THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Introduction to Single Cell RNA-Seq (scRNA-seq)

Basil Khuder
Senior Bioinformatician
UAB Biological Data Science Core

U-BDS Core Members

Liz Worthey, Ph.D.

Associate Professor, Department of Pediatrics and Pathology
eaworthey@uabmc.edu

U-BDS Director
Director, Center for Computational Genomics and Data Sciences
Section Chief, Bioinformatics, director of the Bioinformatics Section in the Division of Genomics Diagnostics and Bioinformatics, Department of Pathology
Associate Director, Hugh Kaul Precision Medicine Institute
Scientist, Center for Clinical and Translational Science

Brittany Lasseigne, Ph.D.

Assistant Professor, Department of Cell, Developmental and Integrative Biology
bnp0001@uab.edu

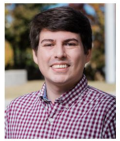
U-BDS Co-director
Associate Scientist, Experimental Therapeutics, O'Neal Comprehensive Cancer Center
Associate Scientist, Center for Clinical & Translational Science
Scientist, Hugh Kaul Precision Medicine Institute
Scientist, Informatics Institute
Scientist, Center for Neurodegeneration and Experimental Therapeutics

Lara Ianov, Ph.D.

Senior Bioinformatician, Civitan International Research Center
lianov@uab.edu

U-BDS Managing Director
Manager, Neurodevelopmental Bioinformatics Initiative, Civitan International Research Center

Austyn Trull, B.Sc.



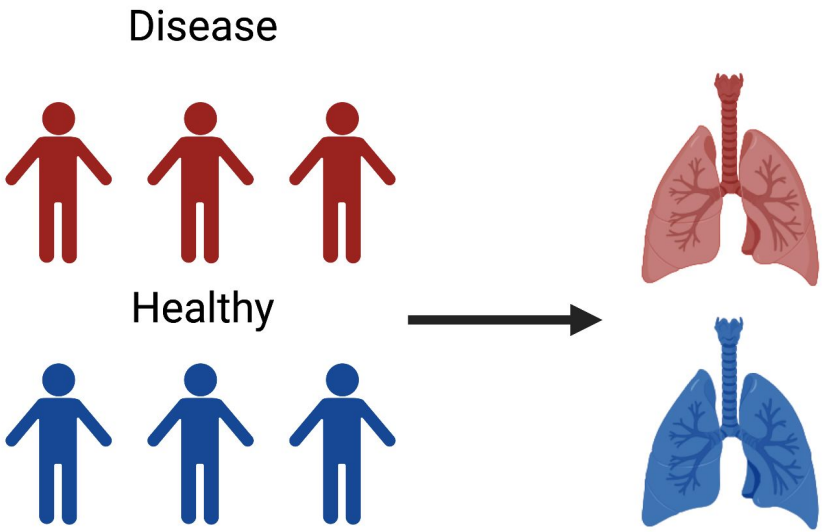
Bioinformatician I, Department of Pediatrics and Pathology
agtrull@uabmc.edu

U-BDS Staff
Bioinformatician I, Department of Pediatrics and Pathology
Software Developer

Overview

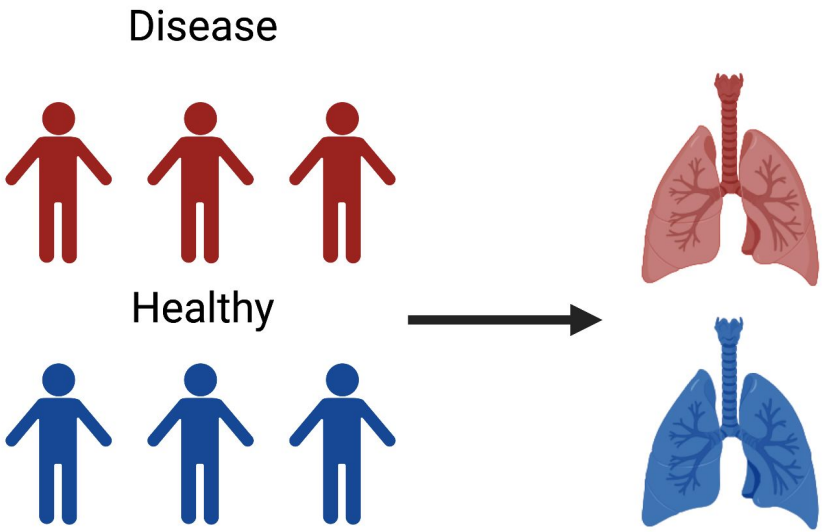
- **What is Single-Cell RNA-Seq?**
  - Most commonly used technologies/platforms
- **Overview of Computational Biology Methods**
  - *Secondary Analysis*
    - 10X Cell Ranger (commercially based pipelines)
    - STARSolo, Alevin (Alevin-fry), kallisto bustools (scientific community)
  - *Tertiary Analysis*
    - Seurat/Scanpy
    - Steps to analyze data
      - Quality control, normalization, dimensionality reduction, cluster identification, data projection, differential expression

Let’s Pretend…



- We’re researchers looking to develop new therapeutics to target an inflammatory lung disease
- Cohort of six individuals *(3 healthy, 3 disease)*

Let's Pretend...



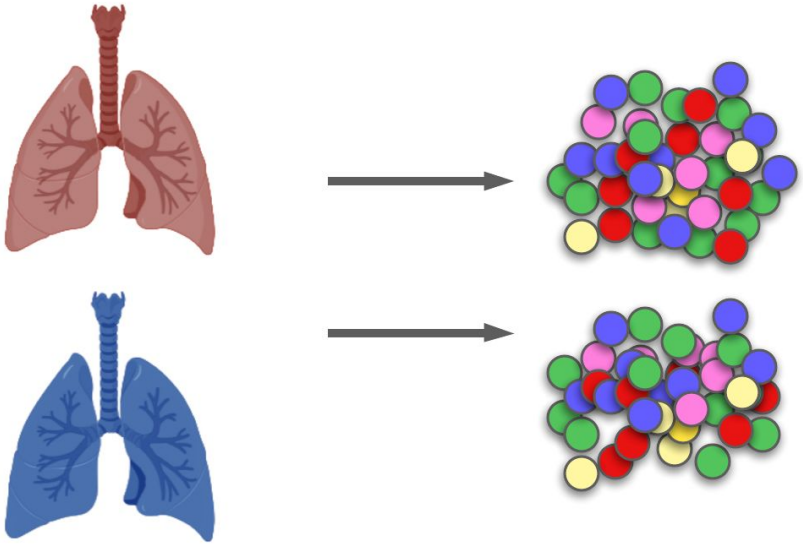
- We have the option of running a single-cell RNA seq analysis or bulk RNA-seq
- We have chosen to go with single-cell…

But why?

Recall from Thursday

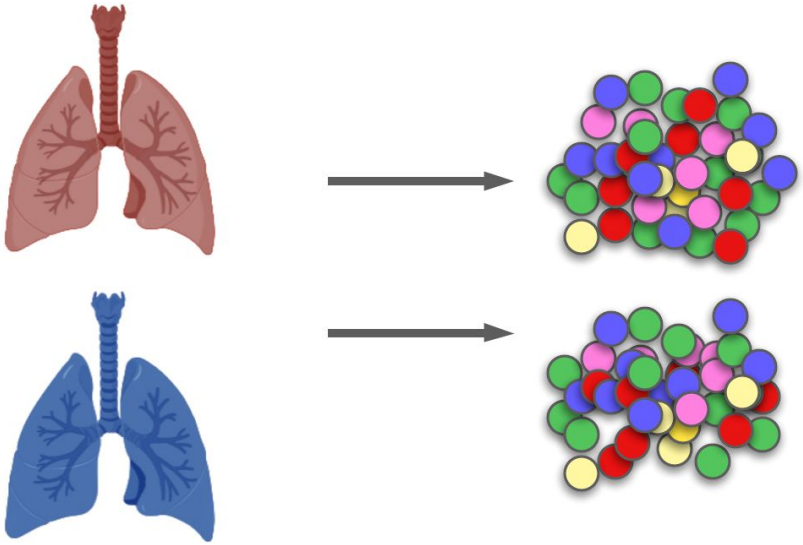
- Bulk RNA-Seq *(or simply "RNA-Seq")*
  - Gives us the average expression across ***all*** cells within a tissue type
  - Confounds any variability within cell-types
  - Gives a high-level picture of the transcriptional landscape
  - Cheaper than single-cell

All Cells Are Not Created Equal



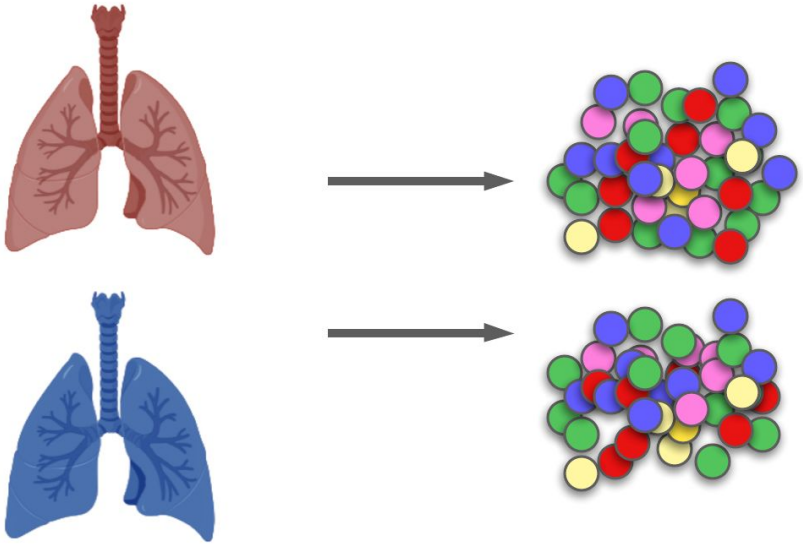
- Lung tissue is comprised of multiple different cell types (over 40..)
- Bulk RNA-Seq doesn’t tell us which cells are the primary drivers of inflammatory gene expression
- Identifying cell-specific expression can lead to a more targeted therapeutic

All Cells Are Not Created Equal



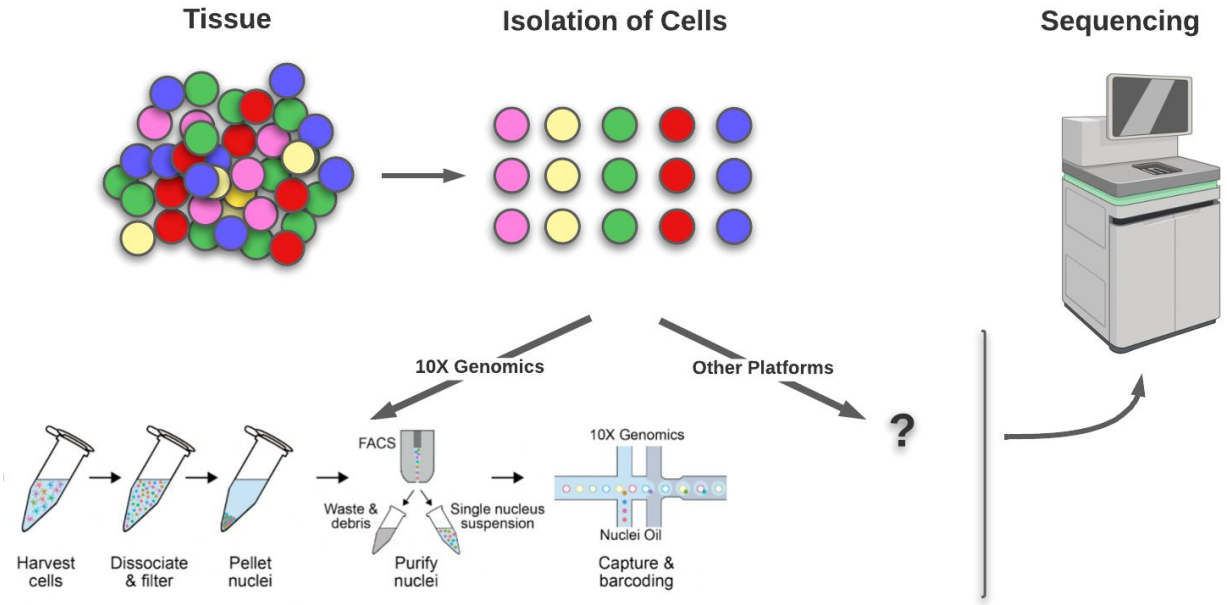
- Through a method called ***Pseudobulk***, we can still answer fundamental question of disease/non-disease and emulate a bulk RNA-seq using single-cell
- We'll touch on this later…

Some More Questions scRNA-Seq Can Answer...



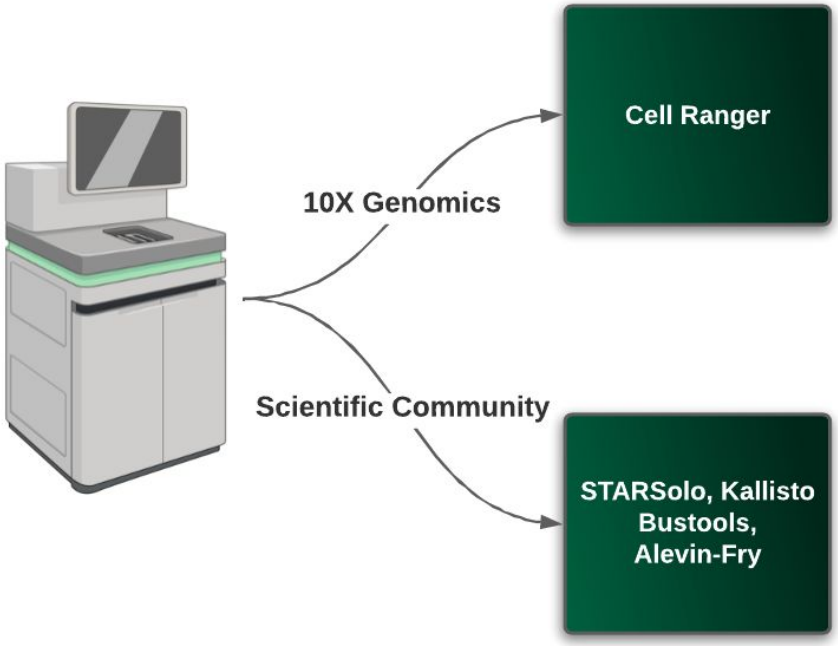
- How does the cellular landscape of healthy individuals differ?
- Are there novel inflammatory cell populations in the disease group?

The Process

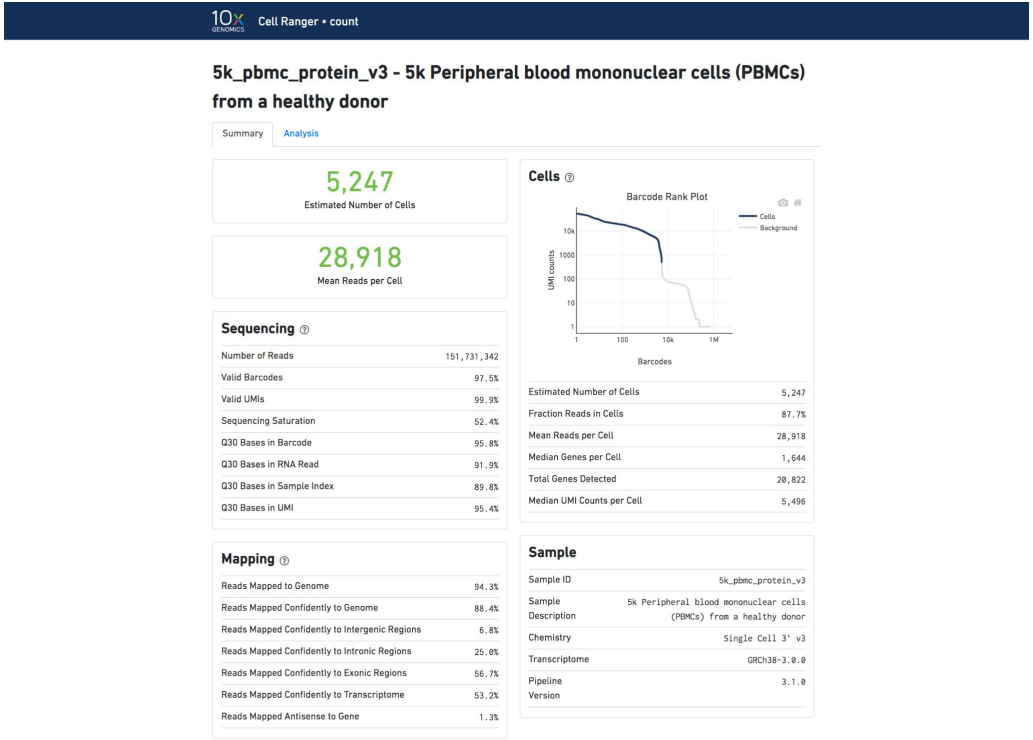


Katherine E. Savell*, Jennifer J. Tuscher*, Morgan E. Zipperly*, Corey G. Duke*, Robert A.Phillips III*, Allison J. Bauman, Saakshi Thukral, Faraz A. Sultan, Nicholas A. Goska, Lara Ianov, Jeremy J. Day (Science Advances, June, 2020). *A dopamine-induced gene expression signature regulates neuronal function and cocaine response* DOI: 10.1126/sciadv.aba4221

The Process



Single-Cell RNA-Seq - 10X Genomics



10x GENOMICS Cell Ranger • count

5k_pbmc_protein_v3 - 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor

Summary | Analysis

5,247
Estimated Number of Cells

28,918
Mean Reads per Cell

Sequencing

Number of Reads	151,731,342
Valid Barcodes	97.5%
Valid UMIs	99.9%
Sequencing Saturation	52.4%
Q30 Bases in Barcode	95.8%
Q30 Bases in RNA Read	91.9%
Q30 Bases in Sample Index	89.8%
Q30 Bases in UMI	95.4%

Mapping

Reads Mapped to Genome	94.3%
Reads Mapped Confidently to Genome	88.4%
Reads Mapped Confidently to Intergenic Regions	6.8%
Reads Mapped Confidently to Intronic Regions	25.0%
Reads Mapped Confidently to Exonic Regions	56.7%
Reads Mapped Confidently to Transcriptome	53.2%
Reads Mapped Antisense to Gene	1.3%

Cells

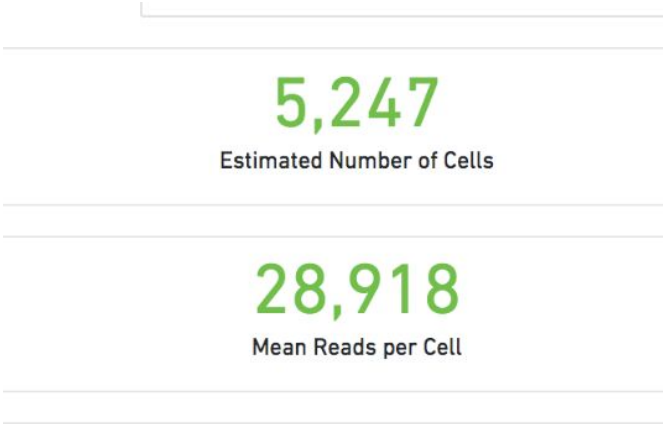
Barcode Rank Plot

Estimated Number of Cells	5,247
Fraction Reads in Cells	87.7%
Mean Reads per Cell	28,918
Median Genes per Cell	1,644
Total Genes Detected	20,822
Median UMI Counts per Cell	5,496

Sample

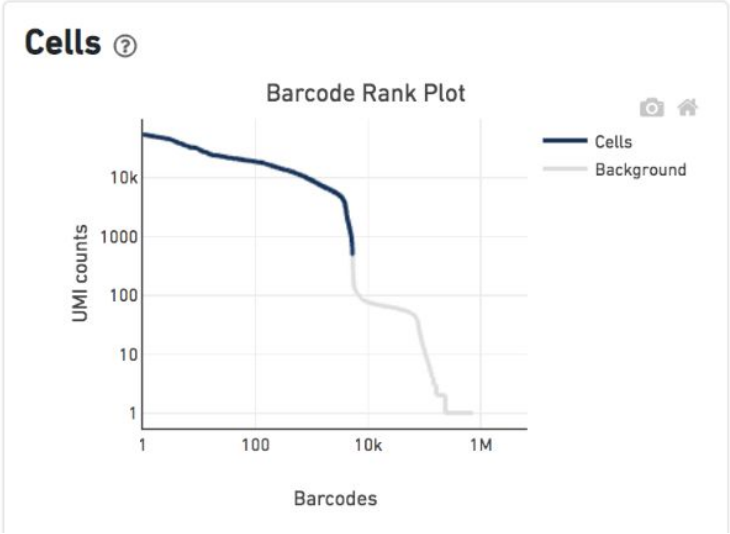
Sample ID	5k_pbmc_protein_v3
Sample Description	5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Chemistry	Single Cell 3' v3
Transcriptome	GRCh38-3.0.0
Pipeline Version	3.1.0

Single-Cell RNA-Seq - 10X Genomics



Single-Cell RNA-Seq - 10X Genomics

- Shows distribution of counts and which cells are associated with barcodes.
- We want to see a steep drop from the cells to the background.
- Indicates good separation.



Single-Cell RNA-Seq - 10X Genomics

Sequencing

Number of Reads	151,731,342
Valid Barcodes	97.5%
Valid UMIs	99.9%
Sequencing Saturation	52.4%
Q30 Bases in Barcode	95.8%
Q30 Bases in RNA Read	91.9%
Q30 Bases in Sample Index	89.8%
Q30 Bases in UMI	95.4%

Mapping

Reads Mapped to Genome	94.3%
Reads Mapped Confidently to Genome	88.4%
Reads Mapped Confidently to Intergenic Regions	6.8%
Reads Mapped Confidently to Intronic Regions	25.0%
Reads Mapped Confidently to Exonic Regions	56.7%
Reads Mapped Confidently to Transcriptome	53.2%
Reads Mapped Antisense to Gene	1.3%

The Process

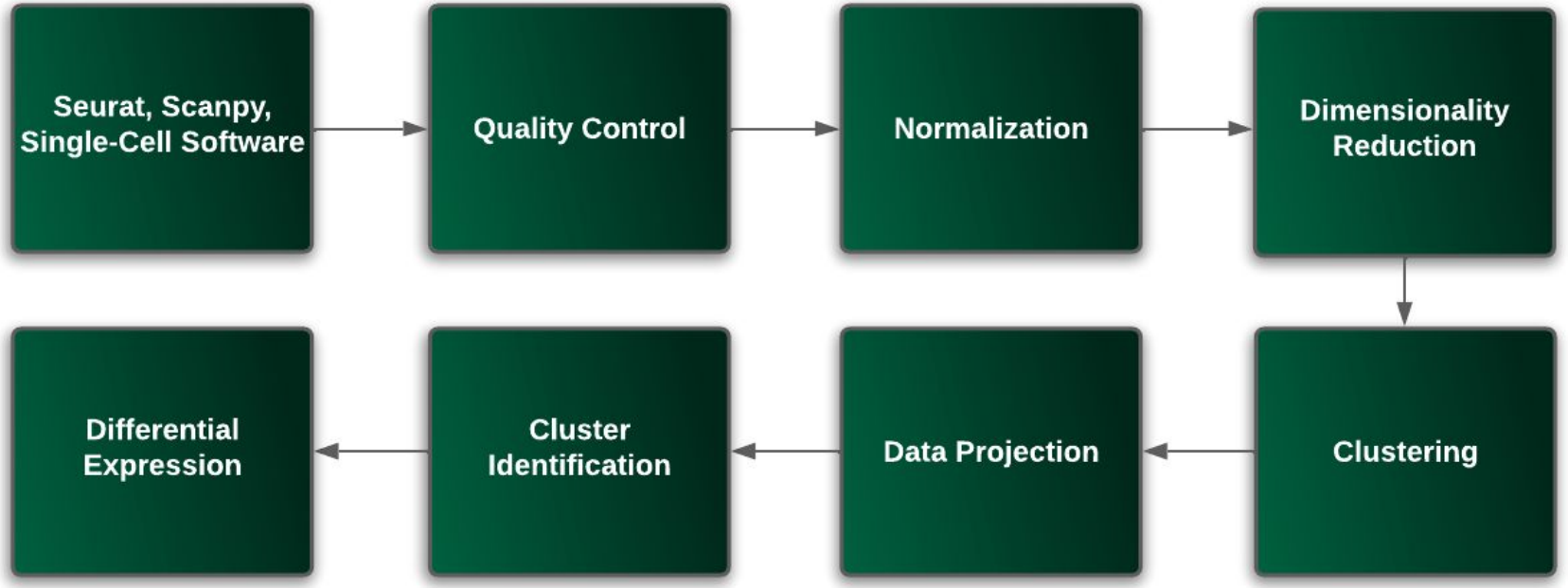
	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Gene 1	0	12	9	0	0	0	41	30
Gene 2	2	0	0	43	0	0	0	97
Gene 3	0	89	0	0	7	0	0	0
Gene 4	0	0	0	0	0	0	0	0
Gene 5	12	0	0	0	52	0	0	19
Gene 6	0	43	0	77	0	0	60	0
Gene 7	0	25	66	0	0	0	0	0

scRNA-Seq vs Bulk RNA-Seq Gene Matrix

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Gene 1	0	12	9	0	0	0	41	30
Gene 2	2	0	0	43	0	0	0	97
Gene 3	0	89	0	0	7	0	0	0
Gene 4	0	0	0	0	0	0	0	0
Gene 5	12	0	0	0	52	0	0	19
Gene 6	0	43	0	77	0	0	60	0
Gene 7	0	25	66	0	0	0	0	0

Gene	Total Expression
Gene 1	12
Gene 2	18
Gene 3	12
Gene 4	10
Gene 5	23
Gene 6	11

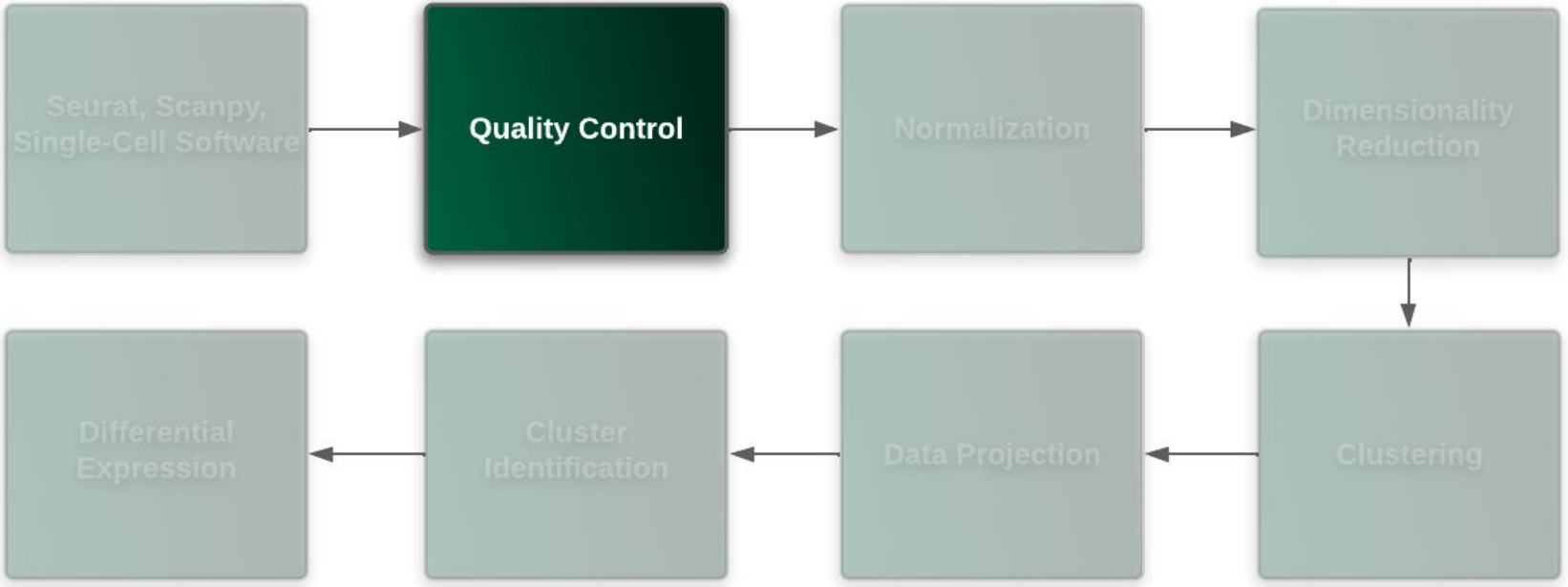
Roadmap



Tertiary Analysis

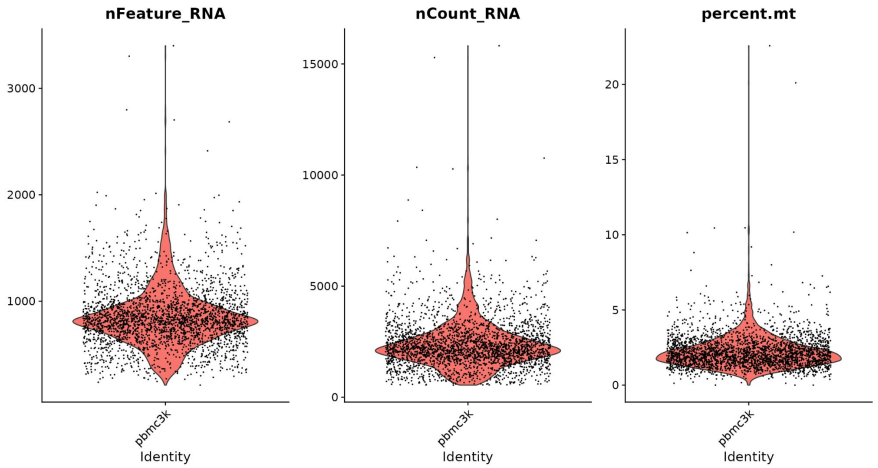
- **Seurat:** R based package for analyzing single-cell data.
- **Scanpy:** Python based package for analyzing single-cell data
- Both employ similar methods,
  - Slight differences in certain algorithms/methods that some their functions use.

Roadmap *(Quality Control)*



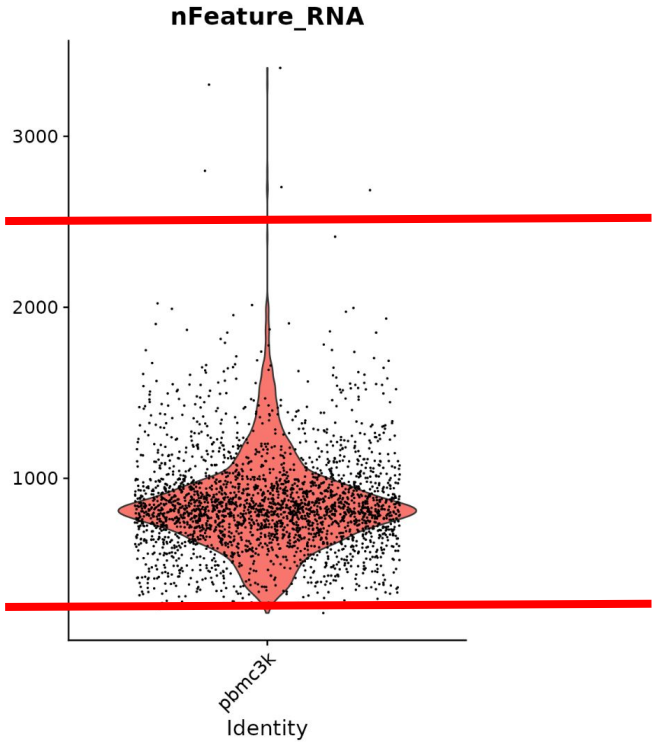
Roadmap *(Quality Control)*

- QC for single-cell is data-driven
- Variable depending on data-type



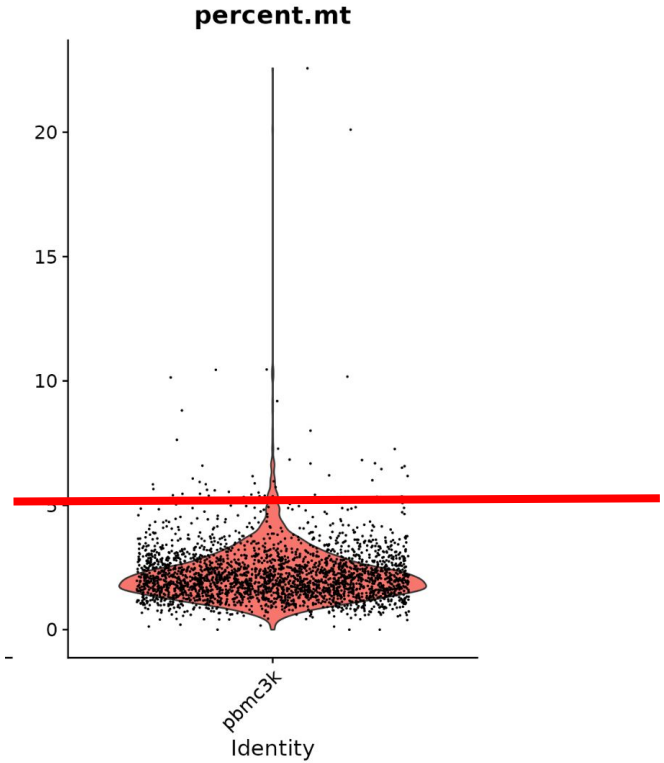
Roadmap *(Quality Control)*

- > 2500 features
- < 200 features

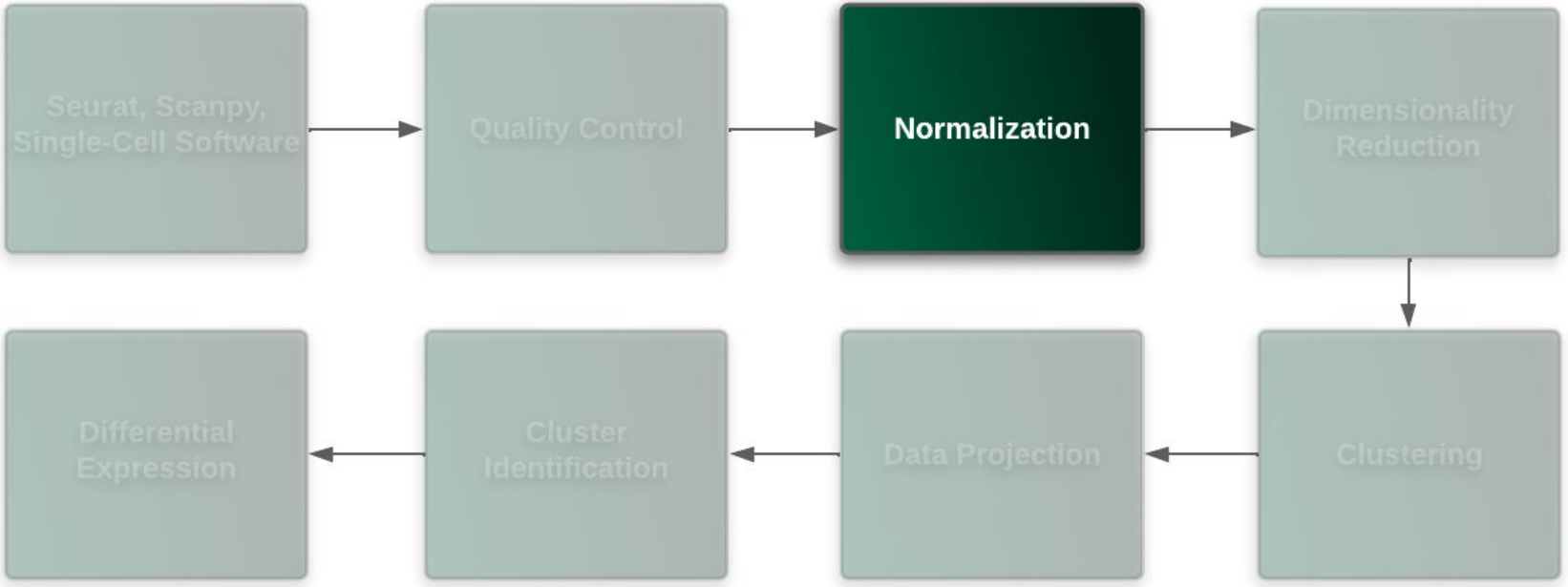


Roadmap *(Quality Control)*

- > 5% MT
- Again variable, cardiomyocytes will have higher percentage MT, etc
- Note: more QCs should be performed beyond what is shown here



Roadmap *(Normalization)*



Roadmap *(Normalization)*

- Inherent bias present.
- Sequencing depth, gene length, etc.
- Normalization removes this bias so that the differences we see from the cells are biological differences.

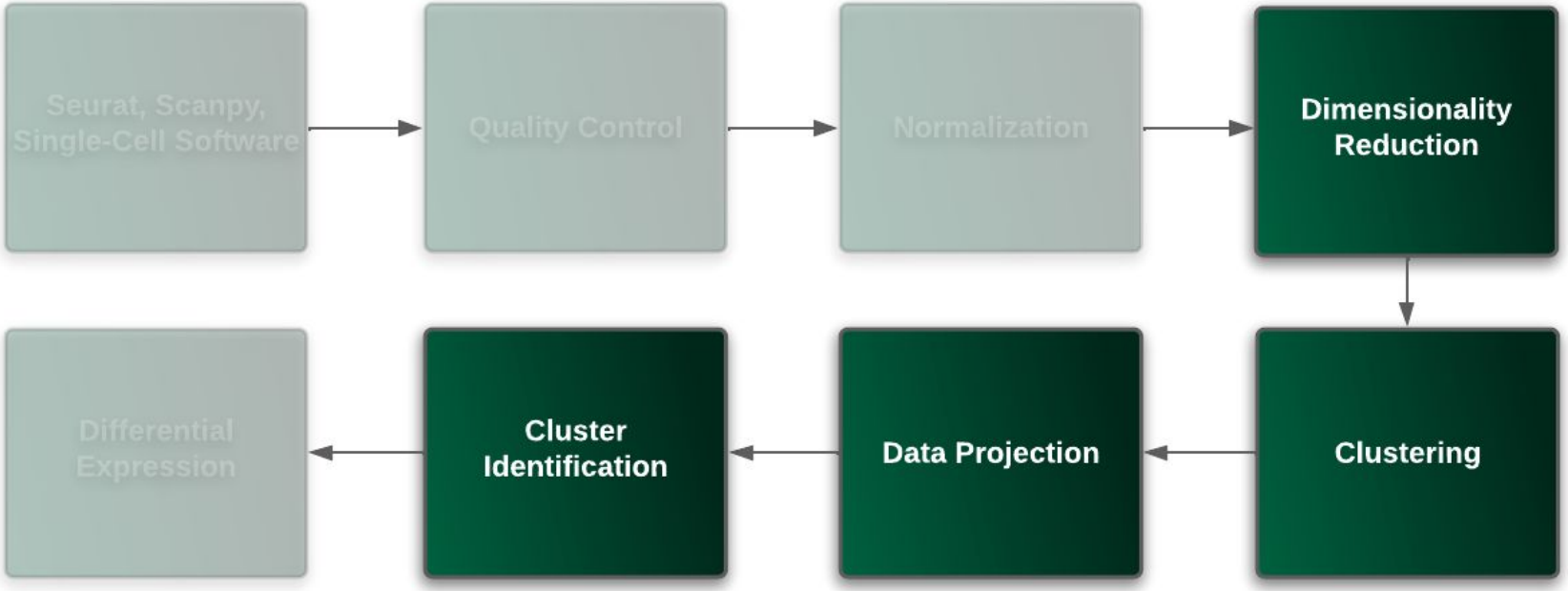
	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7	Cell 8
Gene 1	0	12	9	0	0	0	41	30
Gene 2	2	0	0	43	0	0	0	97
Gene 3	0	89	0	0	7	0	0	0
Gene 4	0	0	0	0	0	0	0	0
Gene 5	12	0	0	0	52	0	0	19
Gene 6	0	43	0	77	0	0	60	0
Gene 7	0	25	66	0	0	0	0	0

Roadmap *(Normalization)*

- **Log-Normalization**
  - Counts for each cell are divided by total cell counts, and multiplied by a scale factor. They are then log-transformed.

- **SCTransform (“V2 flavor”)**
  - Normalization Using a Regularized Negative Binomial Model
  - Most commonly used normalization method.

Roadmap *(Dimensionality Reduction, Cluster Identification)*



Roadmap *(Dimensionality Reduction, Cluster Identification)*

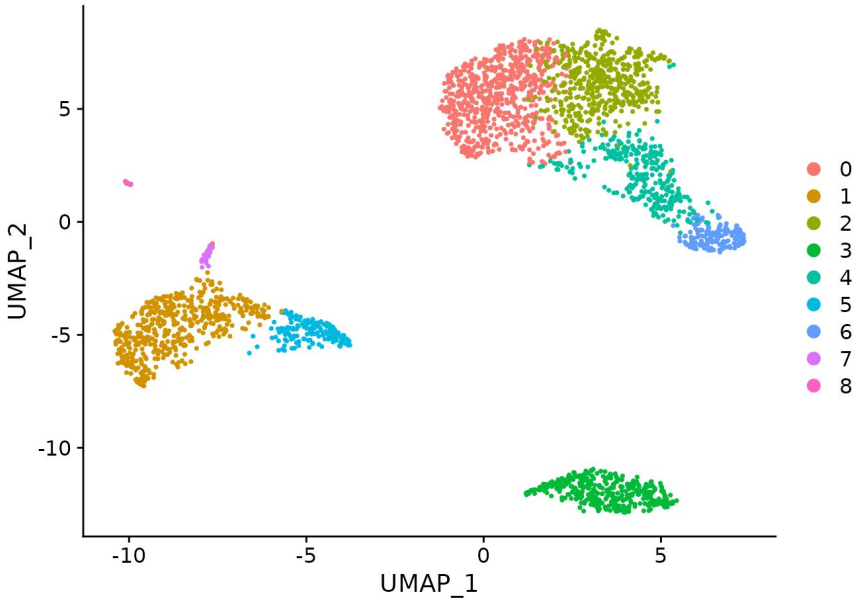
- **Dimensionality Reduction:** Capturing the core variability within the data-set.
- **Clustering:** Finding patterns with cells.
- **Data Projection/Cluster Identification:** Visualizing cell populations and identifying marker genes in each cluster that represent a cell-type.

RoadMap *(Data Projection - UMAP)*

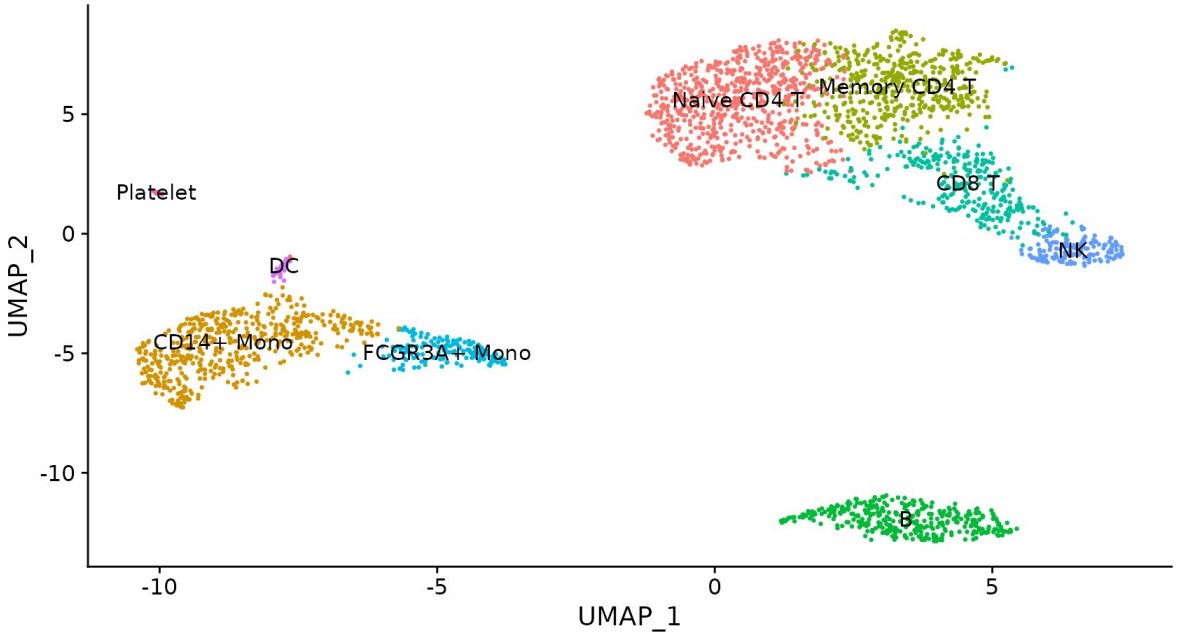
- **UMAP**
  - “data projection.”
  - Non-linear dimensionality reduction technique.
  - Visualizes high-dimensional data in low dimensional space.

- **TSNE**
  - Doesn’t preserve the data as well
  - Computational speed, etc..

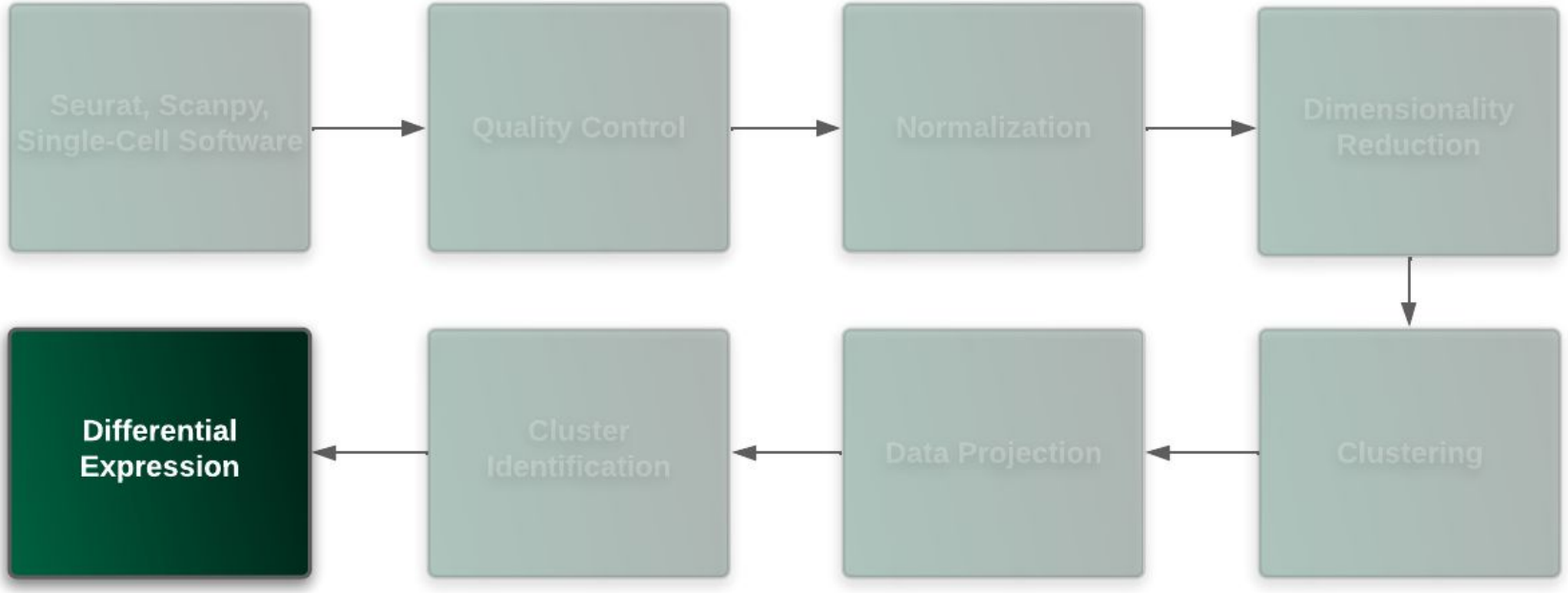
https://pair-code.github.io/understanding-umap/



Roadmap *(Cluster Identification)*



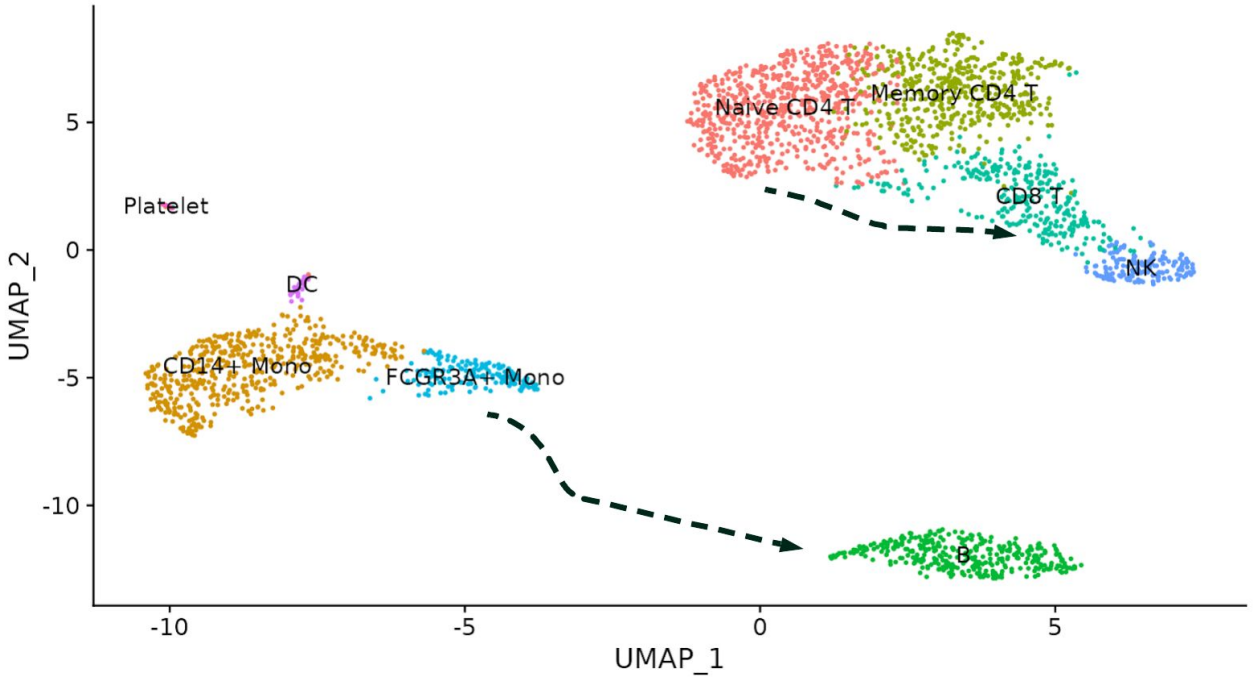
Roadmap *(Differential Expression)*



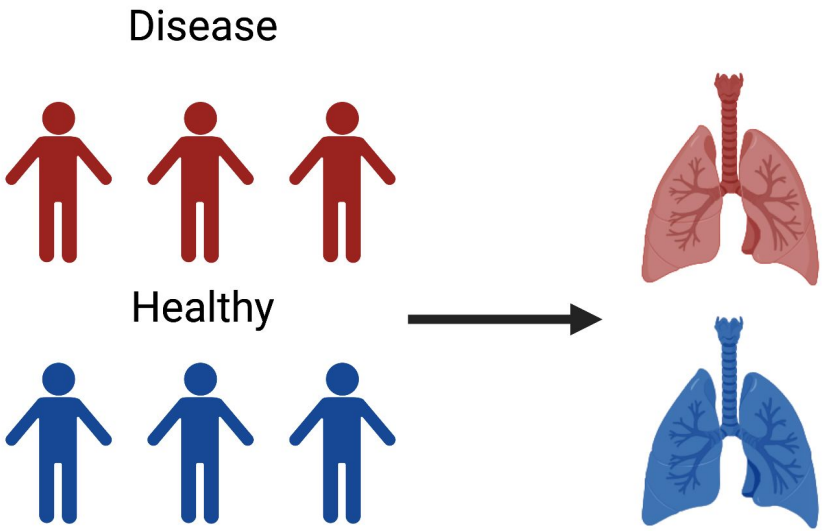
Roadmap *(Differential Expression)*

Differential expression in single-cell can proceed in two ways..

Cluster-Specific DE



Differential Expression Between Phenotypes/Conditions



Pseudobulk

- Aggregation of single-cell counts by cell type for comparison across conditions.
- Allows you to conduct phenotype comparisons (disease vs normal.)
- Useful for situations with lower counts.
- Able to assess biological variation to a greater extent.

Recap on scRNA-Seq

- scRNA-seq allows for novel identification of cell populations, a look into the cell-specific expression landscape, and ability to look at tissue-wide differences (pseudobulk.)
- Single-Cell RNA-Seq is ***evolving fast***.
- Multimodal approach - simultaneously measure transcriptomics and cell-surface protein, chromatin accessibility, etc.

Links To Sample Data

- 10X Genomics Data
- Seurat Data
- Seurat Data Package

U-BDS

Data Science Office Hours

Thursdays 3:30-4:00

Slack ID: C0118620WC8

Training Guides

https://github.com/U-BDS/training_guides