BINF 5003: Data Mining, Modeling, and Biostatistics

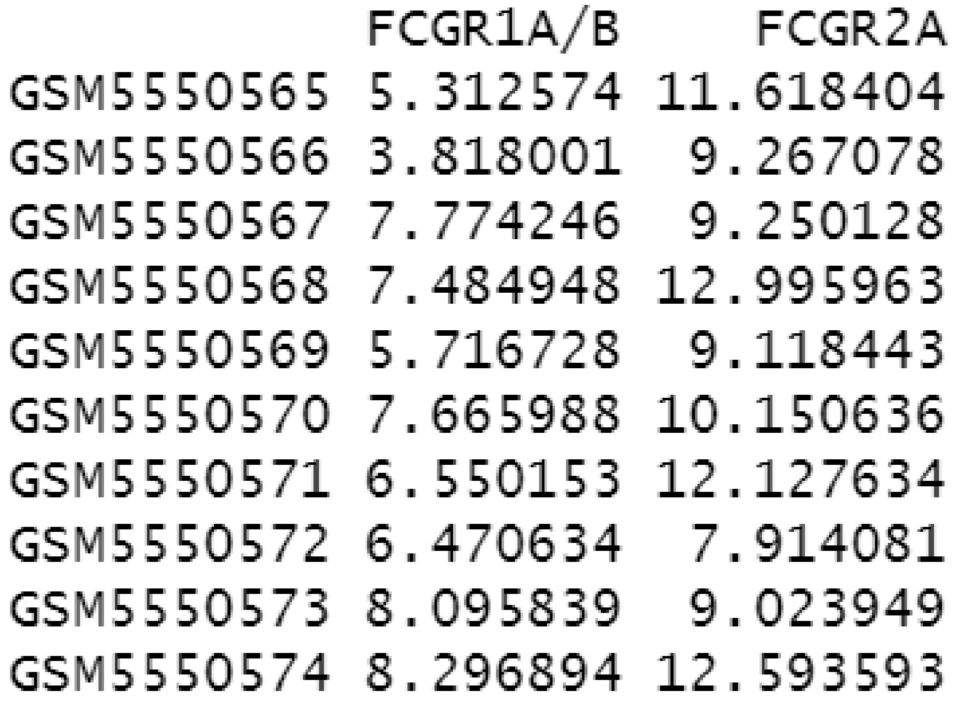
Week 9

Module 6 – Dimensionality Reduction

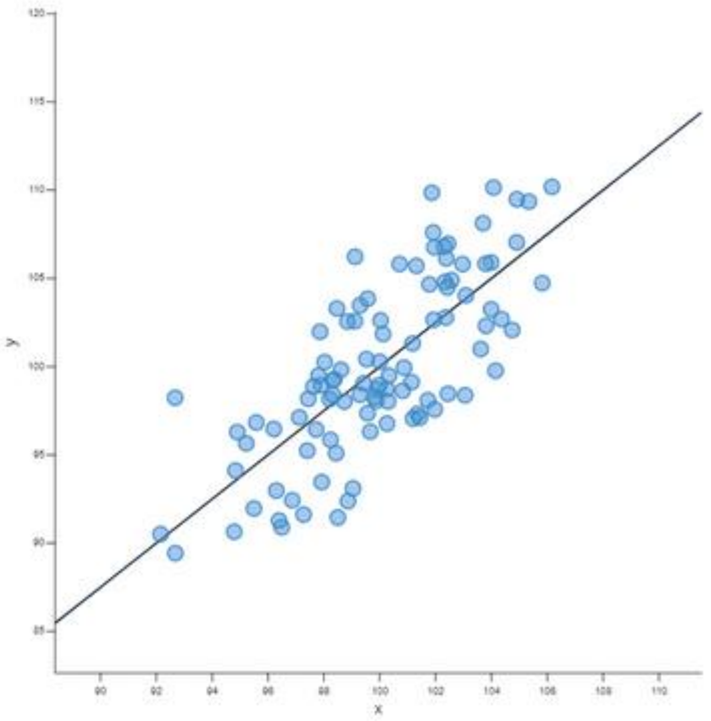
Overview

- Introduce high dimensional data
- Understanding each principal (not principle) component
- Different strategies for dimensionality reduction

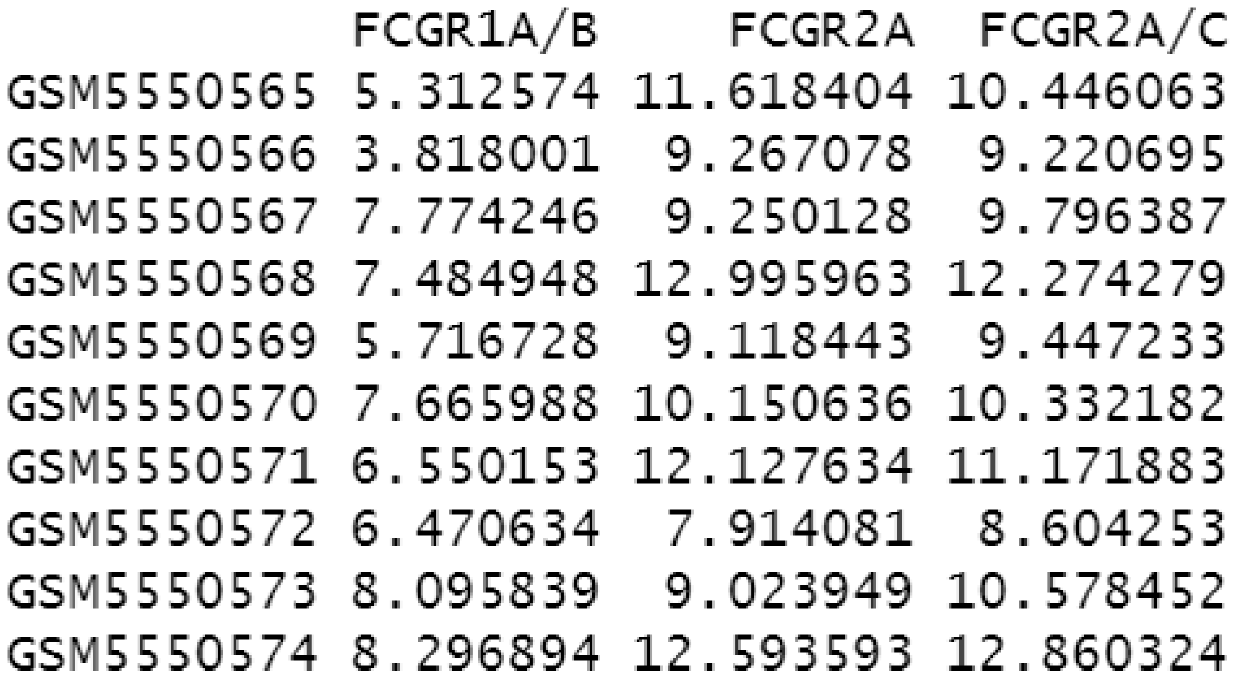
How do you compare the expression between 2 genes?



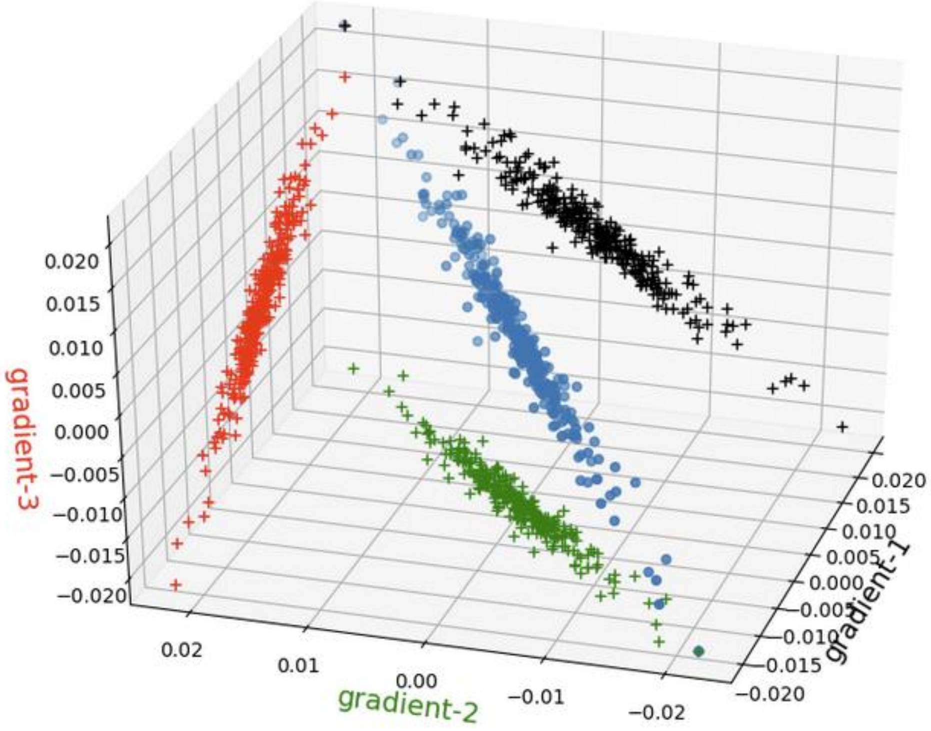
```
            FCGR1A/B    FCGR2A
GSM5550565 5.312574 11.618404
GSM5550566 3.818001  9.267078
GSM5550567 7.774246  9.250128
GSM5550568 7.484948 12.995963
GSM5550569 5.716728  9.118443
GSM5550570 7.665988 10.150636
GSM5550571 6.550153 12.127634
GSM5550572 6.470634  7.914081
GSM5550573 8.095839  9.023949
GSM5550574 8.296894 12.593593
```



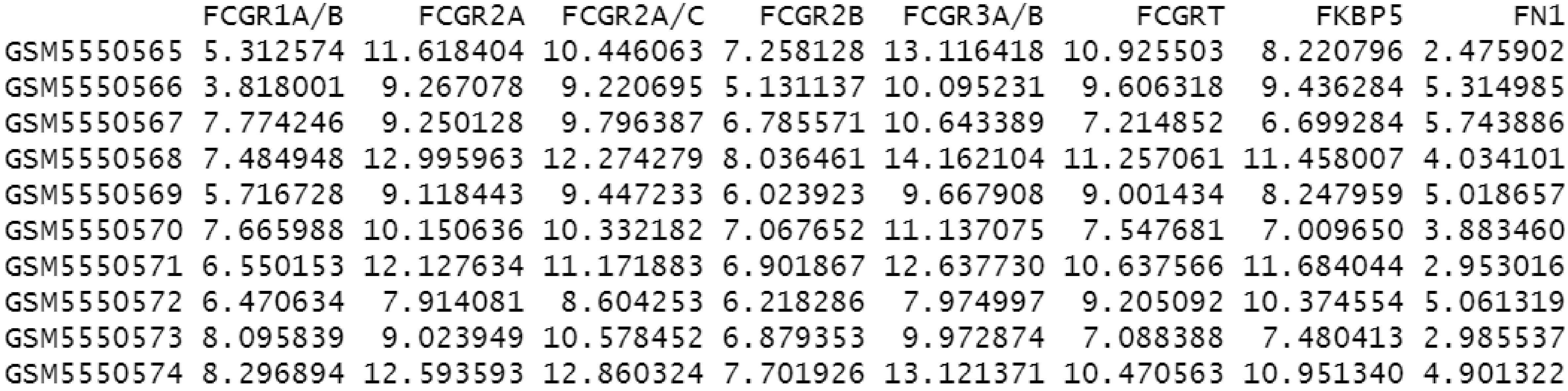
How do you compare the expression between 3 genes?



```
           FCGR1A/B    FCGR2A  FCGR2A/C
GSM5550565 5.312574 11.618404 10.446063
GSM5550566 3.818001  9.267078  9.220695
GSM5550567 7.774246  9.250128  9.796387
GSM5550568 7.484948 12.995963 12.274279
GSM5550569 5.716728  9.118443  9.447233
GSM5550570 7.665988 10.150636 10.332182
GSM5550571 6.550153 12.127634 11.171883
GSM5550572 6.470634  7.914081  8.604253
GSM5550573 8.095839  9.023949 10.578452
GSM5550574 8.296894 12.593593 12.860324
```



How do you compare the expression between 30 genes?



```
            FCGR1A/B    FCGR2A  FCGR2A/C    FCGR2B  FCGR3A/B     FCGRT     FKBP5      FN1
GSM5550565 5.312574 11.618404 10.446063 7.258128 13.116418 10.925503  8.220796 2.475902
GSM5550566 3.818001  9.267078  9.220695 5.131137 10.095231  9.606318  9.436284 5.314985
GSM5550567 7.774246  9.250128  9.796387 6.785571 10.643389  7.214852  6.699284 5.743886
GSM5550568 7.484948 12.995963 12.274279 8.036461 14.162104 11.257061 11.458007 4.034101
GSM5550569 5.716728  9.118443  9.447233 6.023923  9.667908  9.001434  8.247959 5.018657
GSM5550570 7.665988 10.150636 10.332182 7.067652 11.137075  7.547681  7.009650 3.883460
GSM5550571 6.550153 12.127634 11.171883 6.901867 12.637730 10.637566 11.684044 2.953016
GSM5550572 6.470634  7.914081  8.604253 6.218286  7.974997  9.205092 10.374554 5.061319
GSM5550573 8.095839  9.023949 10.578452 6.879353  9.972874  7.088388  7.480413 2.985537
GSM5550574 8.296894 12.593593 12.860324 7.701926 13.121371 10.470563 10.951340 4.901322
```

Dimensionality Reduction

- Each feature (gene) or sample in the dataset adds dimensions to the analysis

- Minimize the dimensions
    - Prevent overfitting
    - Keep the dataset computationally manageable

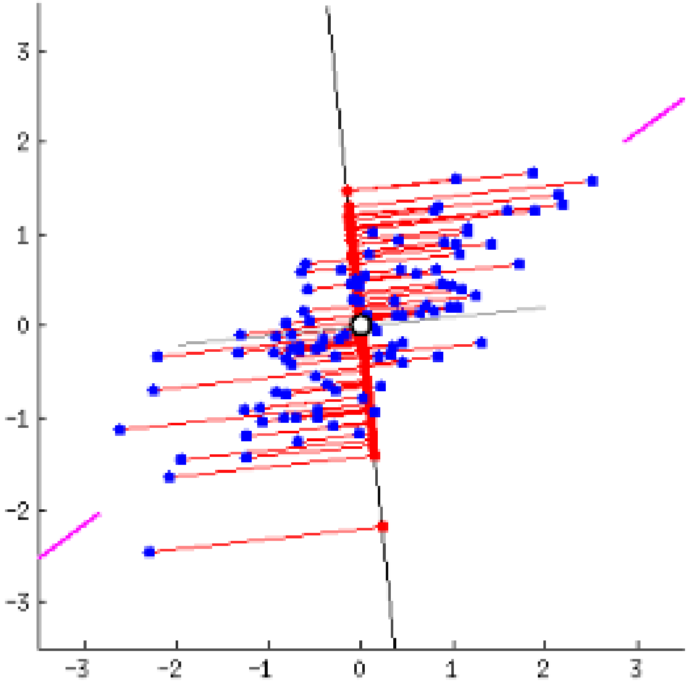
- 2 strategies
    - Eliminate dimensions by removing redundancies or static features
    - Collapse the number of components we are working with

Principal Component Analysis

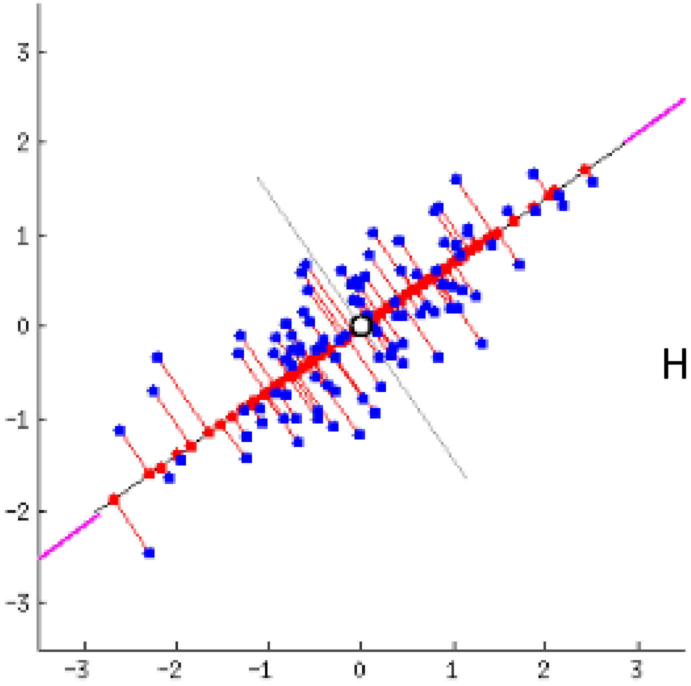
- Original basis of multivariate data analysis
  - Reduce the number of features to more manageable and meaningful numbers
- Draws lines through the data to best account for the variance in the data
  - Difference between observed and expected values on the plane – concept sound familiar??

Principal Component Analysis

- PC determined by maximizing the variance it explains



Low variance

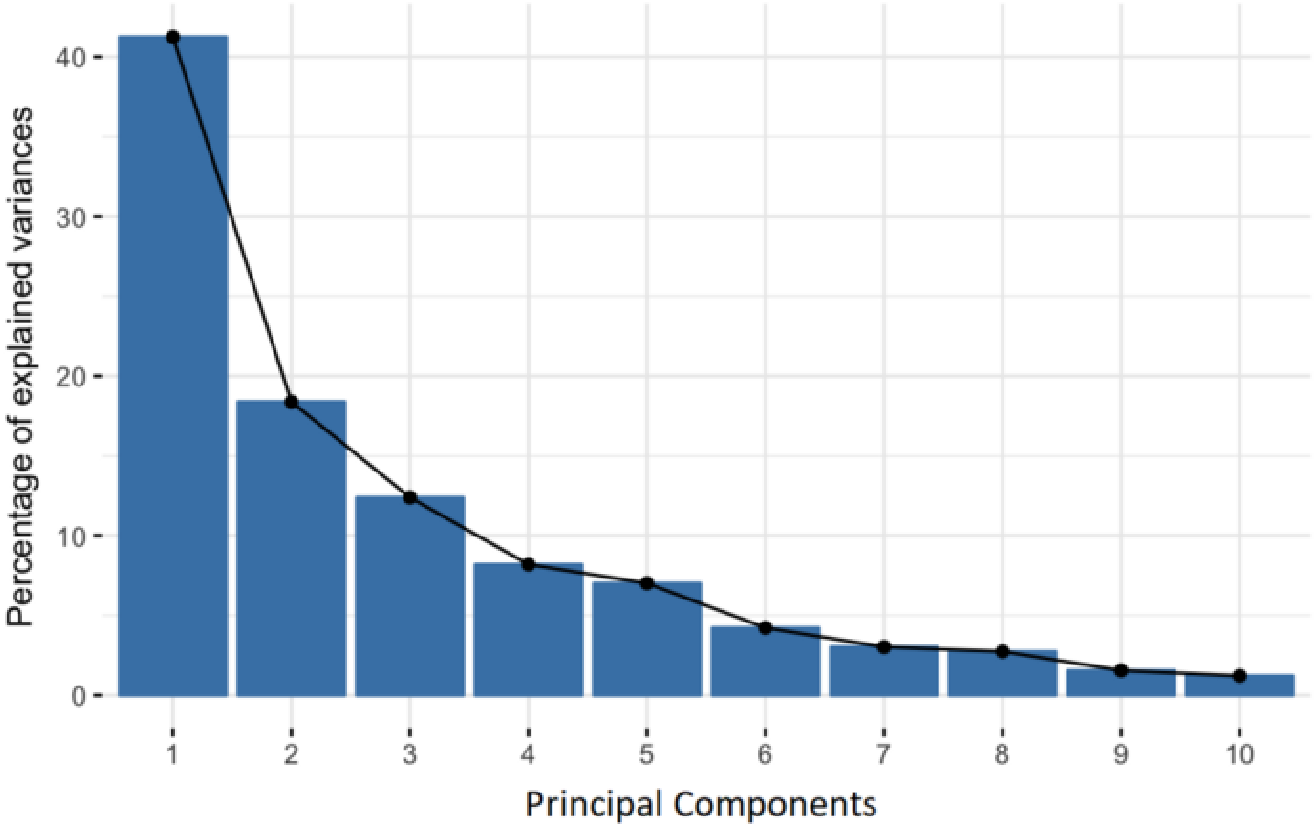


High variance

https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal Component Analysis

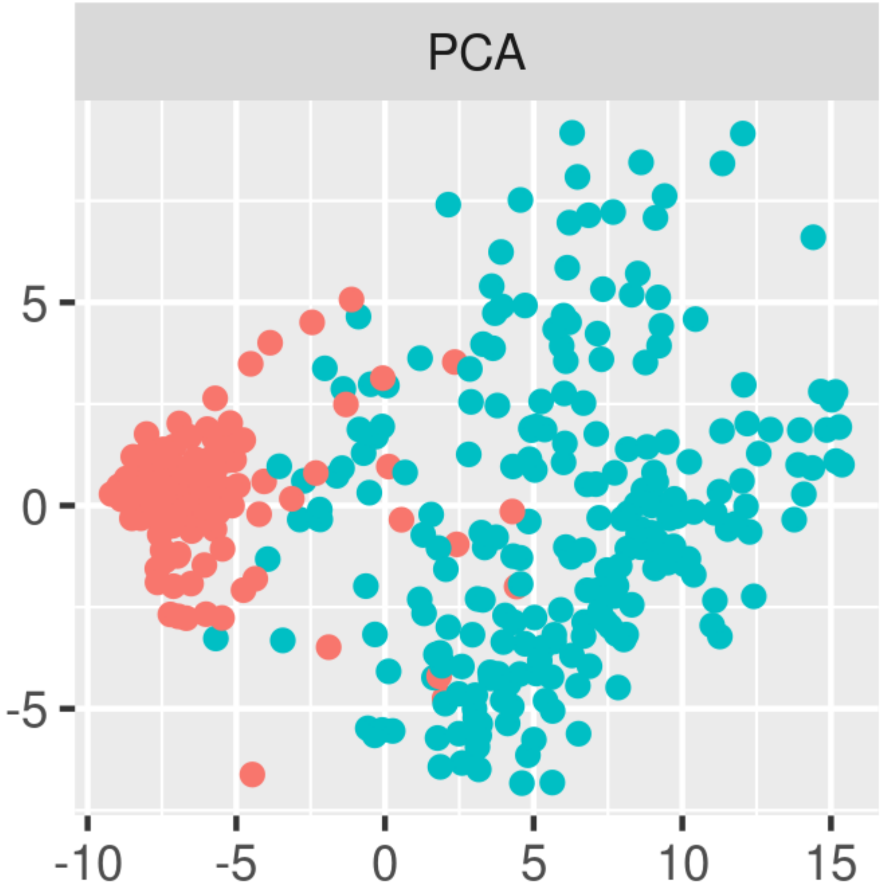
- There are as many principal components as there are features
  - Each feature contributes to each component
  - Ordered from highest to lowest percentage of variance explained



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

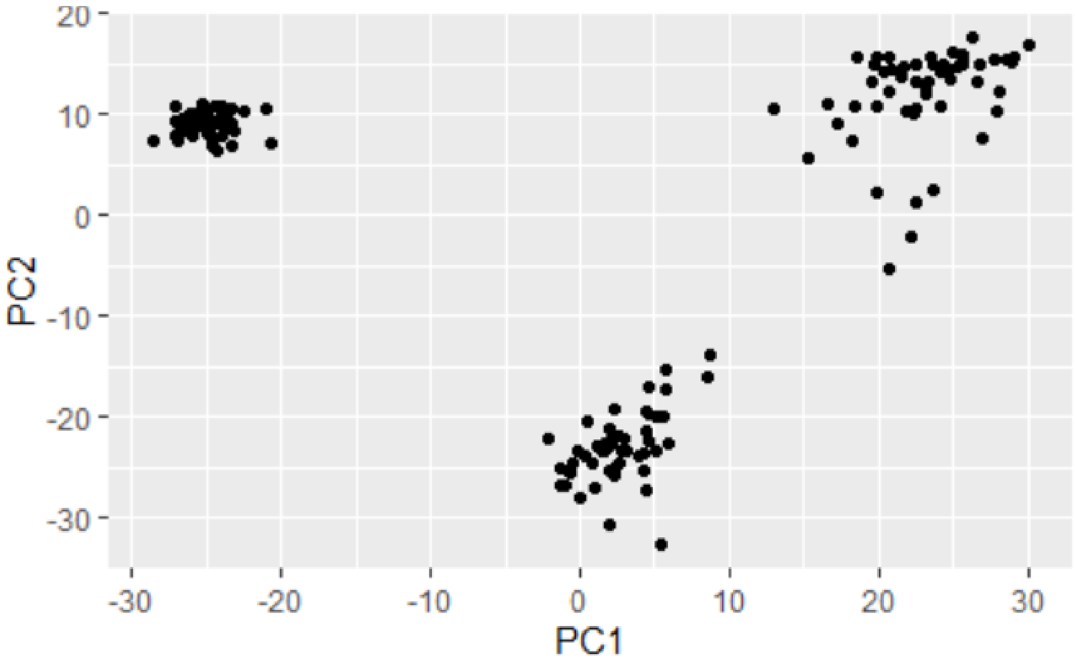
Principal Component Analysis

- Each sample is a dot
- Dots that are more similar are closer together



https://builtin.com/data-science/step-step-explanation-principal-component-analysis

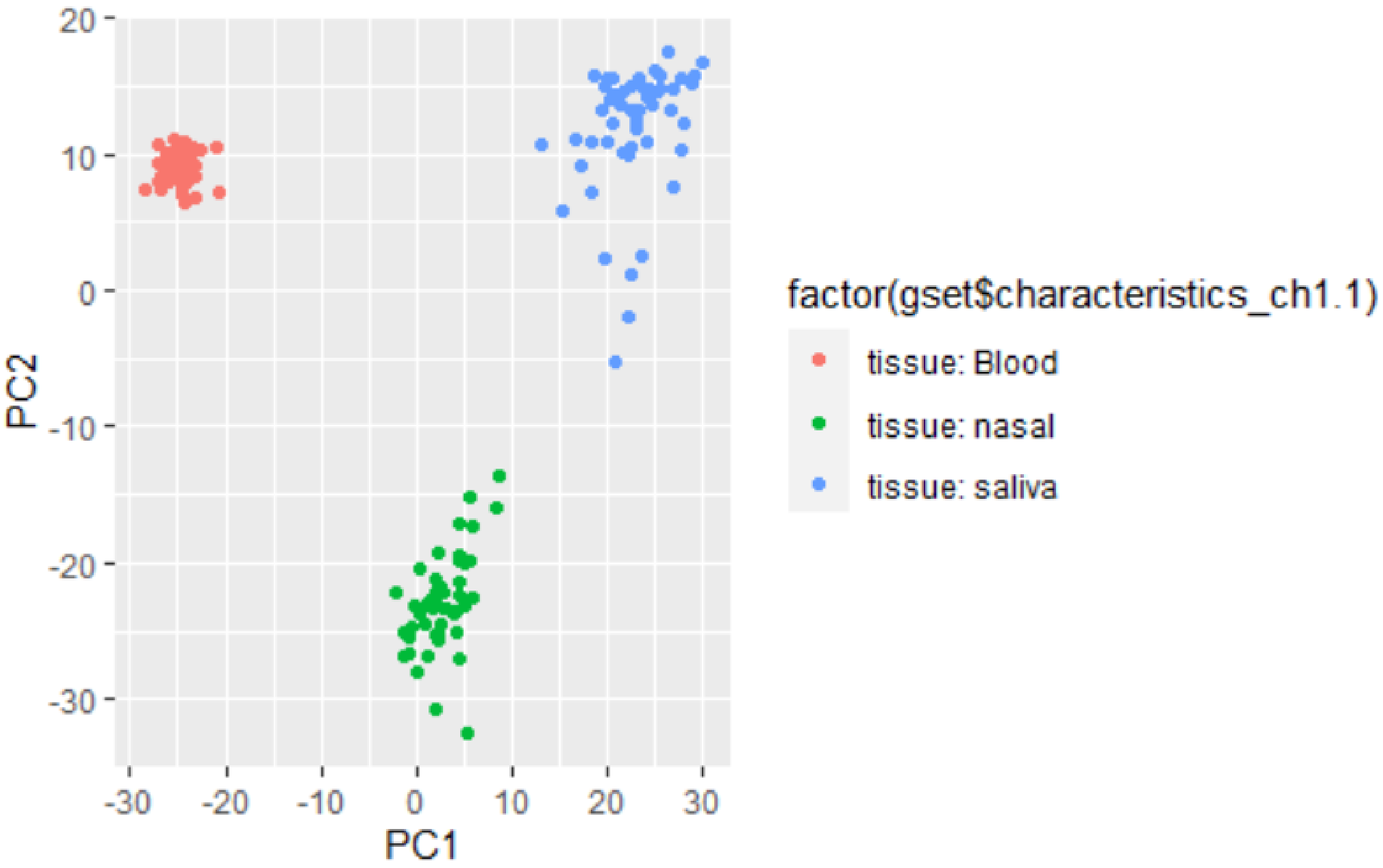
Always start with the first 2 PCs



- Clusters are drawn independent of the clinical phenotypes
- Based on trends of variance in the data

- Next step is to find out how the components are separating the data

Colour based on the phenotypes we have information on

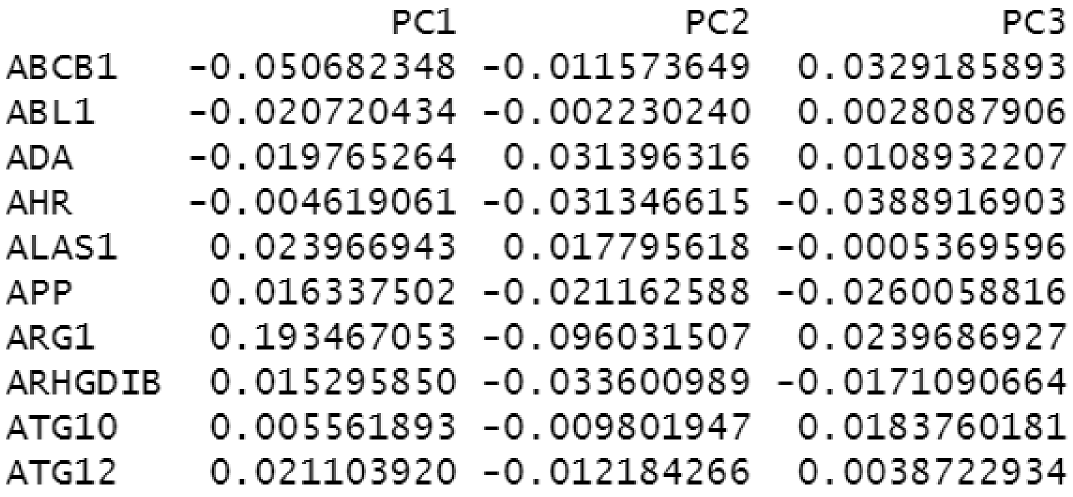


What if the data separated on sex?

- May be an indication of a co-variate
  - Is the pathology appearing different in males and females?
  - Was the data collected differently between males and females?

- May be able to be improved with batch correction, but this will add noise
  - More on this next week!

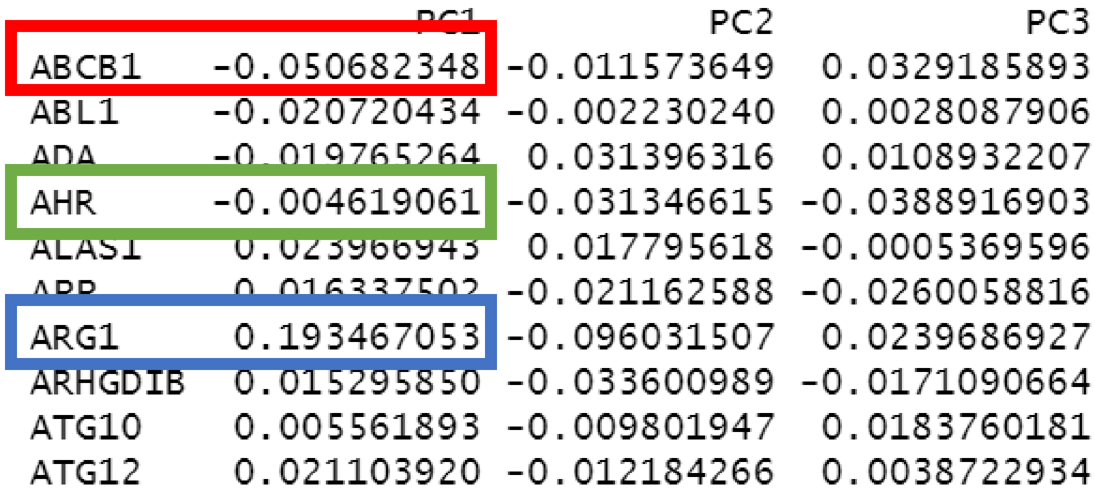
Rotations



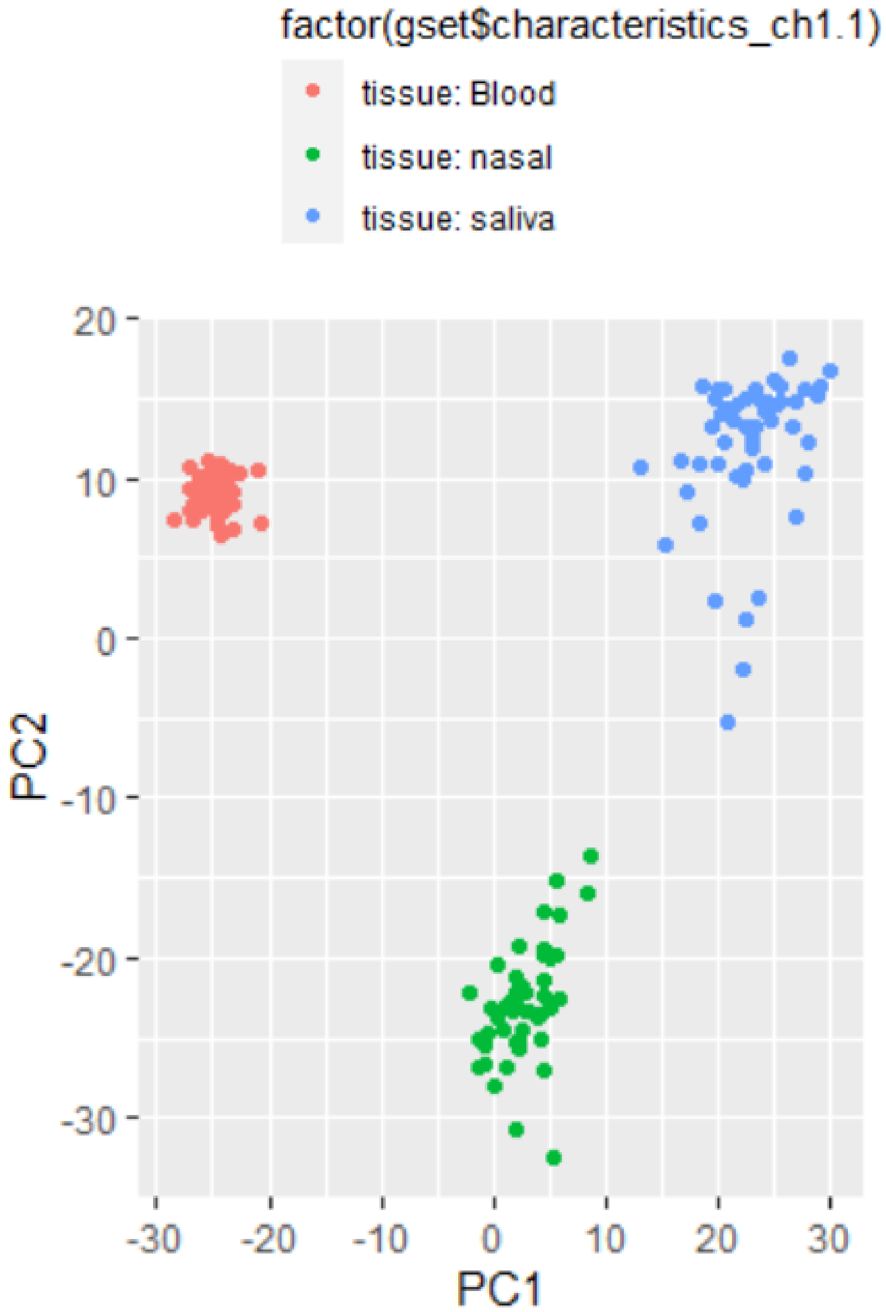
```
                PC1          PC2           PC3
ABCB1   -0.050682348 -0.011573649  0.0329185893
ABL1    -0.020720434 -0.002230240  0.0028087906
ADA     -0.019765264  0.031396316  0.0108932207
AHR     -0.004619061 -0.031346615 -0.0388916903
ALAS1    0.023966943  0.017795618 -0.0005369596
APP      0.016337502 -0.021162588 -0.0260058816
ARG1     0.193467053 -0.096031507  0.0239686927
ARHGDIB  0.015295850 -0.033600989 -0.0171090664
ATG10    0.005561893 -0.009801947  0.0183760181
ATG12    0.021103920 -0.012184266  0.0038722934
```

- Each gene contributes to each PC
- Closer to 0 == lesser contribution
- Higher value in the positive or negative reflects how it is separating the samples

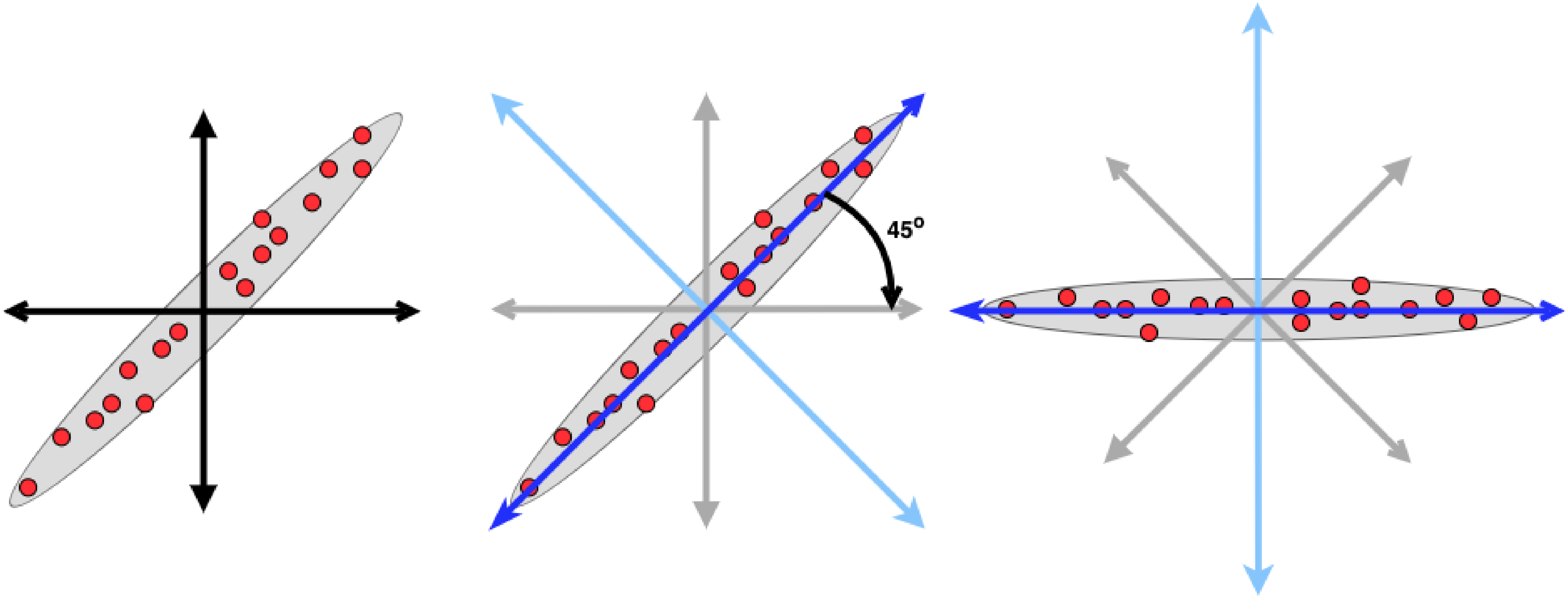
Rotations



```
                 PC1          PC2           PC3
ABCB1   -0.050682348 -0.011573649  0.0329185893
ABL1    -0.020720434 -0.002230240  0.0028087906
ADA     -0.019765264  0.031396316  0.0108932207
AHR     -0.004619061 -0.031346615 -0.0388916903
ALAS1    0.023966943  0.017795618 -0.0005369596
APP      0.016337502 -0.021162588 -0.0260058816
ARG1     0.193467053 -0.096031507  0.0239686927
ARHGDIB  0.015295850 -0.033600989 -0.0171090664
ATG10    0.005561893 -0.009801947  0.0183760181
ATG12    0.021103920 -0.012184266  0.0038722934
```



Rotations

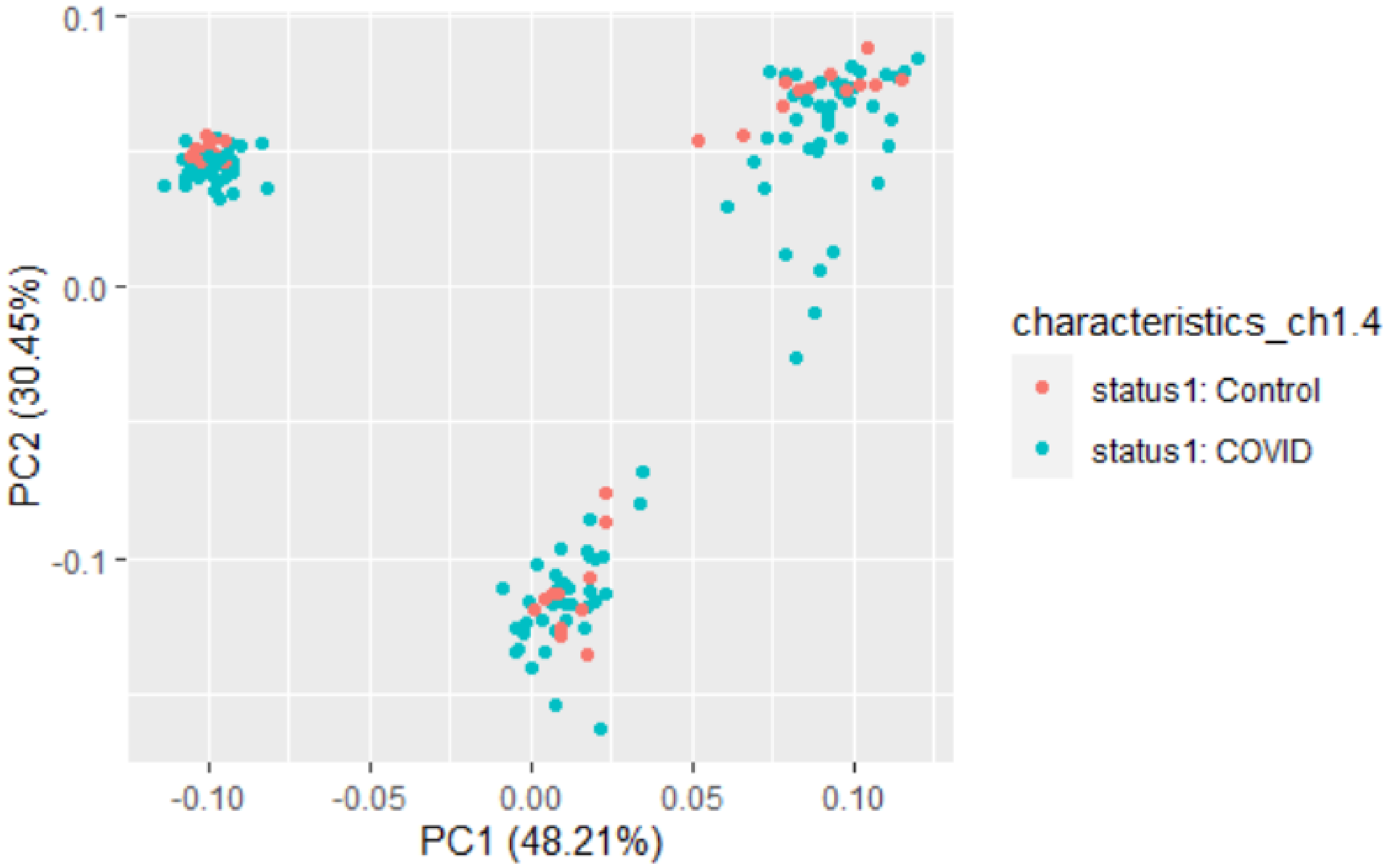


Original Coordinates

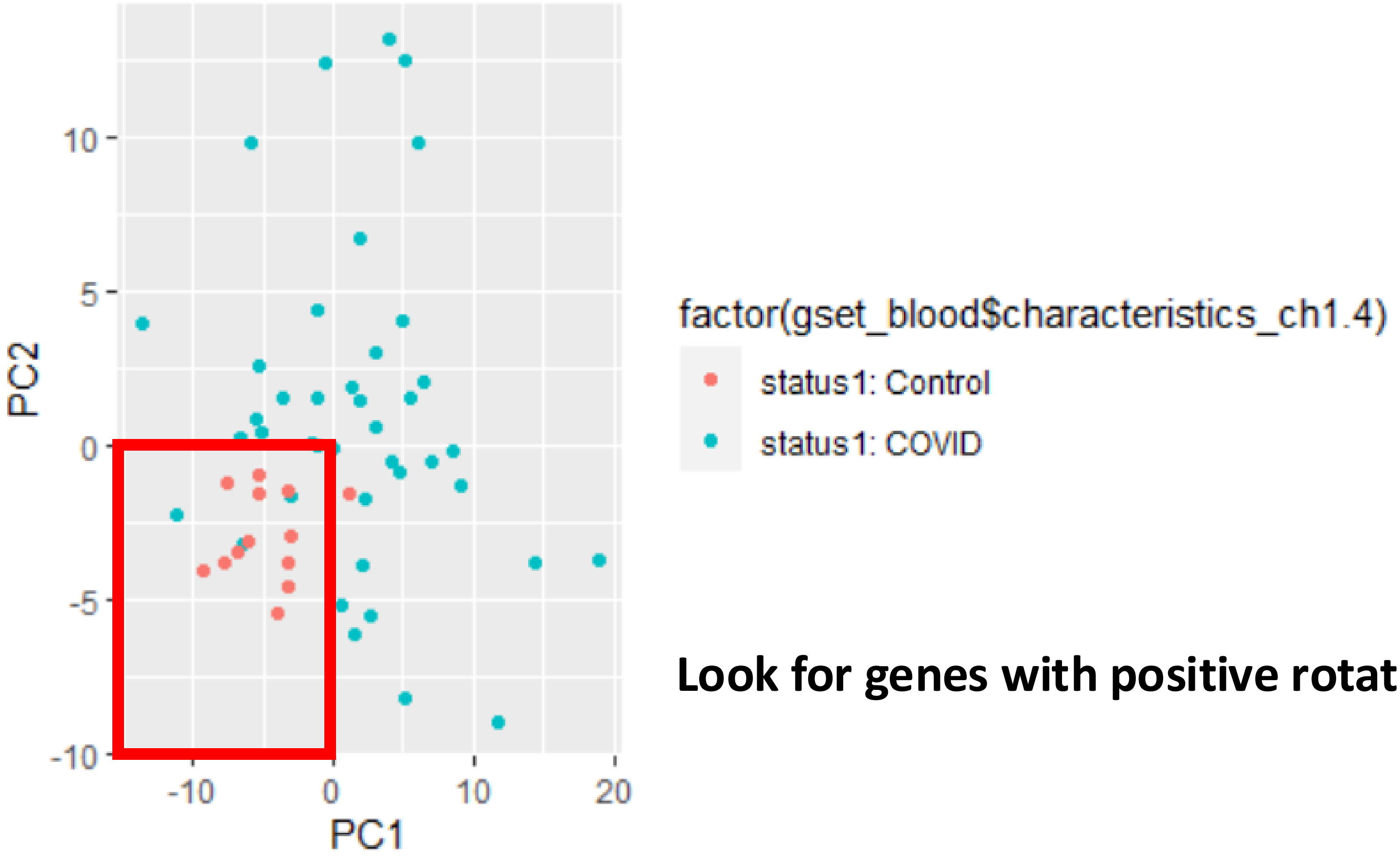
Principle Component

New Rotated Coordinates

What can we do if samples are not separating based on pathology?

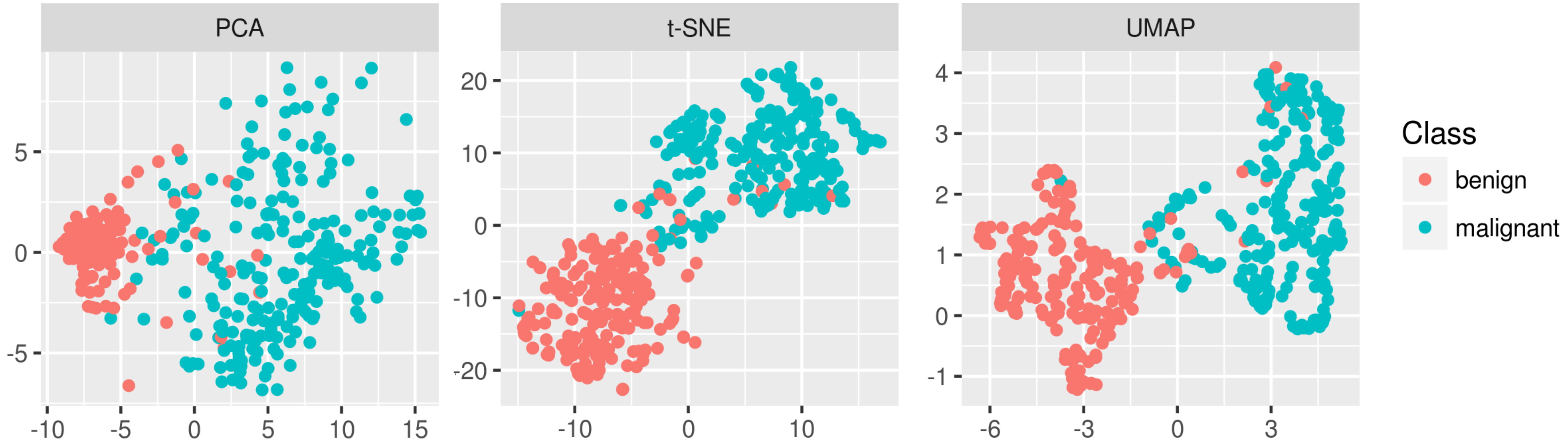


Controls are only in one quadrant



Look for genes with positive rotations in PC1 and/or PC2

Many strategies for Dimensionality Reduction



Also NMDS, PCoA

Wrap-up

- Dimensionality reduction reduces the complexity of data by eliminating repetition or consolidating similarities

- Reduce the number of dimensions by
  - Reducing redundancy
  - Grouping similar features together

- There are as many principal components as there are features
- Each gene contributes to each principal component

- In addition to PCA, TSNE and UMAP are alternate strategies for dimensionality reduction
  - NMDS and PCoA are other ordination methods (large amount of data distilled down to 2D or 3D)