PCA clearly explained (2015)

* Cells with similar transcription should cluster
* PCA: a method for compressing a lot of data into something that captures the essence of the original data
* On a 2D graph, axis stands for two cells,
  + if correlated, meaning genes that are highly transcribed in cell 1 are also highly transcribed in cell 2
  + suppose we have a horizontal straight line, cell 2 genes are all transcribed at the same level, then we can just compress the graph into 1D without too much info loss since both graphs say “the important variation is left to right”
* Each cell we sequence adds another dimension, with some dimensions more important than others
* PCA takes a dataset with a lot of dimensions (i.e. lots of cells) and flattens it to 2 or 3 dimensions
  + Focusing on the things that are different between cells
* 2 cells
  + Spread out along a diagonal line, rotate to get a new pair of axes
  + Data varies a lot left and right, a little up and down
  + Rotated axes: PC1 (captures the direction where most of the variation is) and PC2 (captures the direction with the 2nd most variation)
* X and Y axis
  + PC1: the direction of the most variation in gene expression
  + PC2: the 2nd most variation in gene expression
* Plot cells instead of genes
  + Score genes based on how much they influence PC1, PC2
    - From high to low: the endpoints gene would have high scores because they highly influenced PC1, PC2
    - Rank them numerically
    - Influence in numbers are called loadings and an array of loadings called eigenvector
  + Cell1 PC1 score = (read count\*influence) + … for all genes
* Genes with the largest variation between cells will have the most influence on the principal components
  + Genes highly expressed in some cells and not expressed in others will have a lot of variation and influence on the PCs
* Cells with similar transcription patters will cluster together

StatQuest: A gentle introduction to RNA-seq

* Normal neural cell vs mutated neural cell
  + Want to know what genetic mechanism is causing the difference
  + Look at differences in gene expression
* Each cell has chromosomes, each chromosome has genes, some genes active
  + Wavy lines represent mRNA trancripts
* High throughput (?) sequencing tells us which genes are active, and how much they are transcribed.
* We can use RNA-seq to measure gene expression in normal cells, then use it to measure gene expression in mutated cells
  + Then compare the two cell types and figure out what’s different in the mutated cells.
* 3 main steps for RNA-seq
  + Prepare a sequencing library
    - Isolate the RNA
    - Break the RNA into small fragments
    - Convert the RNA fragments into double stranded DNA
    - Add sequencing adaptors
    - PCR amplify
    - Quality control
  + Sequence
  + Data analysis
    - Plot (PCA)
    - Identify differentially expressed genes between the normal and mutant samples
* Raw data
  + Filter out garbage reads
    - Low quality
    - Artifacts
  + Align the high-quality reads to a genome
  + Count the number of reads per gene
    - Normalize data in the big matrix