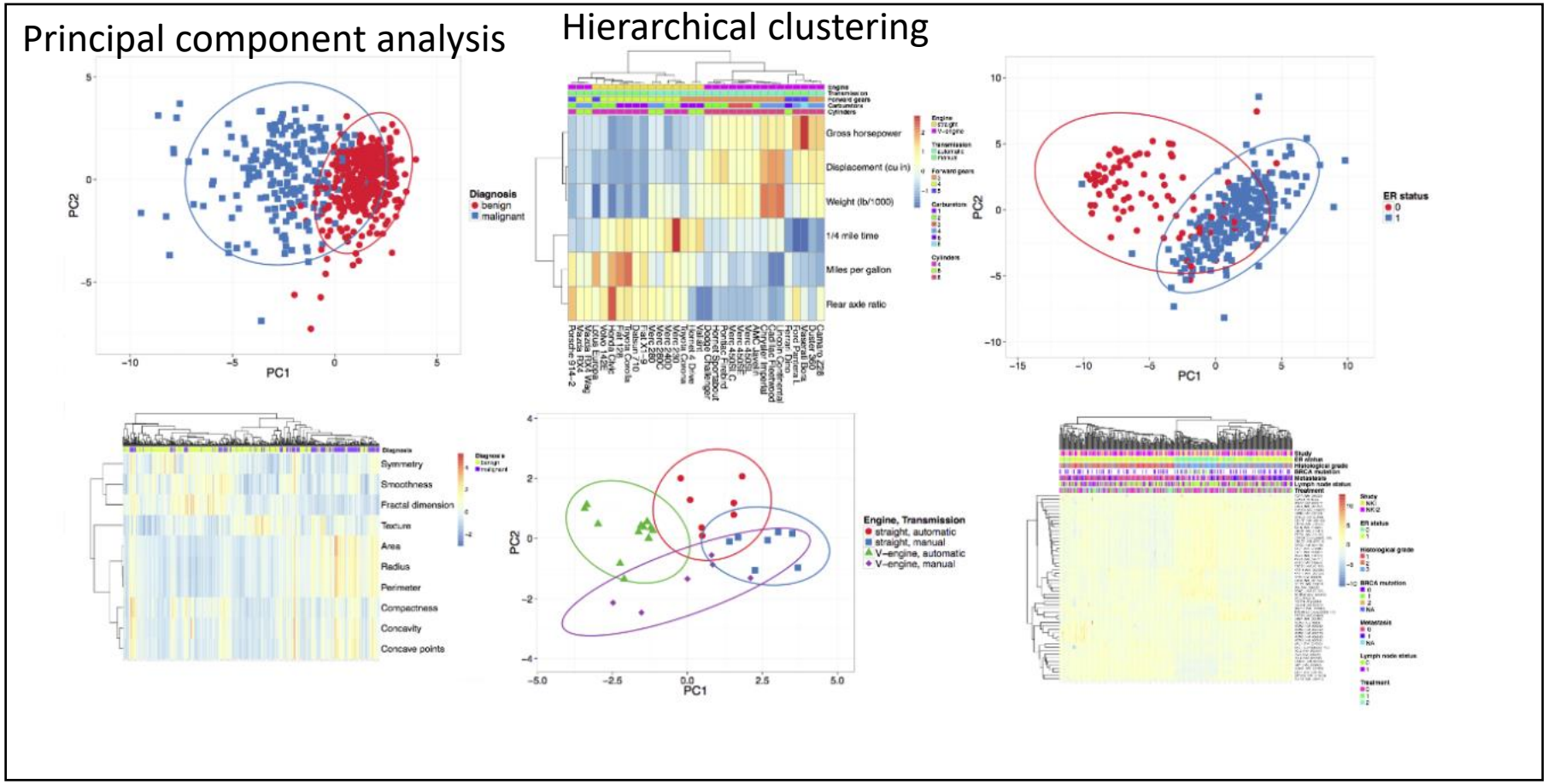
Clustering analysis: correlations and distances

How do we visualize and make sense of large data?



Clustering Methods

- Supervised clustering
- Unsupervised learning: Partitioning clustering (PCA and kNN) and Hierarchical clustering

Distances measures: The classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations. The result of the computation is known as a dissimilarity or distance matrix.

The choice of distance measures is very important, as it has a strong influence on the cluster result.

Distances: Correlations, Euclidian and Manhattan
Methods: Single, Complete, Averages and Ward

Correlations

Correlation tests of whether there is a straight-line relationship between two variables.

- *Use correlation* if individuals are sampled at random from a population and we want to know whether the two variables, e.g. volume of stomach and diameter of mouth, are linearly related in the population on average.

Pearson correlation coefficient

- Pearson correlation coefficient is one such objective measure of linear relation between two variables.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- Correlation quantifies the degree and direction to which two variables are related.
- Correlation does not fit a line through the data points. But simply is computing a correlation coefficient that tells how much one variable tends to change when the other one does.
- When r is 0.0, there is no relationship.
- When r is positive, there is a trend that one variable goes up as the other one goes up. When r is negative, there is a trend that one variable goes up as the other one goes down.

Interpreting Pearson's correlation coefficient

(1) Correlation is independent of the units in which two variables are measured. You can have for example (mL) or (L).

(2) High correlation may indicate a strong association but not causation. This means that the variables x and y are not distinguished as "predictor" and "outcome".

(3) The observed correlation (or lack of it) may be due to a confounding variable. In some situations, the observed associations (or lack of it) may be spurious and, in fact, reflect the effect of a third variable, referred as a " confounding variable".

(4) Correlation is influenced by the range of the X and Y variables. The greater the range of the x and y variables in the sample, the greater the correlation between them. Thus, a single outlying observation might give us a falsely elevated correlation coefficient.

Assumption used in calculating the Pearson correlation

(1) We are assuming that both X and Y are measure on an interval scale.

(2) Both X and Y are assumed to follow a normal probability distribution. This assumption allows us to perform hypothesis tests and construct confidence intervals.

Compute correlation in R
R functions

- **cor()** computes the **correlation coefficient**

- **cor.test()** test for association/correlation between paired samples. It returns both the **correlation coefficient** and the **significance level** (or p-value) of the correlation .

The simplified formats are:

```
cor(x, y, method = c("pearson", "kendall", "spearman"))
cor.test(x, y, method=c("pearson", "kendall", "spearman"))
```

Interpretation

- **t** is the **t-test statistic** value ,
- **df** is the degrees of freedom,
- **p-value** is the significance level of the **t-test**.
- **conf.int** is the **confidence interval** of the correlation coefficient at 95%;
- **sample estimates** is the correlation coefficient .

Preliminary test to check the test assumptions

1. Is the covariation linear? The relationship is linear. In the situation where the scatter plots show curved patterns, we are dealing with nonlinear association between the two variables.

2. Are the data from each of the 2 variables (x, y) follow a normal distribution?

1. Use Shapiro-Wilk normality test –> R function: **shapiro.test**()
2. and look at the normality plot —> R function: **ggpubr::ggqqplot**()

- **Shapiro-Wilk test** can be performed as follow:
  - Null hypothesis: the data are normally distributed
  - Alternative hypothesis: the data are not normally distributed

Example

```
my_data <- mtcars
Head(my_data)
shapiro.test(my_data$mpg)
```

```
data:  my_data$mpg
W = 0.94756, p-value = 0.1229
```

From the output, the two p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

Spearman Rank correlation

Spearman's rank correlation, is another statistics used for measuring the correlation between a pair of variables.
It is called a nonparametric measure (i.e. no assumption) and is preferred when assumptions required for Pearson's correlation coefficient is violated -that is, when X and/or Y are not measured on an interval scale, or when X and/or Y do not follow a normal probability distribution.

To calculate Spearman's correlation coefficient, we need to assign a rank to the individual values of X and Y- that is sort each of X and Y in increasing order and assign them ranks so that the smallest observation has a rank N.

Advantage of using:

An advantage of Spearman's correlation coefficient is that is can be used to evaluate nonlinear relation between variables when the direction of the relationship does not change.

Formula:

- To avoid assumptions about the underlying distribution and to know whether the two variables, are linearly related

Spearman's rank correlation

$$\rho = 1 - \frac{6\sum d_i^{\,2}}{n(n^2-1)} \qquad d_i = x_i - y_i$$

(difference between ranks)

Null hypothesis

The null hypothesis (H_0) and alternative (H_1) for Spearman's rank correlation test are:

- H_0: there is no monotonic association between the two variables
- H_1: there is a monotonic association between the two variables

Kendall rank correlation test

The **Kendall rank correlation coefficient** or **Kendall's tau** statistic is used to estimate a rank-based measure of association. This test may be used if the data do not necessarily come from a bivariate normal distribution.

Kendall rank correlation

$$S = \sum_{i<j} (sign(x[j] - x[i]) * sign(y[j] - y[i]))$$

$$D = \frac{n(n-1)}{2} \qquad \tau = \frac{S}{D}$$

An advantage of the Kendall rank correlation over the Spearman rank correlation is that the score function S nearly normally distributed for small n and the distribution of S is easier to work with.

Kendall rank correlation test

```
result <- cor.test(my_data$wt, my_data$mpg, method="kendall")
result
```

Kendall's rank correlation tau data:
my_data$wt and my_data$mpg z = -5.7981,
p-value = 6.706e-09
alternative hypothesis: true tau is not equal to 0 sample estimates: tau -0.7278321

The **correlation coefficient** between x and y are -0.7278 and the *p-value* is 6.70610^{-9}.

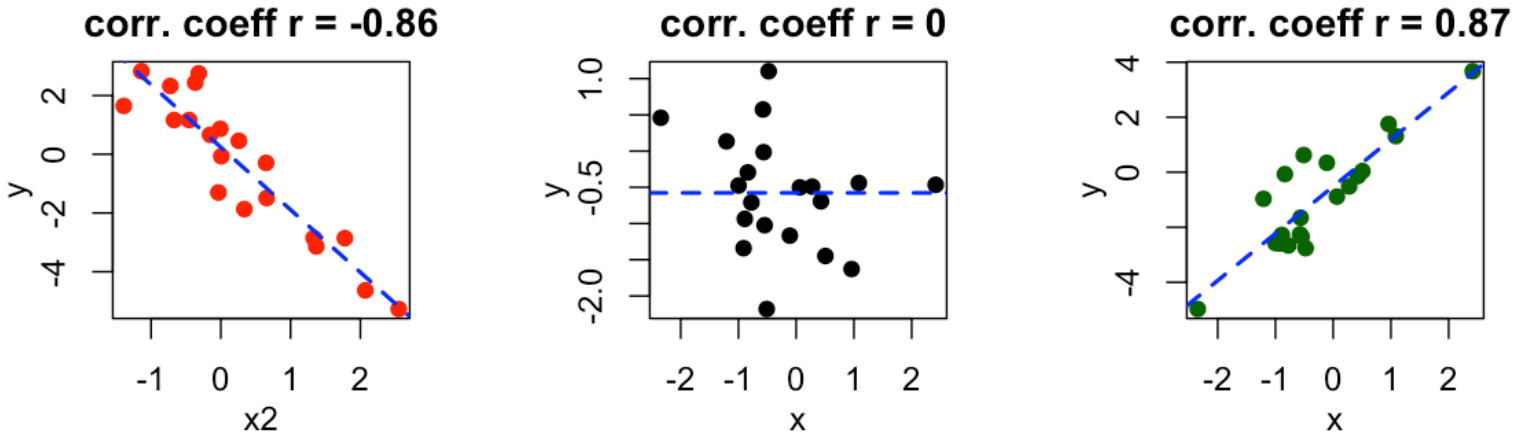
Interpret correlation coefficient

Correlation coefficient is comprised between **-1** and **1**:

-1 indicates a strong **negative correlation** : this means that every time **x increases**, **y decreases** (left panel figure)

0 means that there is no **association** between the two variables (x and y) (middle panel figure)

1 indicates a strong **positive correlation** : this means that **y increases** with **x** (right panel figure)



- Use the function **cor.test**(x,y) to analyze the correlation coefficient between two variables and to get significance level of the correlation.

- Three possible correlation methods using the function **cor.test**(x,y): pearson, kendall, spearman

Distance: 1-Pearson Correlation

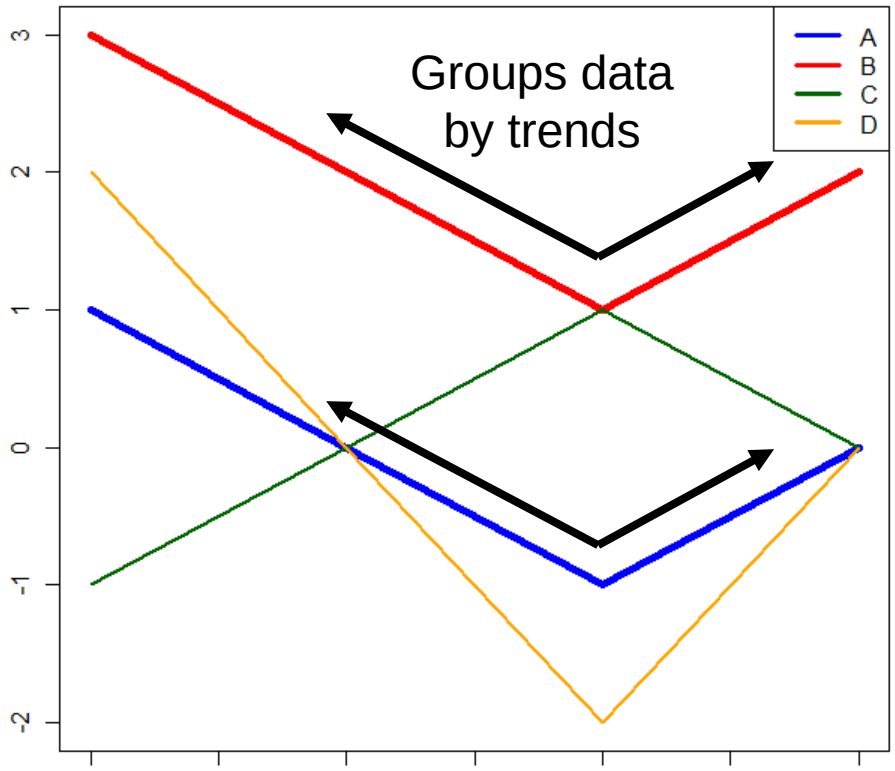
How similar are two observations?

A	1	0	-1	0
B	3	2	1	2
C	-1	0	1	0
D	2	0	-2	0

1 – Pearson Correlation

$$1 - \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} =$$

	A	B	C
B	0		
C	-2	-2	
D	0	0	-2



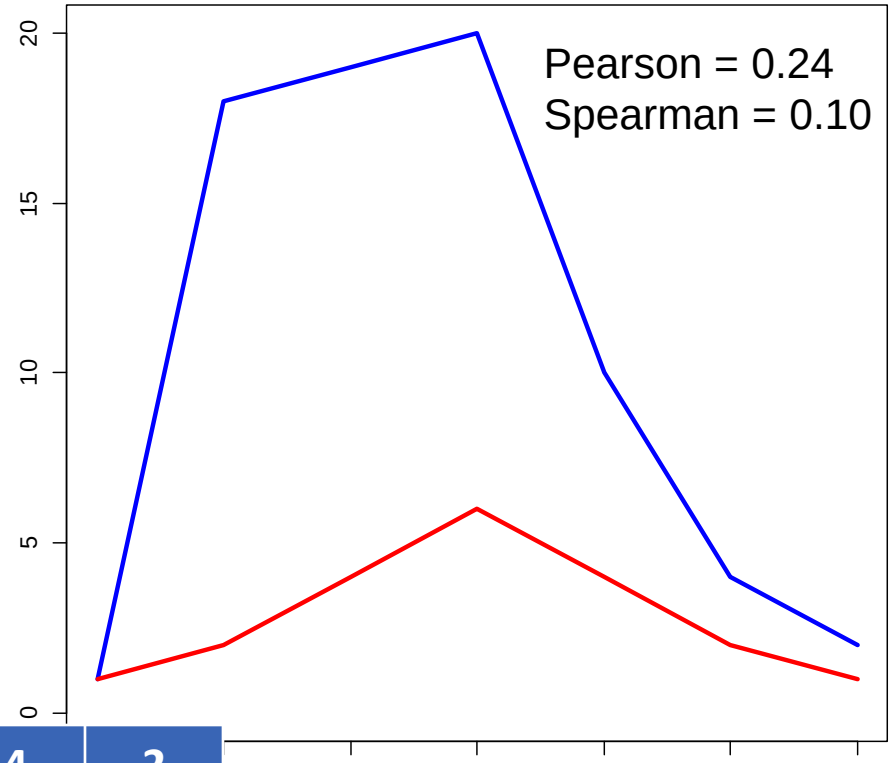
$$dist(A,B) = 1 - \frac{(1-0)(3-2)+(0-0)(2-2)+(-1-0)(1-2)+(0-0)(2-2)}{(4-1)(.816)(.816)} = 0$$

Distance: 1-Spearman Correlation

How similar are two observations?

1 – Spearman Correlation = 1-.90 = 0.10

similar to Pearson correlation but using ranks

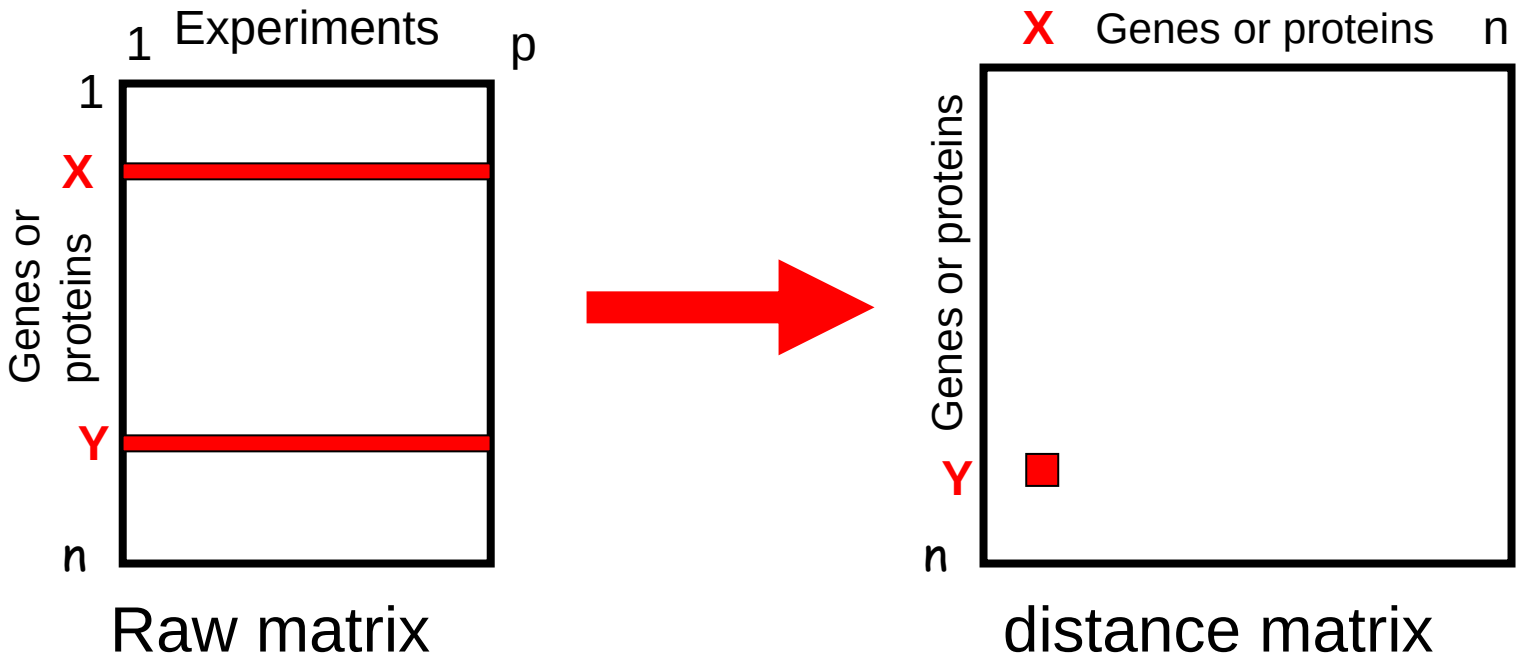


A	1	18	19	20	10	4	2
Rank	1	5	6	7	4	3	2
B	1	2	4	6	4	2	1
Rank	1.5	3.5	5.5	7	5.5	3.5	1.5

Distribution and magnitude of data do not matter

What is a distance measure?

Pair-wise comparison between 2 rows or columns in a matrix



ID	ALL2	ALL2	ALL3	E2A1	E2A2	E2A3
X	8.05	9.36	8.39	6.55	5.94	6.53
Y	10.52	10.27	9.73	9.22	9.45	8.92

Euclidean distance = 5.82

Distance: Euclidean

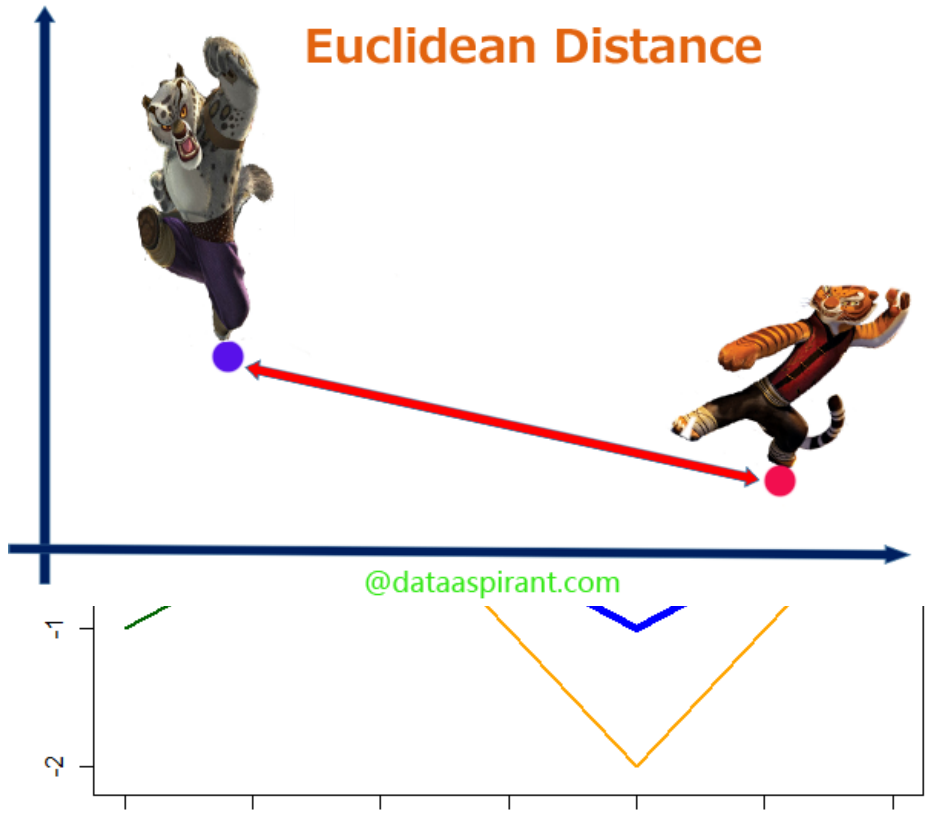
How similar are two observations?

A	1	0	-1	0
B	3	2	1	2
C	-1	0	1	0
D	2	0	-2	0

Euclidean

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} =$$

	A	B	C
B	4		
C	2.8	4.9	
D	1.4	4.2	4.2

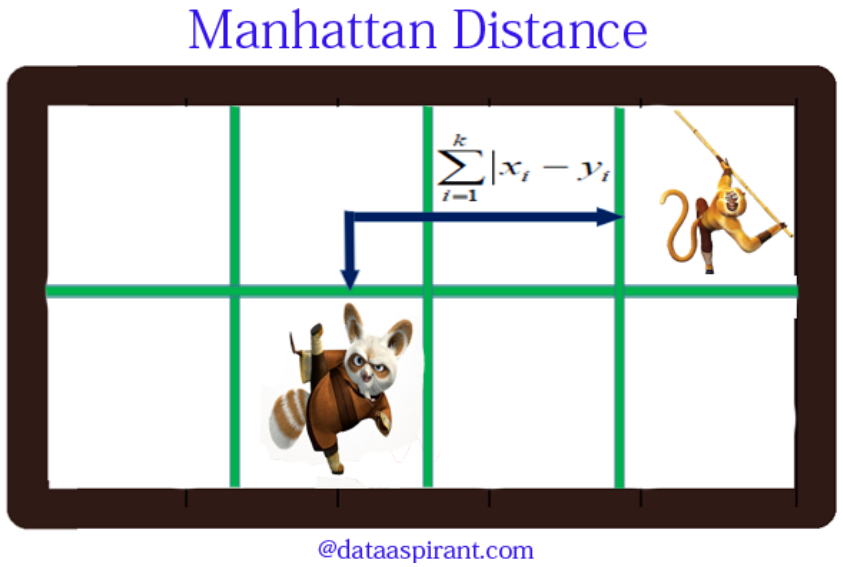


$$dist(A,B) = \sqrt{\left((1-3)^2 + (0-2)^2 + (-1-1)^2 + (0-2)^2\right)} = 4$$

Distance: Manhattan

How similar are two observations?

A	1	0	-1	0
B	3	2	1	2
C	-1	0	1	0
D	2	0	-2	0



Manhattan

$$\sum_{i}^{n}|x_i - y_i| =$$

	A	B	C
B	8		
C	4	8	
D	2	8	8

$$dist(A,B) = |1-3| + |0-2| + |-1-1| + |0-2| = 8$$

Clustering

- Partitioning clustering: Subdivides the data into a set of k groups
- Hierarchical clustering: Identify groups in the data without subdividing it

Clustering Algorithms

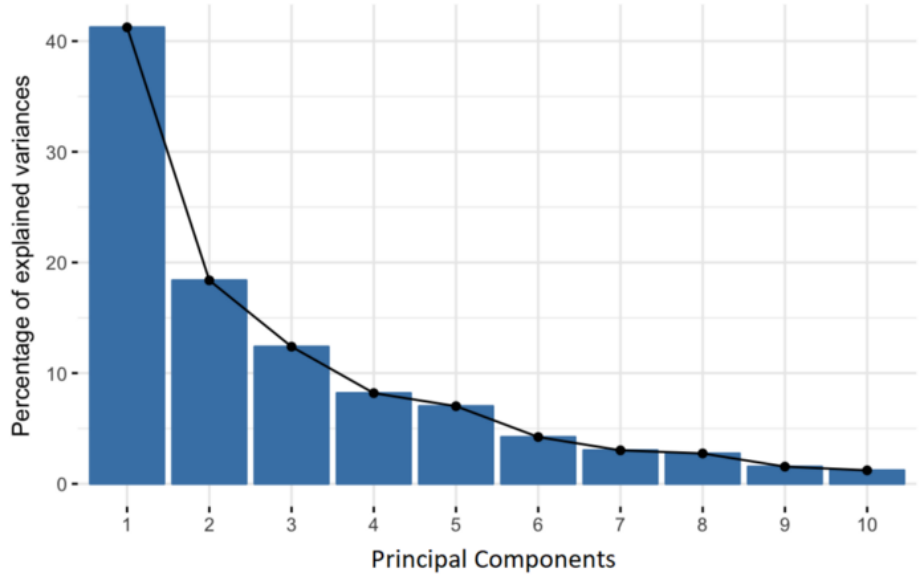
- Hierarchical – produces a hierarchy of clusters displayed as a tree
  - Agglomerative – bottom up approach, adds each element (or cluster) by distance in increasing order
  - Divisive – top down approach, starts with all elements in a one cluster
- Non-hierarchical – Partitions into groups
  - K-means – each element belongs to only one group
  - Fuzzy clustering – each element may belong to multiple groups
- Biclustering – Clustering on both rows and columns at the same time, elements may belong to multiple groups

Principle component analysis

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

- To sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

- Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.
- The idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.



Principal Component Analysis (PCA)

- Transform a set of correlated variables to new set of uncorrelated variables known as principal components
- PCs define direction with largest variation
- Most variation in data should be represented in first few PCs
- Visualize samples with a large number of variables
- Identify patterns in data
- Often used to detect outliers

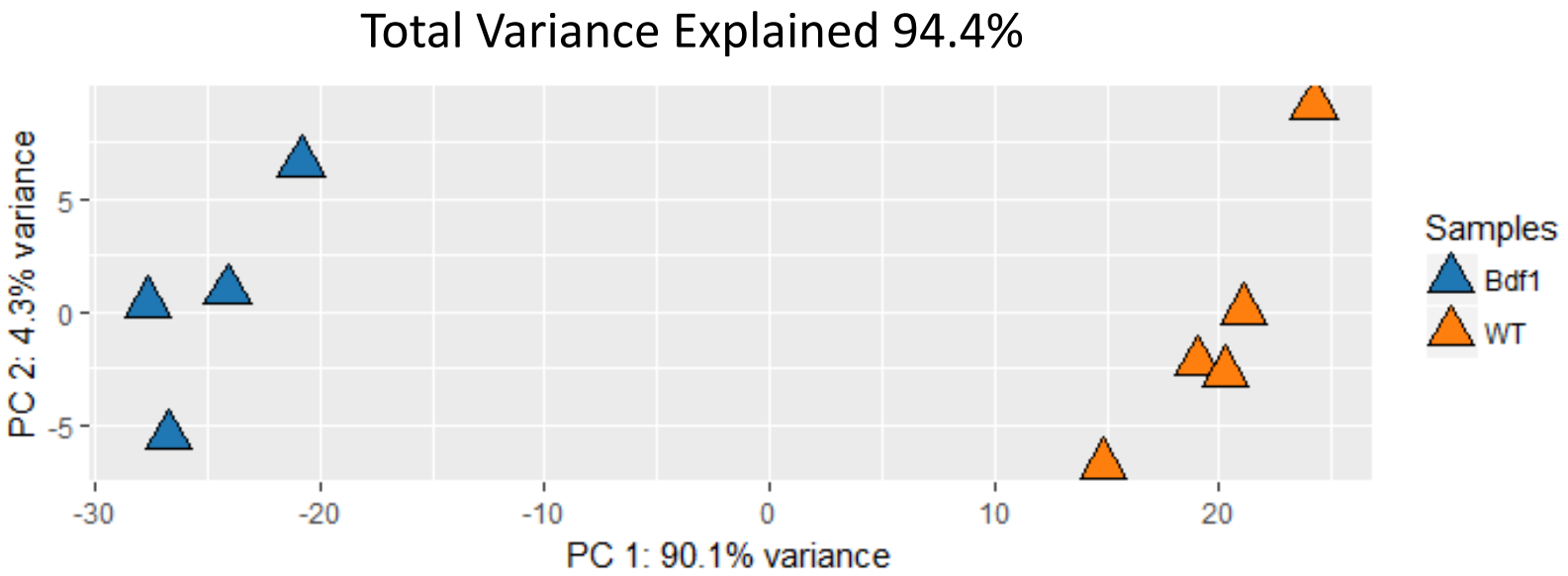
R Functions/Packages

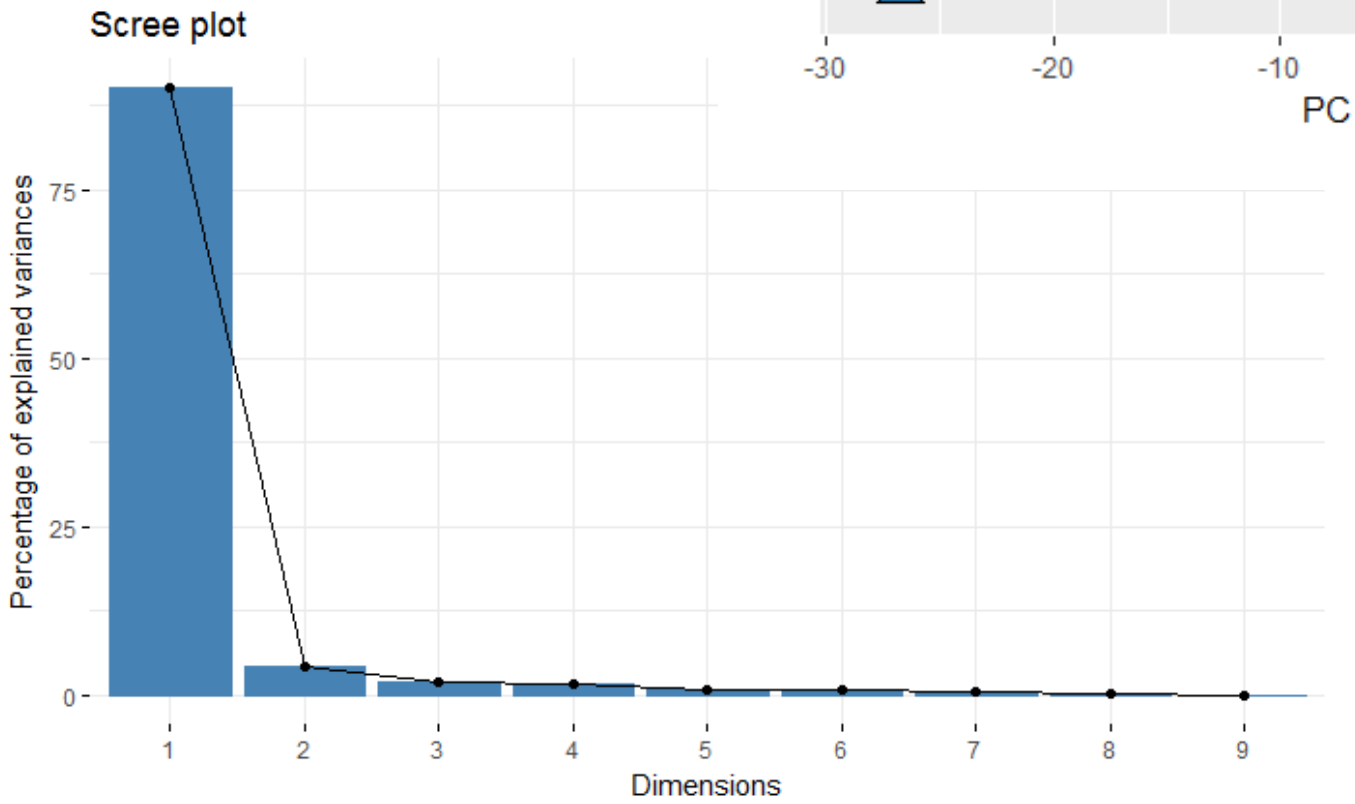
prcomp – calculates PCA using singular value decomposition

factoextra – package with functions to extract and visualize pca results

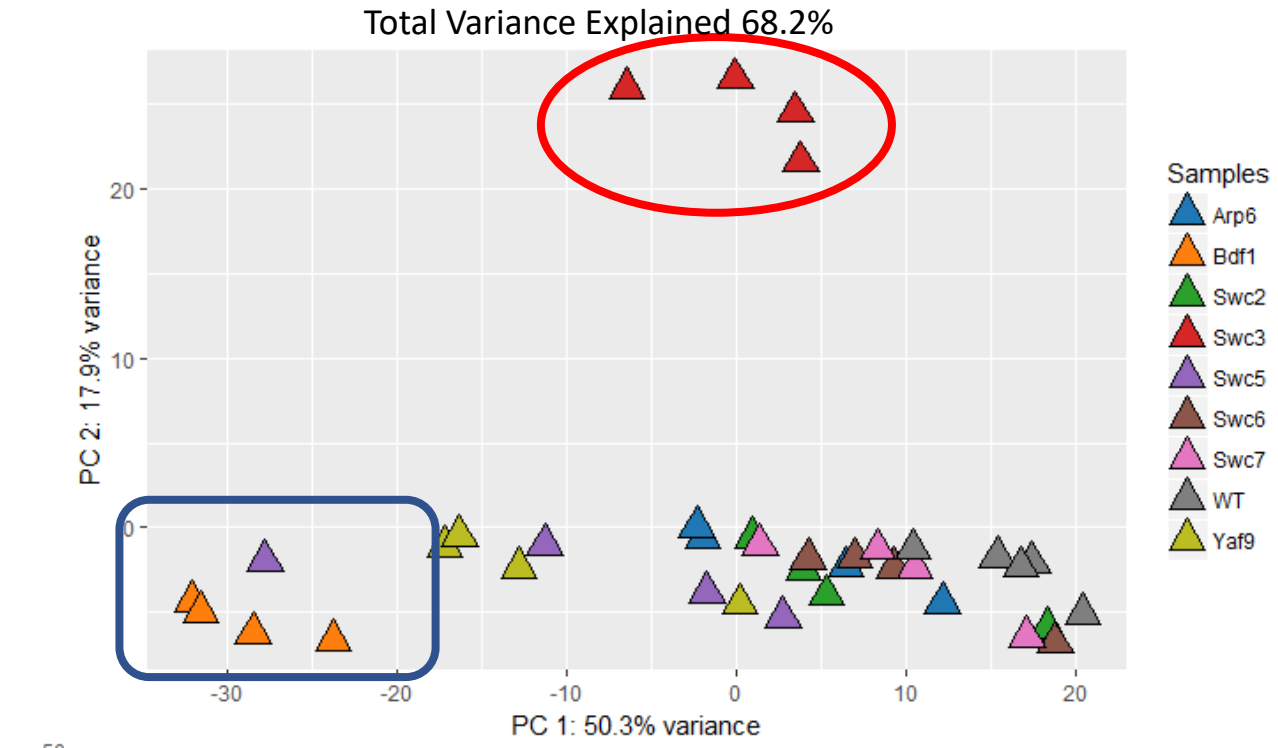
PCA Example

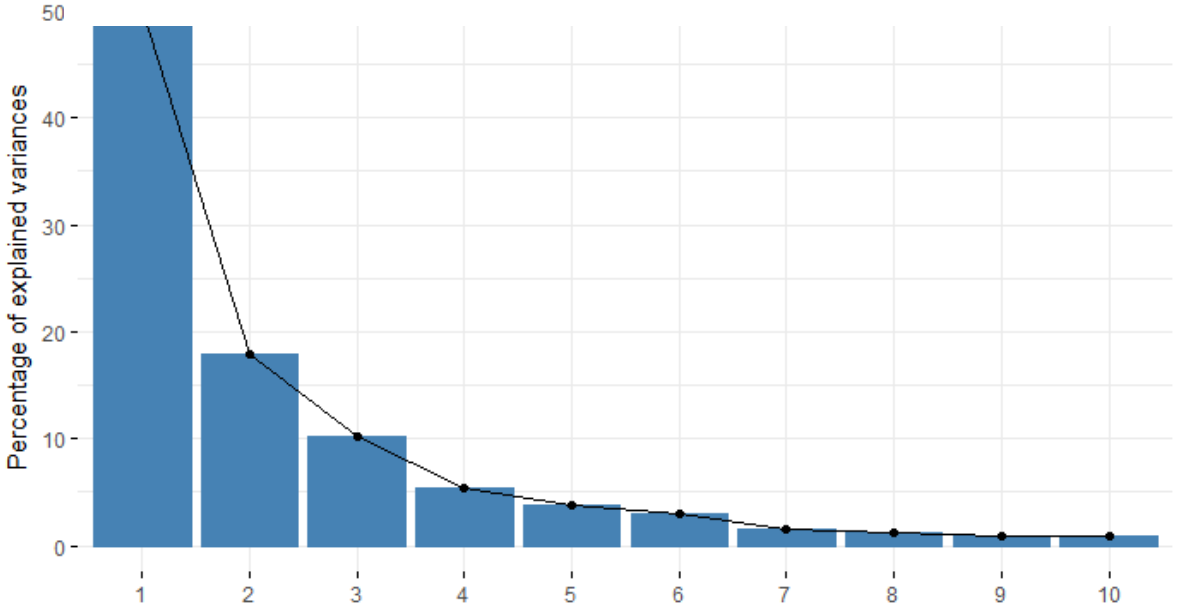
Two dimensional scatterplot to visualize the first 2 principal components

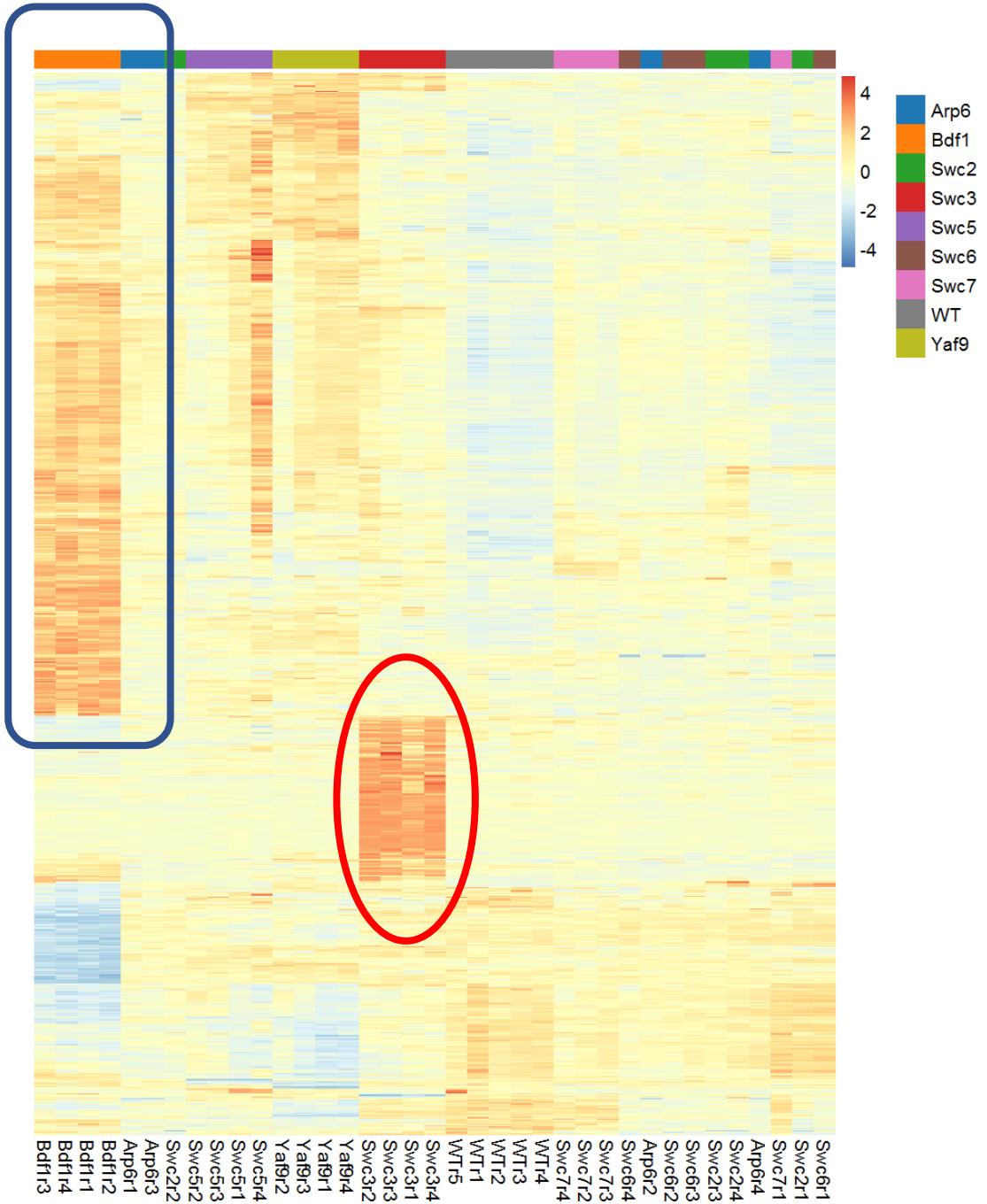




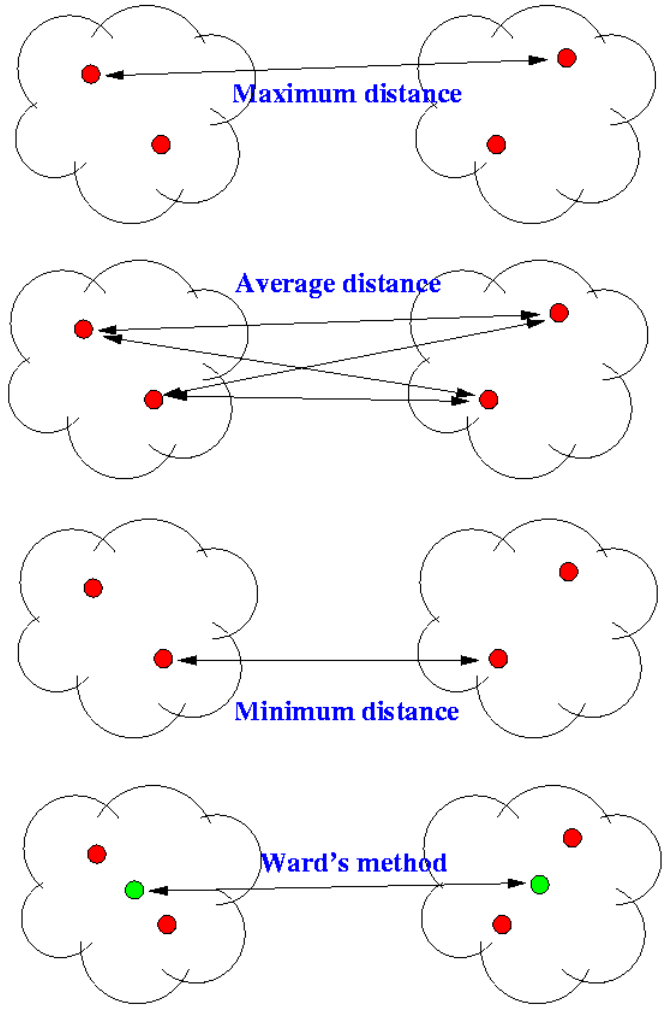
- Each principal component explains a percentage of the variance in the data.
- The number of PCs is variable depending upon sources of variance in the data


Total Variance Explained 68.2%
PC 2: 17.9% variance
PC 1: 50.3% variance
Samples
Arp6
Bdf1
Swc2
Swc3
Swc5
Swc6
Swc7
WT
Yaf9


Percentage of explained variances


Arp6
Bdf1
Swc2
Swc3
Swc5
Swc6
Swc7
WT
Yaf9

Methods for combining clusters



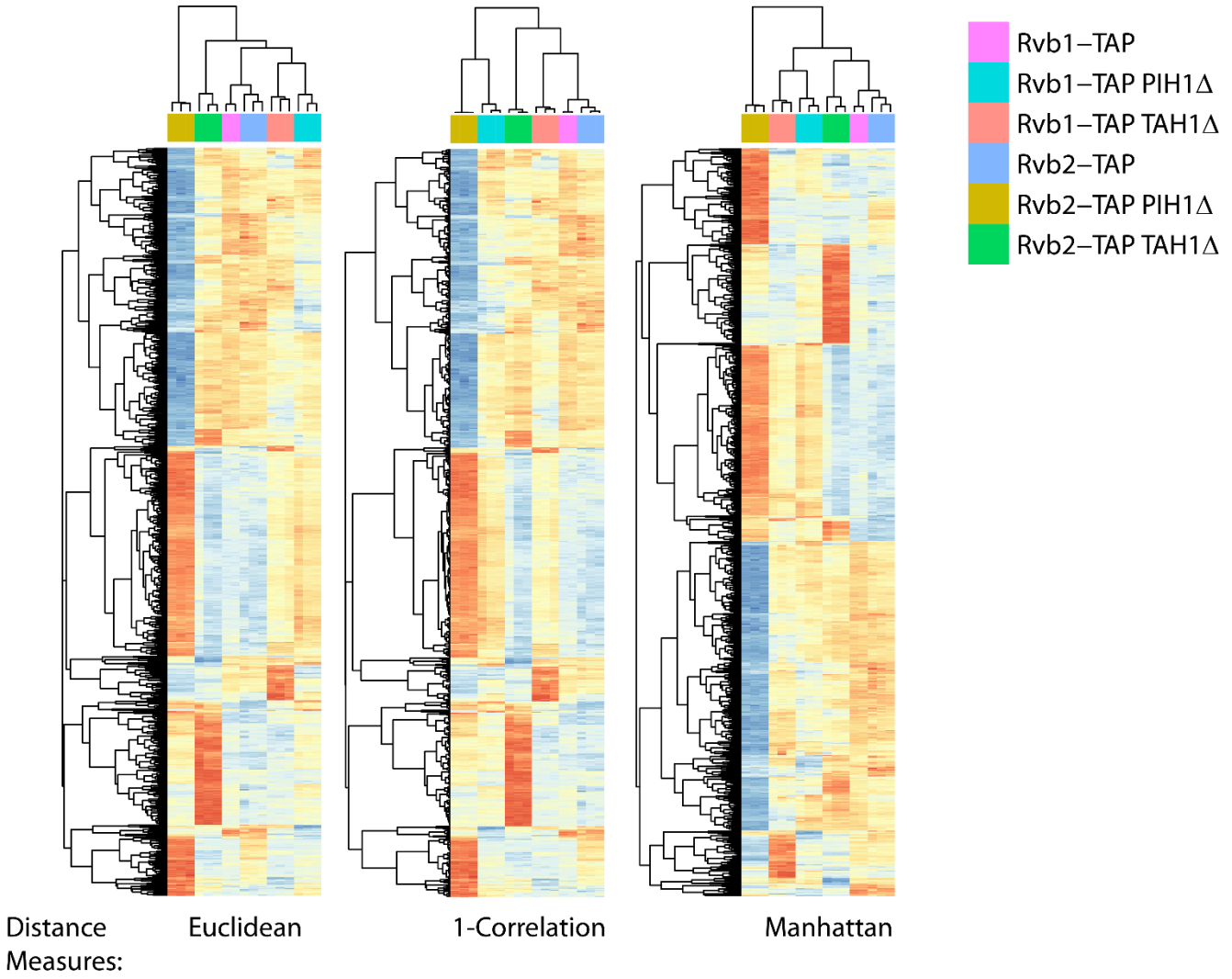
Complete linkage
maximum distance between clusters

Average linkage
average distance between clusters

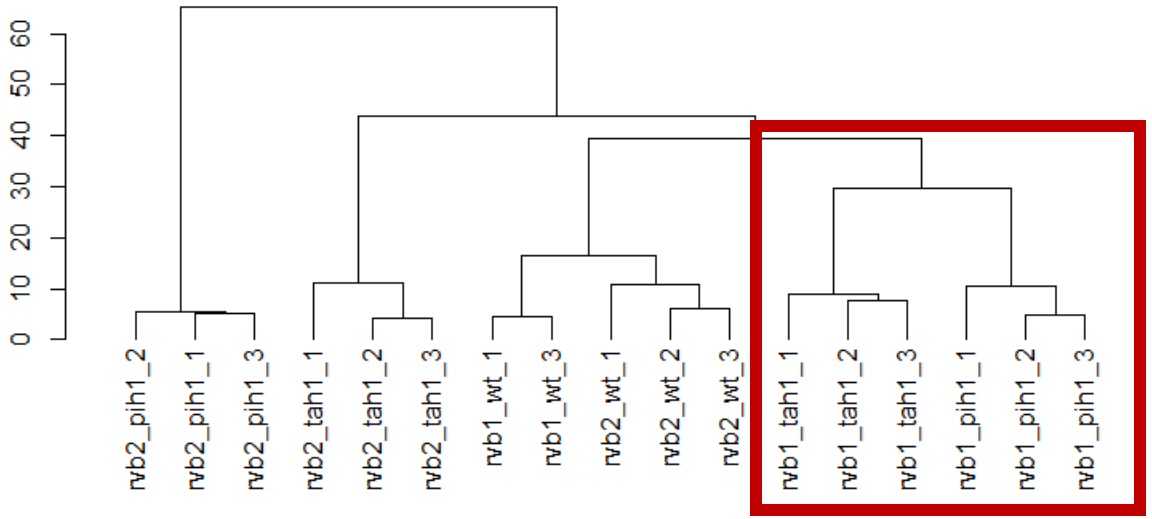
Single linkage
minimum distance between clusters

Ward's method
minimize within cluster variance

Same Data – Different Distances

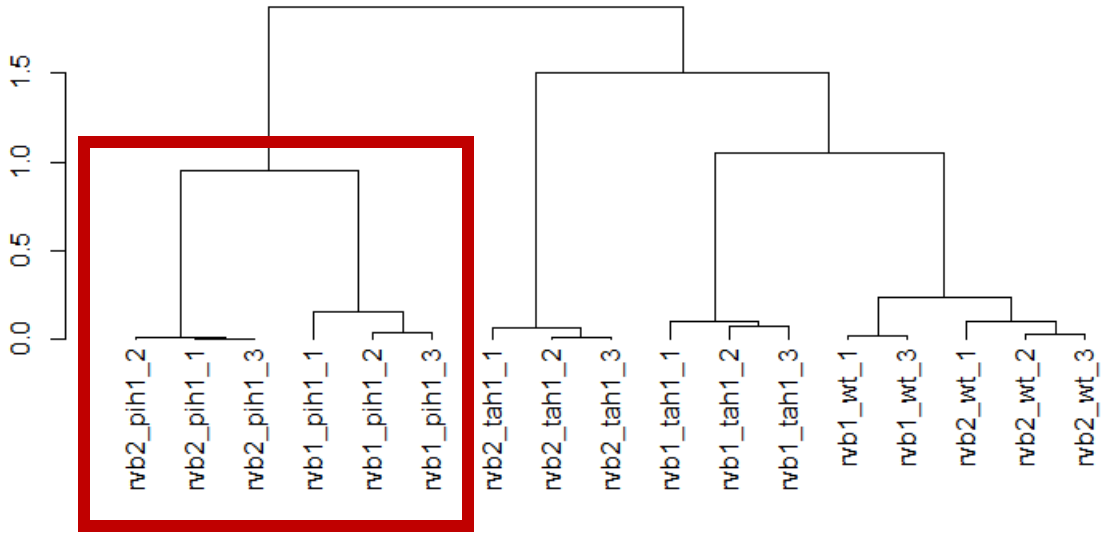


Dendrogram View



Euclidean

Rvb1 deletions cluster together



Correlation

Pih1 deletions cluster together

Different Methods for Combining Clusters