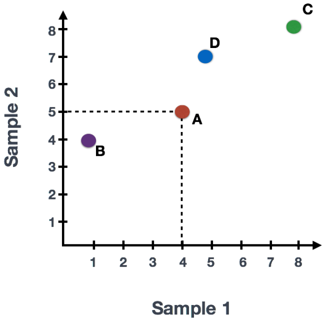
STAT588/BIOL588: Genomic Data Science

Lecture 22: Single-cell RNA-seq Normalization and Clustering

Yen-Yi Ho (hoyen@stat.sc.edu)

PCA

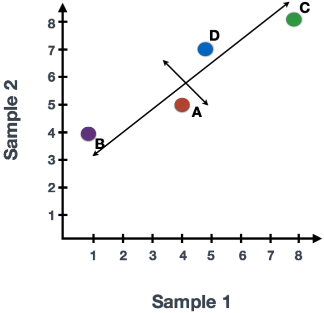
If we draw a line through the data in the direction representing the **most variation**, which is on the diagonal in this example. The maximum variation in the dataset is between the genes that make up the two endpoints of this line.



	Sample 1	Sample 2
Gene A	4	5
Gene B	1	4
Gene C	8	8
Gene D	5	7

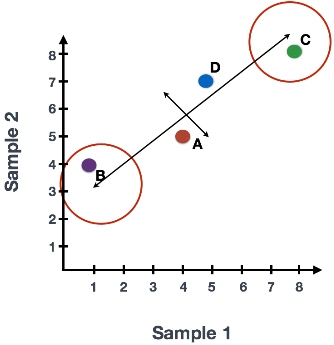
PCA

We also see the genes vary somewhat above and below the line. We could draw another line through the data representing **the second most amount of variation** in the data, since this plot is in 2D (2 axes).

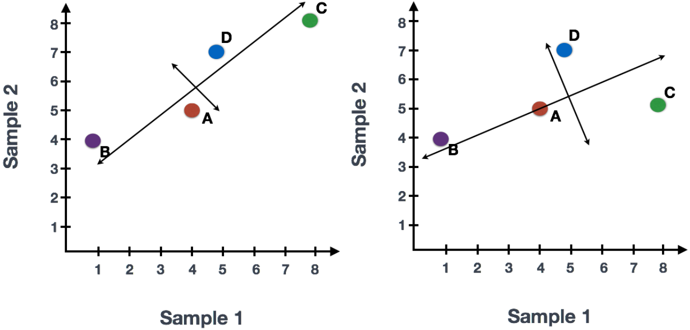


PCA

The genes near the ends of each line would be those with the highest variation; these genes have the **greatest influence** on the direction of the line, mathematically.

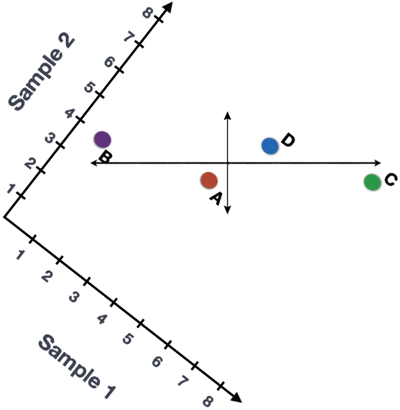


PCA


Sample 2
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
Sample 1
A
B
C
D
Sample 2
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
Sample 1
A
B
C
D

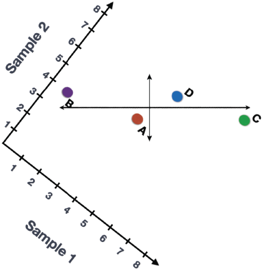
Rotation

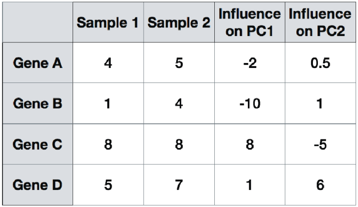
These axes are essentially the "Principal Components", with PC1 representing the most variation in the data and PC2 representing the second most variation in the data.



PC scores

PC scores are calculated for all sample-PC pairs as described in the steps and schematic below:





	Sample 1	Sample 2	Influence on PC1	Influence on PC2
Gene A	4	5	-2	0.5
Gene B	1	4	-10	1
Gene C	8	8	8	-5
Gene D	5	7	1	6

PC scores

(2) Once the influence has been determined, the score for each sample is calculated using the following equation:

Sample1 PC1 score = (read count * influence) + . . . for all genes

For our 2-sample example, the following is how the scores would be calculated:

```
## Sample1
PC1 score = (4 * -2) + (1 * -10) + (8 * 8) + (5 * 1) = 51
PC2 score = (4 * 0.5) + (1 * 1) + (8 * -5) + (5 * 6) = -7

## Sample2
PC1 score = (5 * -2) + (4 * -10) + (8 * 8) + (7 * 1) = 21
PC2 score = (5 * 0.5) + (4 * 1) + (8 * -5) + (7 * 6) = 8.5
```

PC scores

Here is a schematic that goes over the first 2 steps:

Gene expression in each Cell/Sample

$$\begin{array}{c|cccc} & \text{Gene 1} & \text{Gene 2} & \text{Gene 3} & \text{Gene 4} \\ \hline \text{Cell 1} & 4 & 1 & 8 & 5 \\ \text{Cell 2} & 5 & 4 & 8 & 7 \end{array}$$

Dim 2 x 4

×

Every gene's influence on the PC

$$\begin{array}{c|cc} & \text{PC1} & \text{PC2} \\ \hline \text{Gene 1} & -2 & 0.5 \\ \text{Gene 2} & -10 & 1 \\ \text{Gene 3} & 8 & -5 \\ \text{Gene 4} & 1 & 6 \end{array}$$

Dim 4 x 2

# of PCs = # of cells!

=

PC score

$$\begin{array}{c|cc} & \text{PC1} & \text{PC2} \\ \hline \text{Cell 1} & 4\times(-2) + 1\times(-10) + 8\times8 + 5\times1 & 4\times0.5 + 1\times1 + 8\times(-5) + 5\times6 \\ \text{Cell 2} & 5\times(-2) + 4\times(-10) + 8\times(-5) + 5\times6 & 5\times0.5 + 4\times1 + 8\times(-5) + 7\times6 \end{array}$$

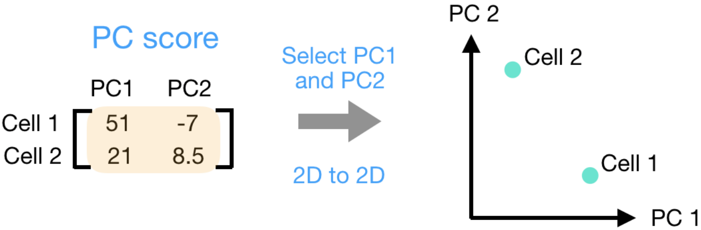
OR

$$\begin{array}{c|cc} & \text{PC1} & \text{PC2} \\ \hline \text{Cell 1} & 51 & -7 \\ \text{Cell 2} & 21 & 8.5 \end{array}$$

Dim 2 x 2

2D plot

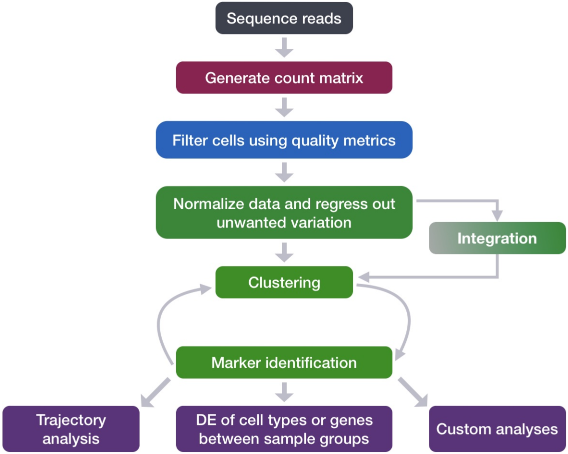
Once these scores are calculated for all the PCs, they can be plotted on a simple scatter plot. Below is the plot for the example here, going from the 2D matrix to a 2D plot:



PC score

$$\begin{array}{ccc} & \text{PC1} & \text{PC2} \\ \text{Cell 1} & 51 & -7 \\ \text{Cell 2} & 21 & 8.5 \end{array}$$

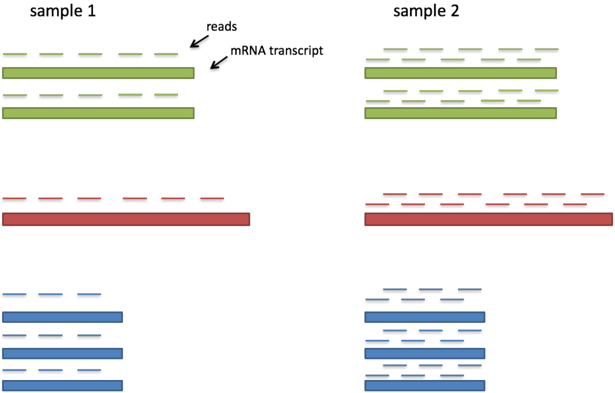
scRNAseq Workflow


Sequence reads
Generate count matrix
Filter cells using quality metrics
Normalize data and regress out unwanted variation
Integration
Clustering
Marker identification
Trajectory analysis
DE of cell types or genes between sample groups
Custom analyses

scRNAseq Normalization

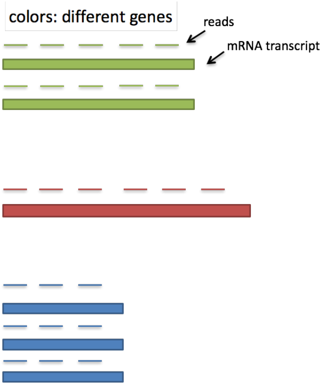
- Sequencing Depth
- Gene Length
- Mitochondria Ratio

Sequencing Depth



MI Love: RNA-seq statistical analysis

Gene Length

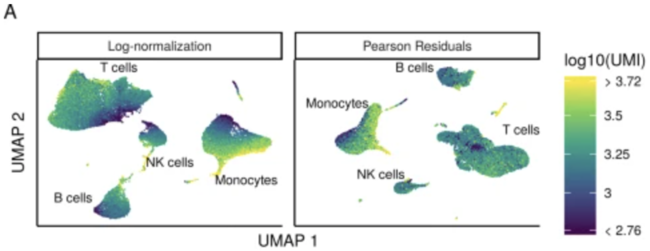


Slide adapted from MI Love: RNA-seq statistical analysis

Normalization

SCTransform method constructs a generalized linear model (GLM) for each gene with UMI counts as the response and sequencing depth as the explanatory variable.

Fig. 6



PCA

Consider a single-cell RNA-seq dataset with 12,000 cells and you have quantified the expression of 20,000 genes.

Gene expression in each Cell/Sample

$$\begin{array}{c|cccc|} & \text{Gene 1} & \text{Gene 2} & \dots & \text{Gene 20,000} \\ \text{Cell 1} & & & & \\ \text{Cell 2} & & \dots & & \\ \dots & & & & \\ \text{Cell 12,000} & & & & \end{array}$$

Dim 12,000 x 20,000

×

Every gene's influence on the PC

$$\begin{array}{c|cccc|} & \text{PC1} & \text{PC2} & \dots & \text{PC12,000} \\ \text{Gene 1} & & & & \\ \text{Gene 2} & & & & \\ & & \dots & & \\ \dots & & & & \\ \text{Gene 20,000} & & & & \end{array}$$

Dim 20,000 x 12,000

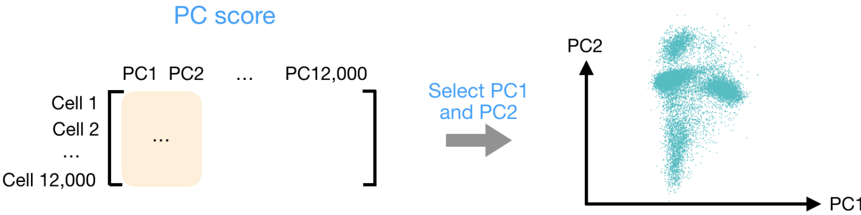
# of PCs = # of cells!

=

Cell

Clustering

After the PC scores have been calculated, you are looking at a matrix of 12,000 x 12,000 that represents the information about relative gene expression in all the cells. You can select the PC1 and PC2 columns and plot that in a 2D way.



PC score

PC1 PC2 … PC12,000

Cell 1

Cell 2

…

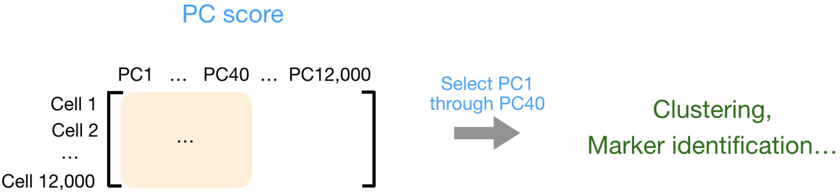
Cell 12,000

…

Select PC1 and PC2

PCs

Use PC scores from the first 40 PCs for downstream analysis like clustering, marker identification etc., since these represent the majority of the variation in the data. We will be talking a lot more about this later in this workshop.



PC score

PC1 ... PC40 ... PC12,000

Cell 1

Cell 2

...

Cell 12,000

...

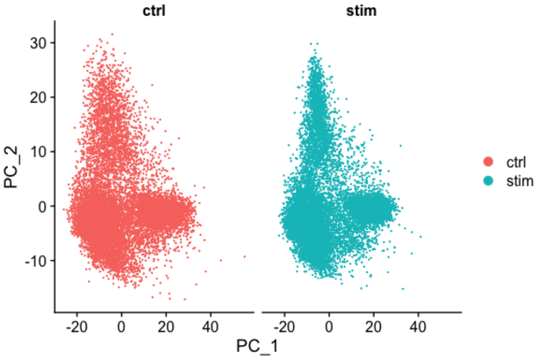
Select PC1 through PC40

Clustering,

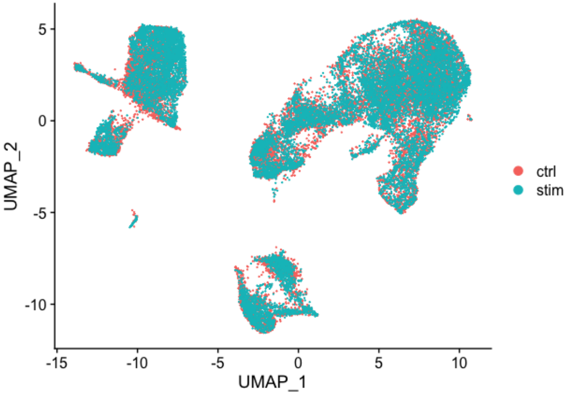
Marker identification…

UMAP

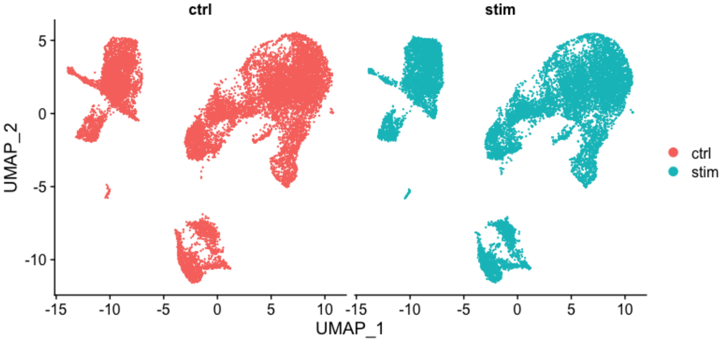
A popular dimensionality reduction technique for scRNAseq is called **Uniform Manifold Approximation and Projection (UMAP)**. UMAP takes the information from any number of top PCs to arrange the cells in this multidimensional space. It will take those distances in multidimensional space and plot them in two dimensions working to preserve local and global structure.



UMAP



Side-by-side comparison of clusters


ctrl
stim
UMAP_2
UMAP_1
● ctrl
● stim