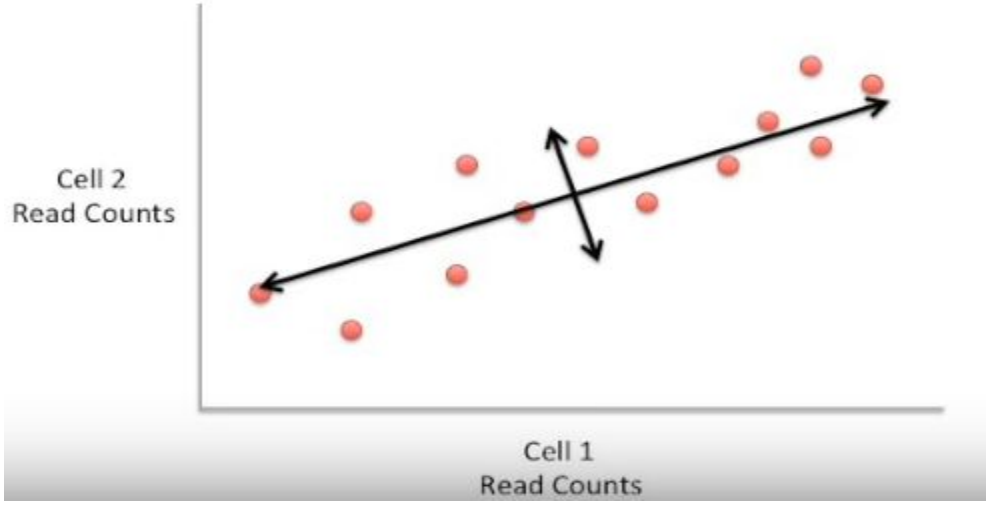
PCA

1. Consider the following 2d cell readings:

Gene	Cell 1	Cell 2
A	10	8
B	0	2
C	14	10
….	...	...

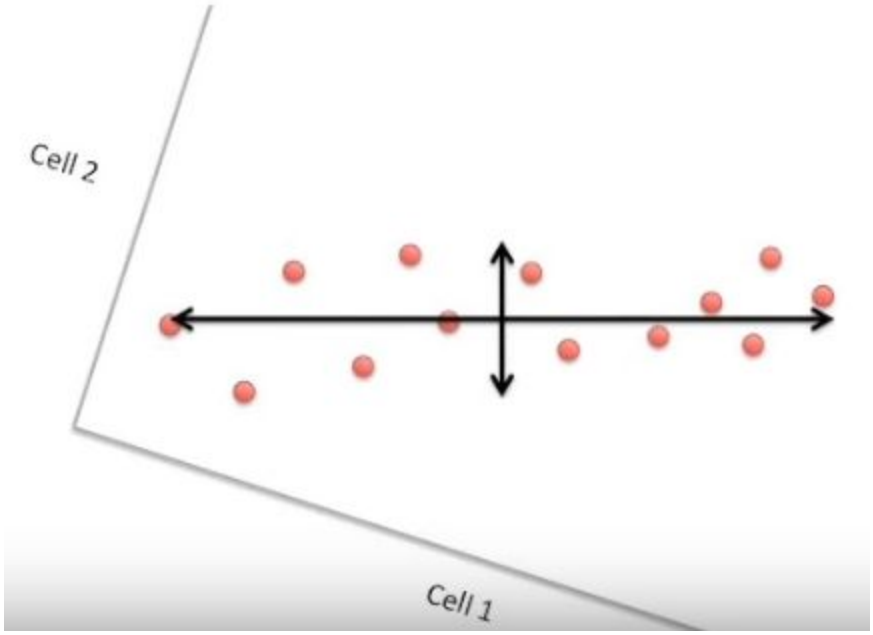
Plot theses values in a 2d graph (X axis-> cell 1 readings and Y axis->cell 2 readings).

2. Plot a diagonal through these values:



Rotate the x and y axes in such a way that the newly drawn line becomes

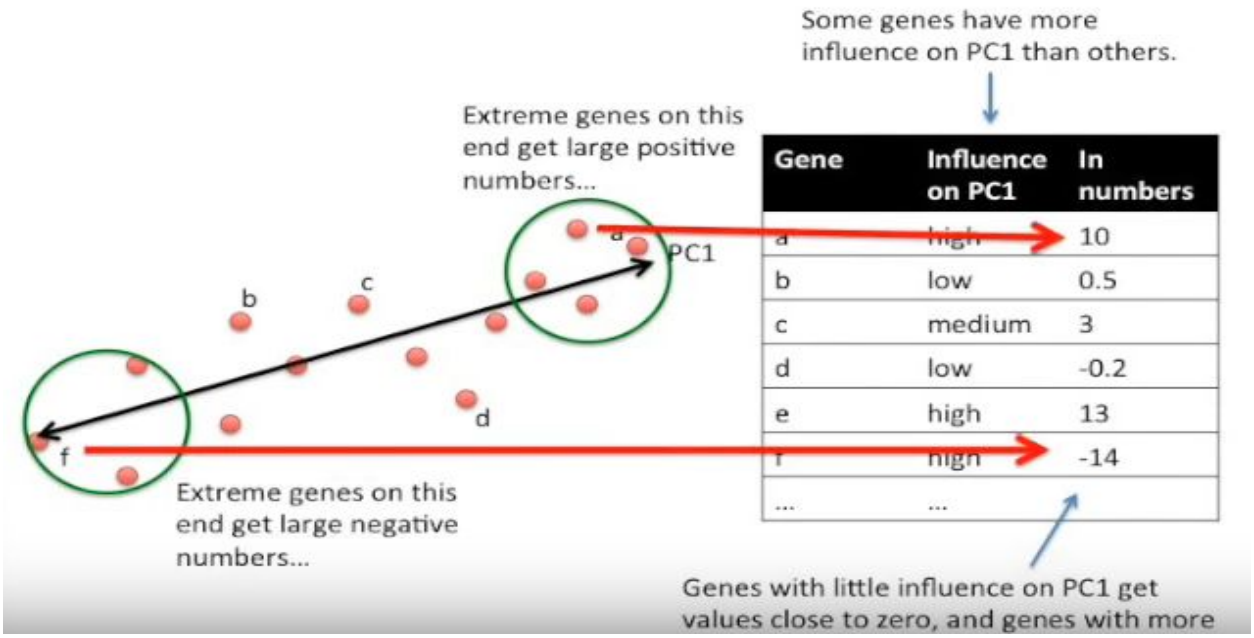
the new axis:



The axis in which most variations in cell reading happen is PC1 (Principal Component 1) and the second most variations occur is PC2 and it continues…

3. Calculate how much each gene is influenced by Principal Component axis:

- The genes that are near to the endpoints of a PC will have higher influence value compared to those near to the midpoint.
- The genes that are on different sides of mid point of a PC will differ in sign.



Influence numbers are weights given to each gene according to their influence on PC.

In PCA terminology, these weights are called “loadings” and array of loadings for a PC is called “eigenvector”.

4. Plot each cell reading in PC1 vs PC2 axes graph:

Combine the read count and influence value for all genes in a cell to get a single value:

In the above example,

For Cell1,

```
PC1 score = sigma (readCount of gene i * influence
                   value on PC1 by gene i)
```

Similarly Cell1, PC2 score is calculated.

This value is plotted in PC1 vs PC2 graph.

5. By plotting each cell in principal component axis:

- Genes with largest variation between cells will have most influence on principal component axis.
- Cells with similar transcription pattern will cluster together.