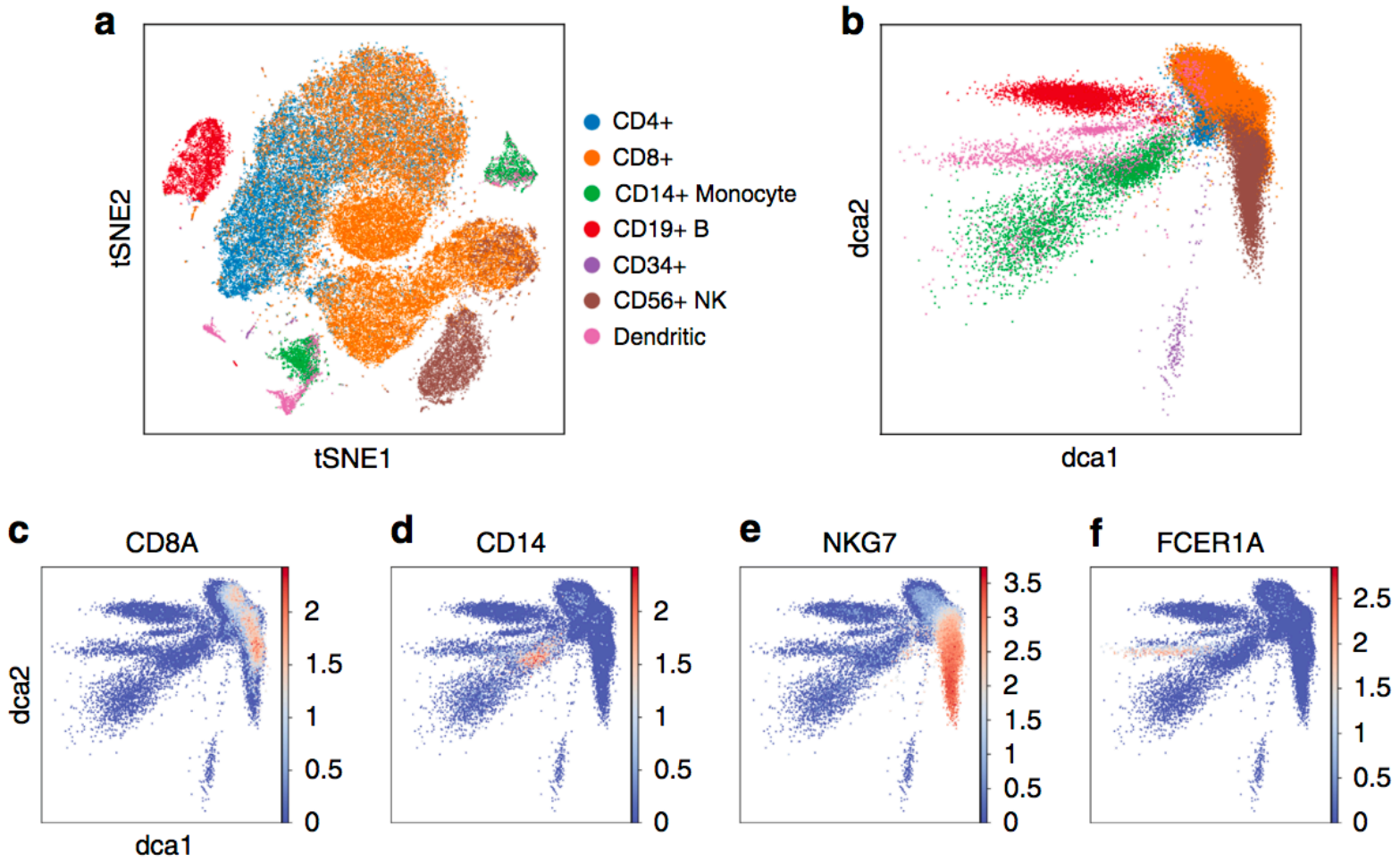
Available imputation tools include:

- dca (Deep count autoencoder) (Erslan, et al. 2019)
- SCRABBLE (Peng, et al. 2019)
- SAVER (Huang, et al. 2018)
- DrImpute (Gong, et al. 2018)
- scImpute (Li and Li 2018)
- bayNorm (Tang, et al. 2018)
- knn-smooth (Wagner, Yan and Yanai 2018)
- MAGIC (van Dijk, et al. 2017)
- CIDR (Lin, Troup, and Ho 2017)

Genes per Barcode (dca)



Imputation effects on clusters



Imputation effects on gene expression