Statistical Analysis of Metabolomics Data

Xiuxia Du
Department of Bioinformatics & Genomics
University of North Carolina at Charlotte

Outline

- Introduction
- Data pre-treatment
  1. Normalization
  2. Centering, scaling, transformation
- Univariate analysis
  1. Student's *t*-tes
  2. Volcano plot
- Multivariate analysis
  1. PCA
  2. PLS-DA
- Machine learning
- Software packages

Results from data processing

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Sample	ko15	ko16	ko18	ko19	ko21	ko22	wt15	wt16	wt18	wt19	wt21	wt22
2	Label	ko	ko	ko	ko	ko	ko	wt	wt	wt	wt	wt	wt
3	1	4534353.62	4980914.48	5290739.14	4564262.9	4733236.08	3931592.59	349660.885	491793.181	645526.705	634108.849	1438254.45	1364627.84
4	2	962353.43	1047934.14	1109303.04	946943.393	984787.205	806171.473	86450.4116	120096.52	143007.948	137319.686	218483.143	291392.971
5	3	180780.817	203926.952	191015.911	190626.85	156869.08	220288.622	16269.096	43677.784	54739.1289	76318.0077	54726.1154	49679.9425
6	4	432037.001	332159.073	386966.751	334951.453	294816.236	373577.608	7643.13807	10519.9422	26472.293	33598.3228	8030.46734	0
7	5	165830.903	183665.01	150844.995	134637.117	136452.456	167008.053	24302.8057	16631.3863	19213.6611	12822.5391	12452.1286	18789.754
8	6	236249.547	255168.616	212710.776	180690.967	191746.741	152861.186	29530.1749	17037.2075	35133.1225	27706.2092	47228.7374	40193.5296
9	7	1108851.28	950126.529	674222.714	677091.267	772290.305	1013977.91	58898.7043	21990.98	27642.9741	31727.8261	45299.168	8715.57767
10	8	4809521.45	3931304.64	2913711.55	2819100.93	3284987.28	4346409.7	259229.273	314154.323	161900.584	200939.74	247976.865	237853.621
11	9	980012.03	1319875.61	1375014.96	949318.774	899985.582	799401.791	137450.635	203674.39	223544.907	160796.141	95447.3273	110763.407
12	10	304741.632	426745.765	437071.769	304811.059	285571.596	275335.275	57976.3758	92450.81	86924.8392	79420.9306	56605.4873	65708.7501
13	11	238208.464	272654.011	378948.167	273593.949	247266.87	179812.074	23463.0482	30685.2109	62726.3335	42656.2854	32328.4535	31575.6239
14	12	144480.636	149070.729	95314.6844	136823.134	83447.365	163903.723	15781.0387	14619.5489	19411.8176	23885.3028	33471.3336	28779.3836
15	13	1135610.95	904522.365	832780.32	532867.709	570306.904	729939.966	112167.159	154475.666	51277.225	91243.2653	79838.1181	72094.7396
16	14	201588.695	167042.482	143015.845	105447.719	92547.8575	95737.3812	18621.1758	18071.6701	3745.92163	17251.6947	9713.90424	8609.33088
17	15	456678.246	324908.659	190365.213	192504.5	328211.717	439481.04	38080.0815	27120.0761	17746.3781	17365.864	21200.2003	31711.0758
18	16	6463382.87	5630552.38	4398403.51	3085430.38	3289887.97	2626186.41	402409.001	573880.805	403748.596	412858.155	385802.429	297208.544
19	17	1521141.36	1315446.89	990625.769	707798.716	759301.22	605854.743	114347.574	140213.418	111356.479	69186.0779	88438.2624	43885.9531
20	18	328565.947	567776.99	391437.155	254805.159	305032.229	218120.799	54773.6586	148383.91	77678.498	66290.1789	26578.2065	69723.9671
21	19	334039.86	257216.862	222655.68	126090.44	154979.578	218179.86	67313.1827	77539.4099	50070.8683	41084.4371	39098.8359	50169.4484
22	20	710639.895	580437.414	433480.221	286404.148	317220.06	323722.843	102300.92	94799.875	91915.8527	57353.3102	55472.9859	37194.5061
23	21	221263.214	388214.929	186860.976	162564.375	148367.965	182694.97	29717.6663	50643.6507	26721.2749	14828.8956	27497.5713	21940.7572
24	22	1189376.08	2242782.41	1535138.25	932580.306	1206594.77	823710.047	181130.138	521478.345	258986.27	237846.961	197582.43	246931.896

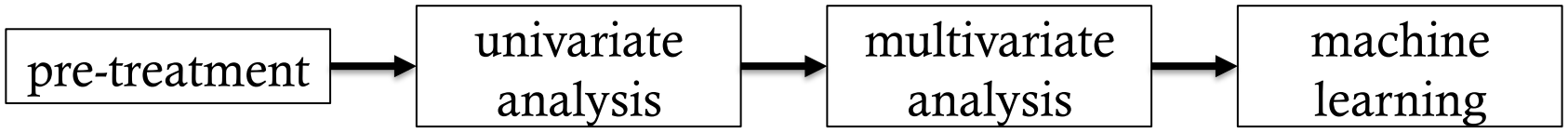
Next, analysis of the quantitative metabolite information …

Metabolomics data analysis

- Goals
  - biomarker discovery by identifying significant features associated with certain conditions
  - Disease diagnosis via classification

- Challenges
  - Limited sample size
  - Many metabolites / variables

- Workflow



Data Pre-treatment

Normalization (I)

- Goals
  - to reduce systematic variation
  - to separate biological variation from variations introduced in the experimental process
  - to improve the performance of downstream statistical analysis

- Sources of experimental variation
  - sample inhomogeneity
  - differences in sample preparation
  - ion suppression

Normalization (II)

- Approaches
  - Sample-wise normalization: to make samples comparable to each other
  - Feature/variable-wise normalization: to make features more comparable in magnitude to each other.
- Sample-wise normalization
  - normalize to a constant sum
  - normalize to a reference sample
  - normalize to a reference feature (an internal standard)
  - sample-specific normalization (dry weight or tissue volume)
- Feature-wise normalization (i.e., centering, scaling, and transformation)

Centering, scaling, transformation

Class	Method	Formula	Unit	Goal	Advantages	Disadvantages
I	Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	*O*	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
II	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
	Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{max}} - x_{i_{min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
	Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	*O*	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
	Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
	Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
III	Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij}\right)$ $\bar{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log *O*	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
	Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\bar{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√*O*	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary

Centering

- Converts all the concentrations to fluctuations around zero instead of around the mean of the metabolite concentrations
- Focuses on the fluctuating part of the data
- Is applied in combination with data scaling and transformation

Scaling

- Divide each variable by a factor
- Different variables have a different scaling factor
- Aim to adjust for the differences in fold differences between the different metabolites.
- Results in the inflation of small values
- Two subclasses
  - Uses a measure of the data dispersion
  - Uses a size measure

Scaling: subclass 1

- Use data dispersion as a scaling factor
  - auto: use the standard deviation as the scaling factor. All the metabolites have a standard deviation of one and therefore the data is analyzed on the basis of correlations instead of covariance.
  - pareto: use the square root of the standard deviation as the scaling factor. Large fold changes are decreased more than small fold changes and thus large fold changes are less dominant compared to clean data.
  - vast: use standard deviation and the coefficient of variation as scaling factors. This results in a higher importance for metabolites with a small relative sd.
  - range: use (max-min) as scaling factors. Sensitive to outliers.

Scaling: subclass 2

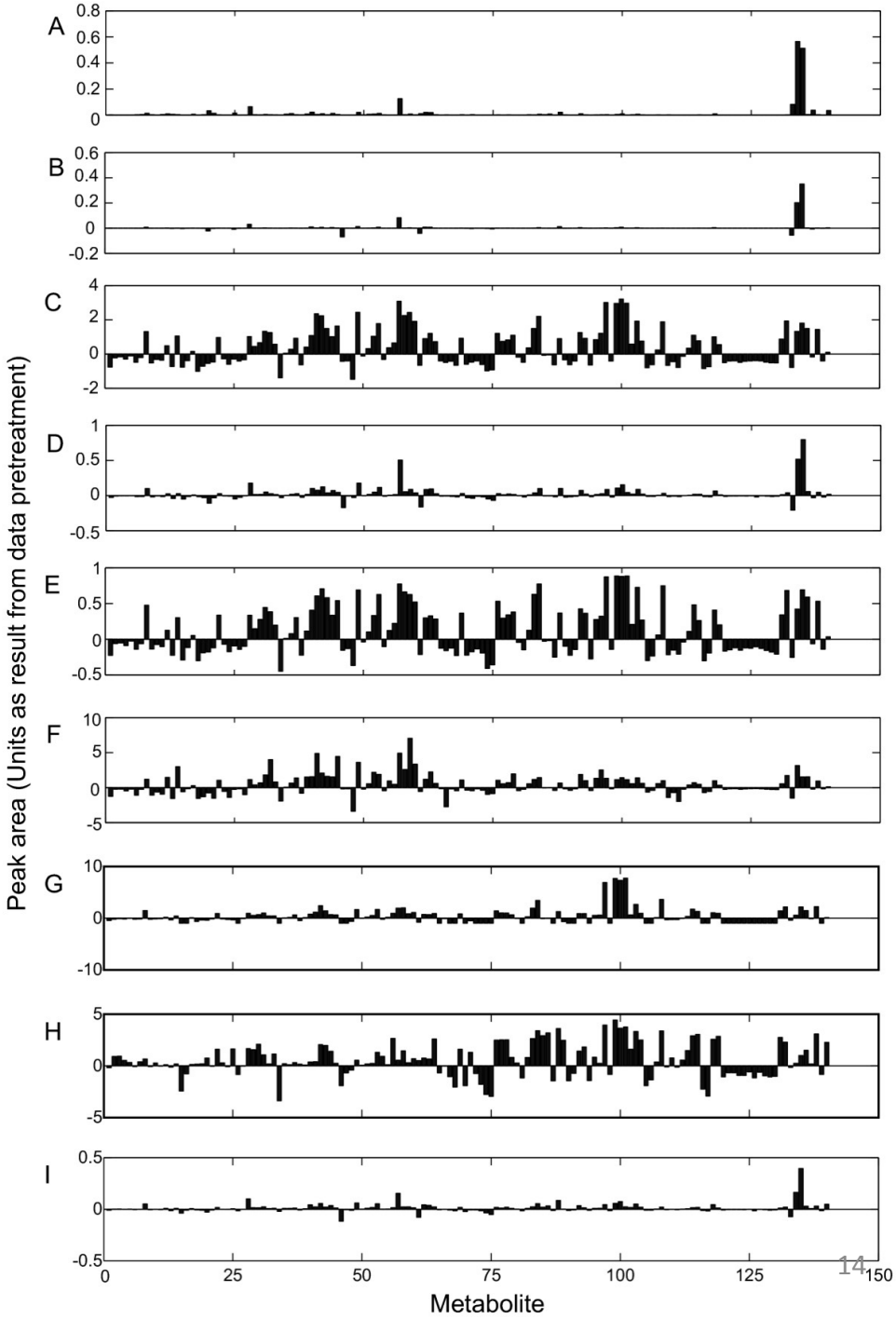
- Use average as scaling factors
  - The resulting values are changes in percentages compared to the mean concentration.
  - The median can be used as a more robust alternative.

Transformation

- Log and power transformation
- Both reduce large values relatively more than the small values.
- Log transformation
  - pros: removal of heteroscedasticity
  - cons: unable to deal with the value zero.
- Power transformation
  - pros: similar to log transformation
  - cons: not able to make multiplicative effects additive

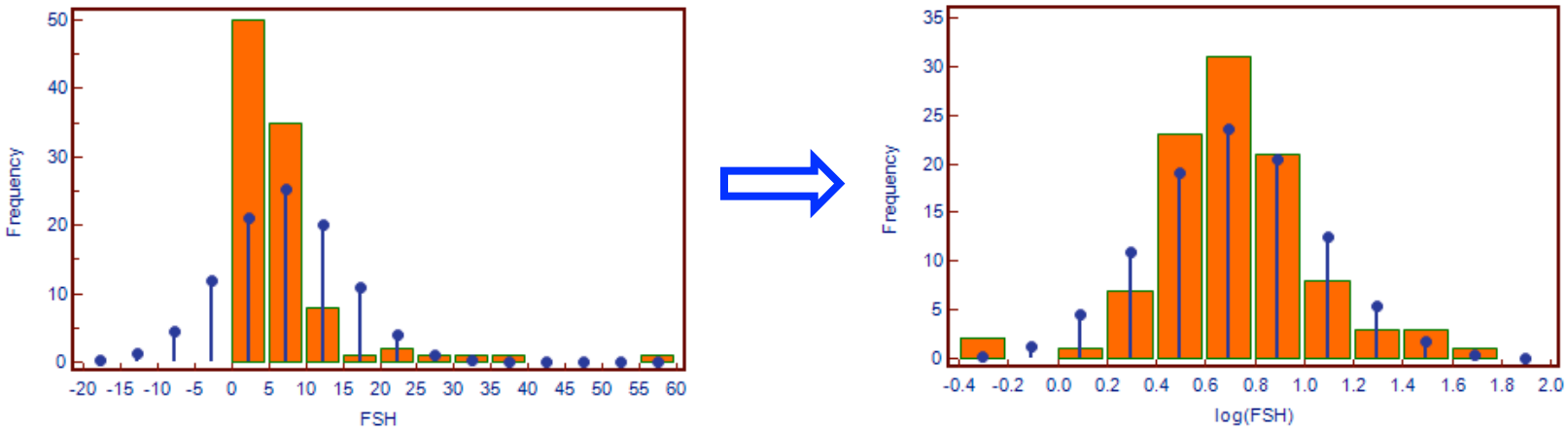
Centering, scaling, transformation

A. original data
B. centering
C. auto
D. pareto
E. range
F. vast
G. level
H. log
I. power



Log transformation, again

- Hard to do useful statistical tests with a skewed distribution.



- A skewed distribution or exponentially decaying distribution can be transformed into a Gaussian distribution by applying a log transformation.

http://www.medcalc.org/manual/transforming_data_to_normality.php

Univariate vs. multivariate analysis

- Univariate analysis examines each variable separately.
  - ***t*-tests**
  - **volcano plot**
- Multivariate analysis considers two or more variables simultaneously and takes into account relationships between variables.
  - PCA: Principle Component Analysis
  - PLS-DA: Partial Least Squares-Discriminant Analysis
- Univariate analyses are often first used to obtain an overview or rough ranking of potentially important features before applying more sophisticated multivariate analyses.

Univariate Statistics

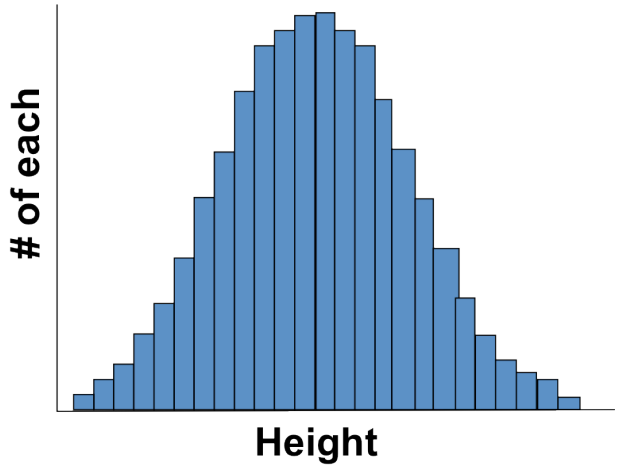
t-test

volcano plot

Univariate statistics

- A basic way of presenting univariate data is to create a frequency distribution of the individual cases.

Due to the Central Limit Theorem, many of these frequency distributions can be modeled as a normal/Gaussian distribution.



Gaussian distribution

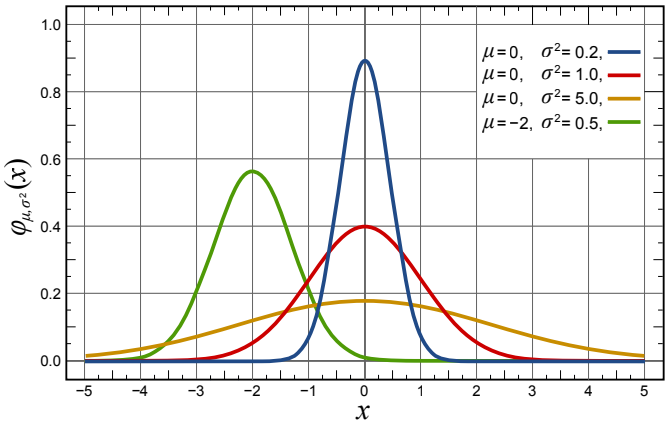
- The total area underneath each density curve is equal to 1.

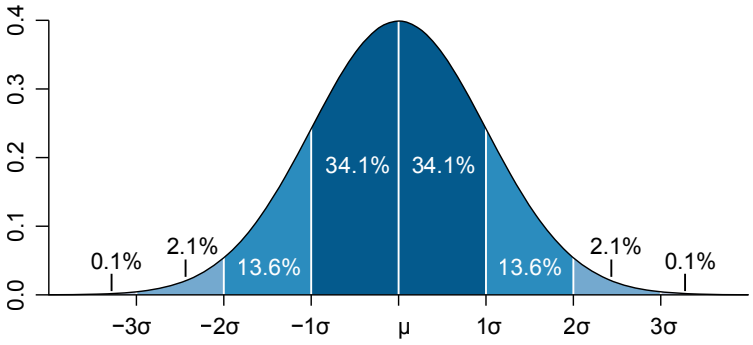
$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\text{mean} = \mu$$

$$\text{variance} = \sigma^2$$

$$\text{standard deviation} = \sigma$$





https://en.wikipedia.org/wiki/Normal_distribution

Sample statistics

Sample mean: $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

Sample variance: $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$

Sample standard deviation: $S=\sqrt{S^2}$

https://en.wikipedia.org/wiki/Normal_distribution

t-test (I)

- One-sample *t*-test: is the sample drawn from a known population?

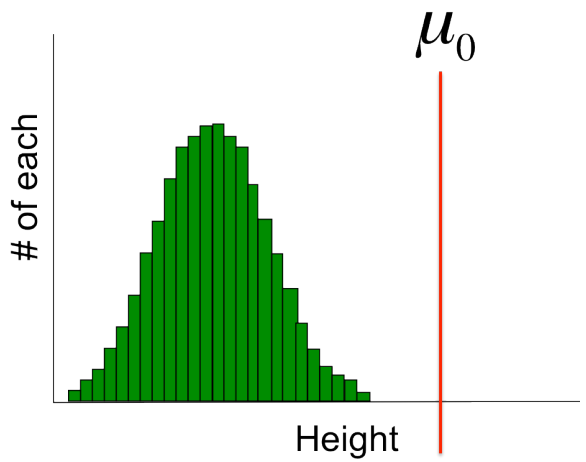
Null hypothesis H_0: $\mu = \mu_0$

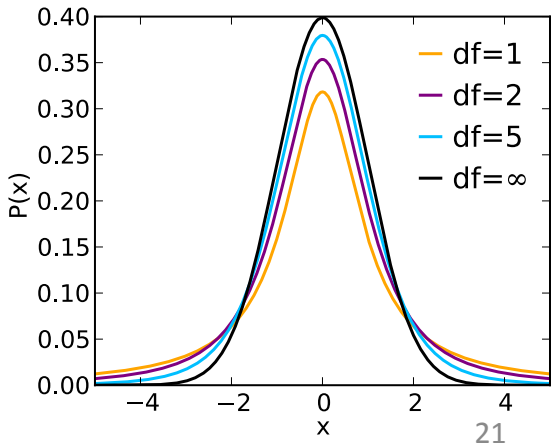
Alternative hypethesis H_1: $\mu < \mu_0$

Test statistic: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Sample standard deviation: $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

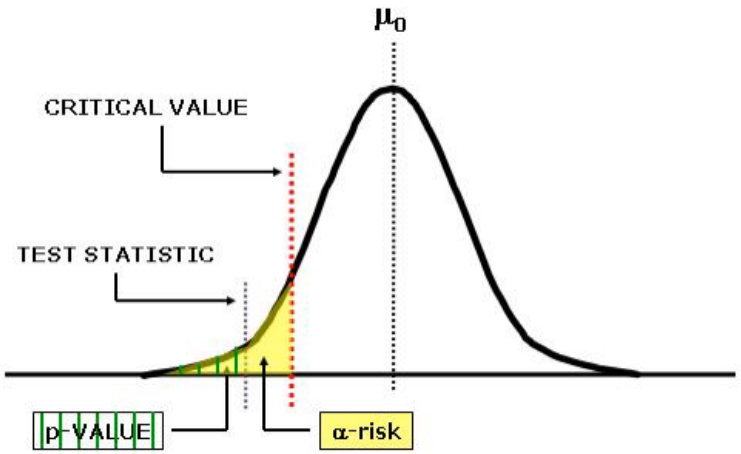
The test statistic *t* follows a student's t distribution. The distribution has *n-1* degrees of freedom.

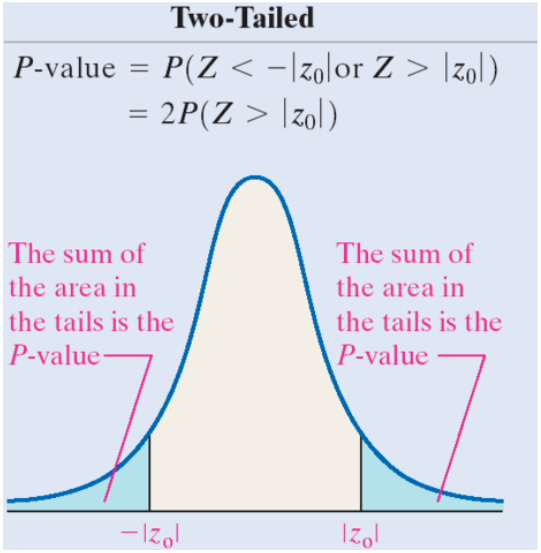


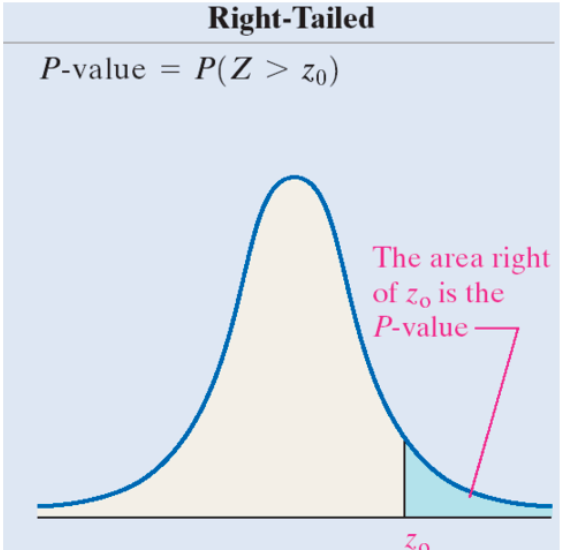


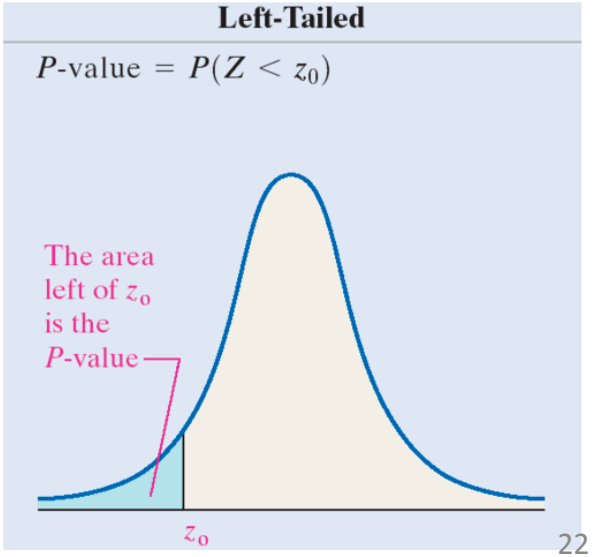
t-test (II): *p*-value

When the null hypothesis is rejected, the result is said to be statistically significant.









t-test (III)

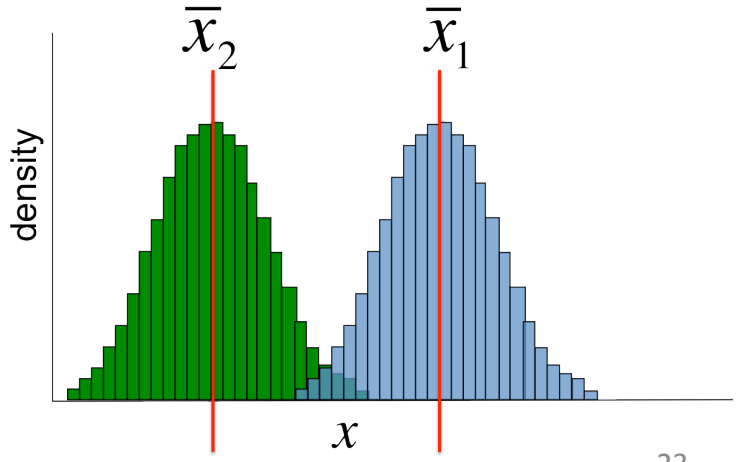
- Two-sample *t*-test: are the two populations different?

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$

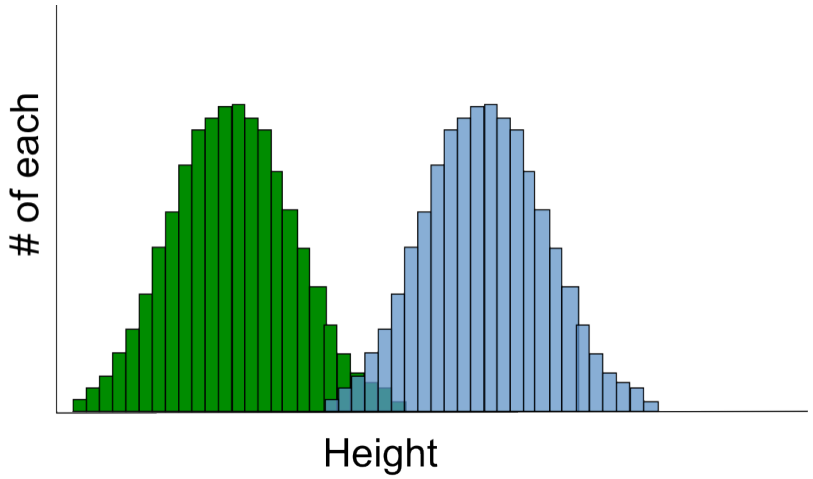
Alternative hypethesis H_1: $\mu_1 - \mu_2 \neq 0$

Test statistic: $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

- The two samples should be independent.



t-test (IV)

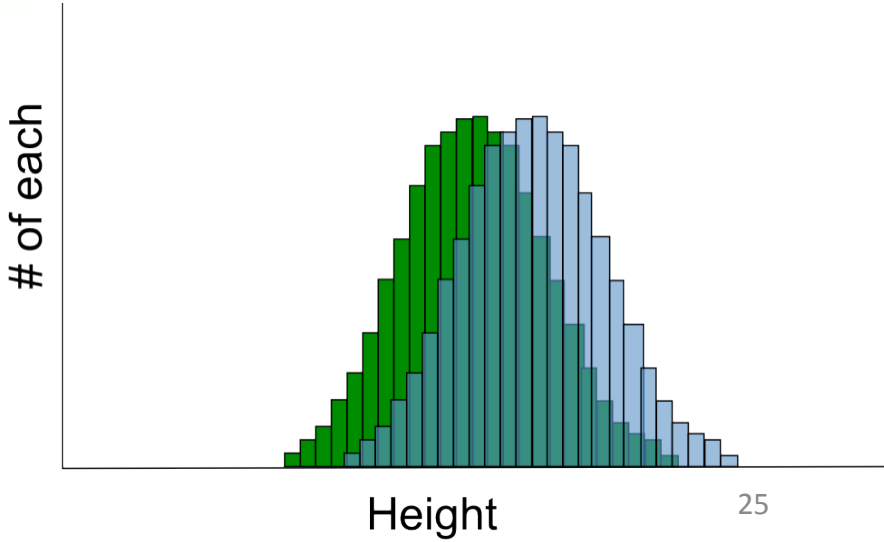


Equivalent statements:

- The *p*-value is small.
- The difference between the two populations is unlikely to have occurred by chance, i.e. is statistically significant.

t-test (V)

- The *p*-value is big.
- The difference between the two populations are said **NOT** to be statistically significant.



t-test (VI)

- Paired *t*-test: what is the effect of a treatment?
- Measurements made on the same individuals before and after the treatment.

Example: Subjects participated in a study on the effectiveness of a certain diet on serum cholesterol levels.

Subject	Before	After	Difference
1	201	200	-1
2	231	236	+5
3	221	216	-5
5	260	243	-17
6	228	224	-4
7	245	235	-10

$$H_0 : \mu_d = 0$$

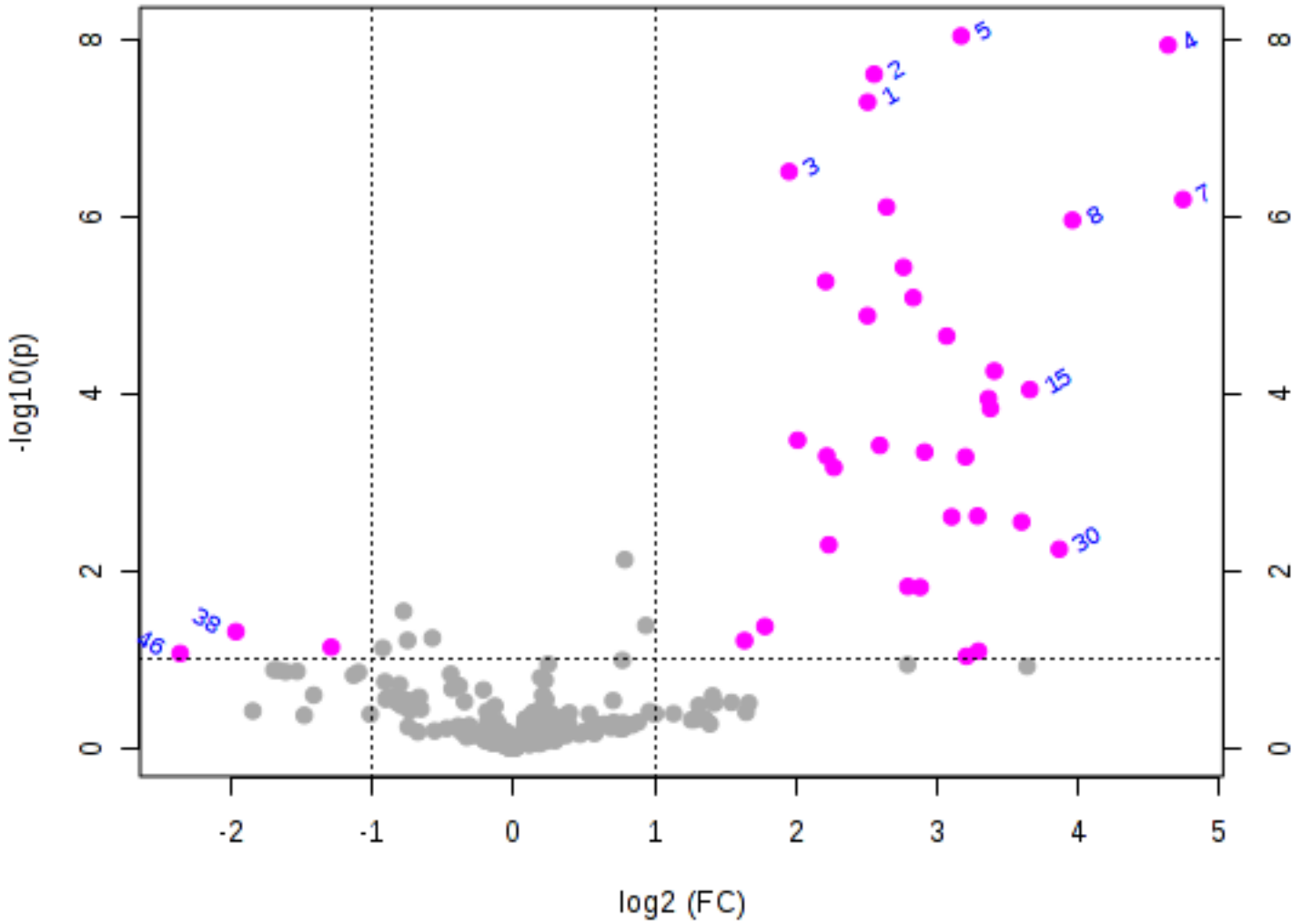
$$H_a : \mu_d \neq 0$$

Test statistic: $t = \dfrac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$

Volcano plot (I)

- Plot fold change vs. significance
- y-axis: negative log of the *p*-value
- x-axis: log of the fold change so that changes in both directions (up and down) appear equidistant from the center
- Two regions of interest: those points that are found towards the top of the plot that are far to either the left- or the right-hand side.

Volcano plot (II)



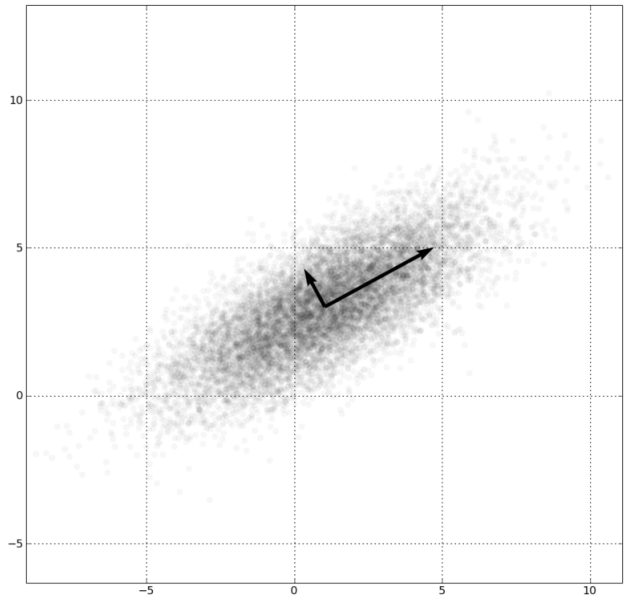
Multivariate statistics

PCA

PLS-DA

PCA (I)

- PCA is a statistical procedure to transform a set of correlated variables into a set of linearly uncorrelated variables.



PCA (II)

- The uncorrelated variables are ordered in such a way that the first one accounts for as much of the variability in the data as possible and each succeeding one has the highest variance possible in the remaining variables.

- These ordered uncorrelated variables are called principle components.

- By discarding low-variance variables, PCA helps us reduce data dimension and visualize the data.

PCA (III)

- The transformation matrix

$$\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n]$$

- The transformation

$$\boldsymbol{Y} = \boldsymbol{P}^T \boldsymbol{X} = \begin{bmatrix} \boldsymbol{p}_1^T \cdot \boldsymbol{X} \\ \boldsymbol{p}_2^T \cdot \boldsymbol{X} \\ \vdots \\ \boldsymbol{p}_n^T \cdot \boldsymbol{X} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1^{new} \\ x_2^{new} \\ \vdots \\ x_n^{new} \end{bmatrix}$$

- $p_1, p_2, \cdots, p_n$ are called the 1st, 2nd, and n^{th} principle components, respectively.

PCA (IV)

- Each original sample is represented by an *n*-dimensional vector:

$$S_{\text{before transformation}} = [x_1, x_2, \cdots, x_n]$$

- After the transformation
  - If all of the principle components are kept, then each sample is still represented by an *n*-dimensional vector:

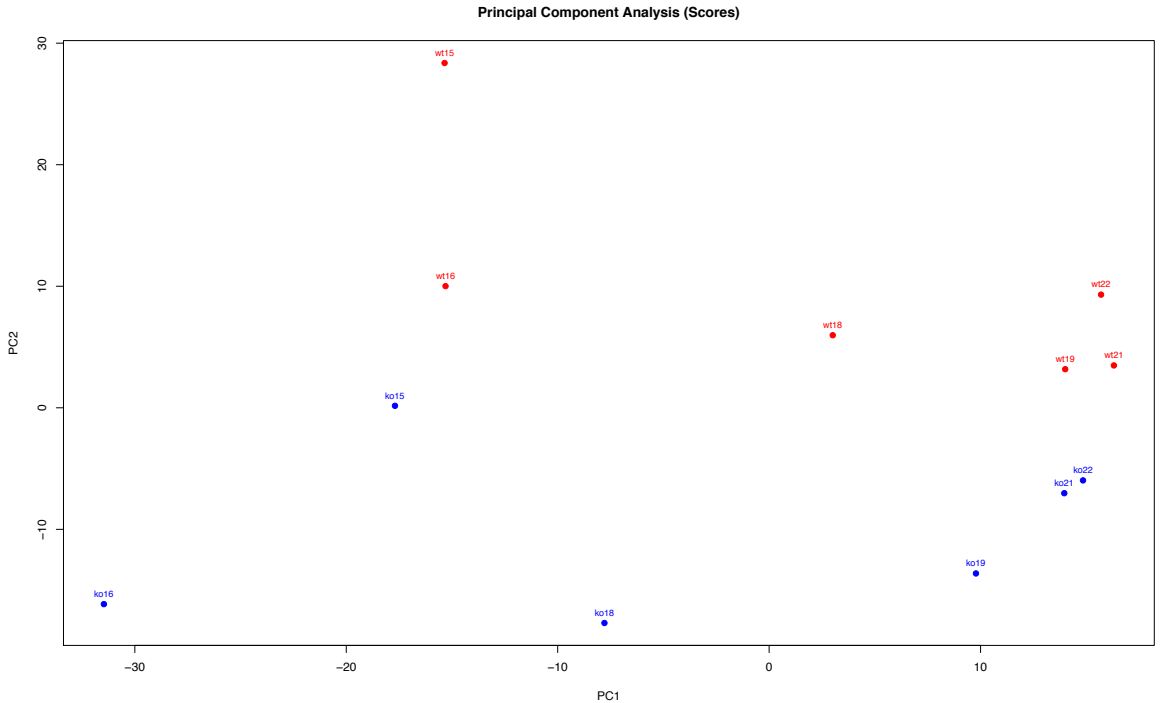
$$S_{\text{after transformation}} = [y_1, y_2, \cdots, y_n]$$

  - If only $m < n$ principle components are kept, then each sample will be represented by an *m*-dimensional vector:

$$S_{\text{after transformation}} = [y_1, y_2, \cdots, y_m]$$

PCA (V)

- *y* are called scores.
- For visualization purpose, *m* is usually chosen to be 2 or 3.
- As a result, each sample will be represented by a 2- or 3-dimenational point in the score plot.



PCA (VI)

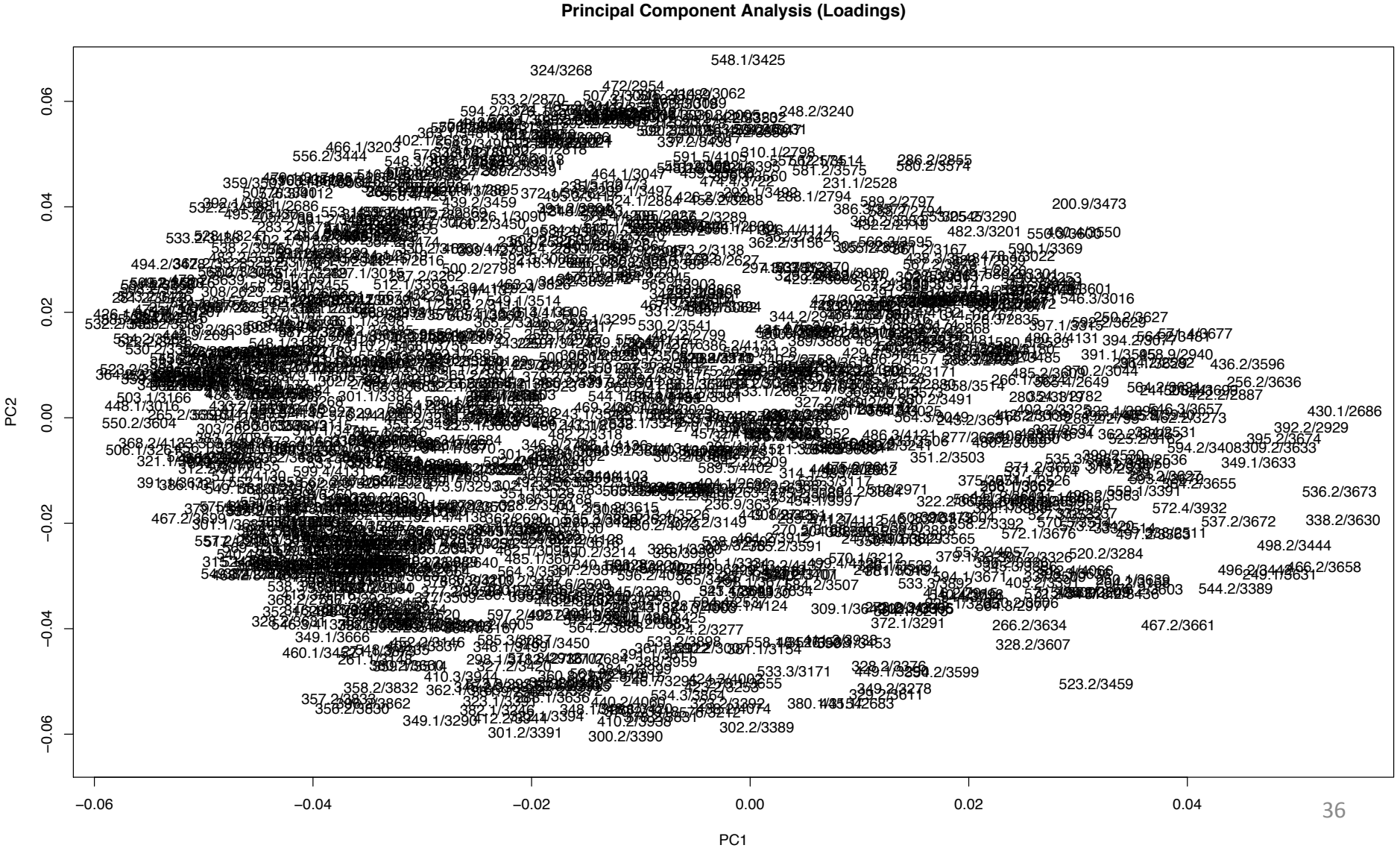
- Loadings

Variables	Components Y_1	Y_2	$\cdots$	Y_n
X_1	p_{11}	p_{12}	$\cdots$	p_{1n}
X_2	p_{21}	p_{22}	$\cdots$	p_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_n	p_{n1}	p_{n2}	$\cdots$	p_{nn}
Eigenvalues	λ_1	λ_2	$\cdots$	λ_n
Eigenvectors	$\boldsymbol{p}_1$	$\boldsymbol{p}_2$	$\cdots$	$\boldsymbol{p}_n$

- $[p_{11}, p_{12}], [p_{21}, p_{22}], \cdots, [p_{n1}, p_{n2}]$ are denoted as points in the loadings plot

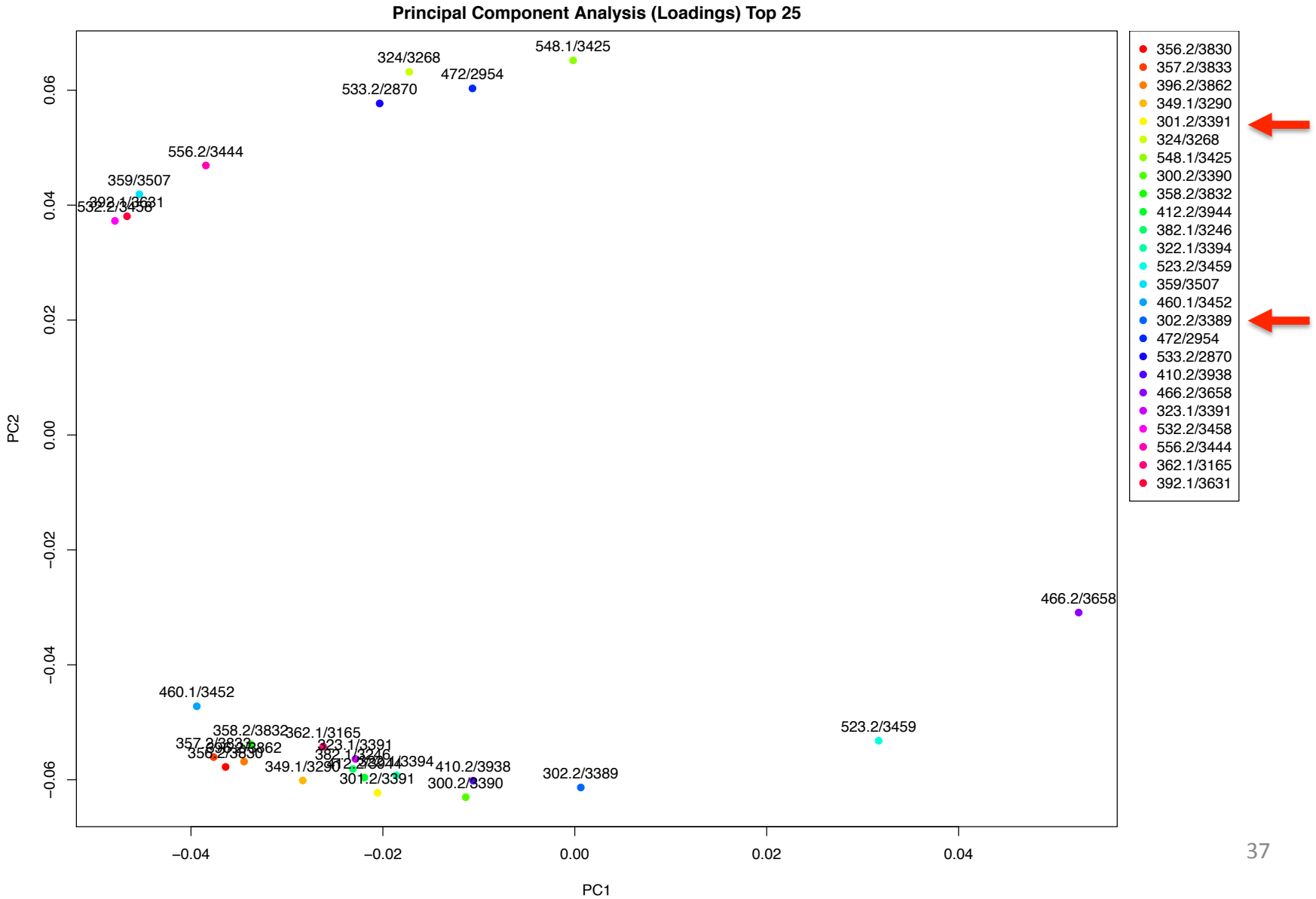
Loadings plot for all variables

PCA (VII)



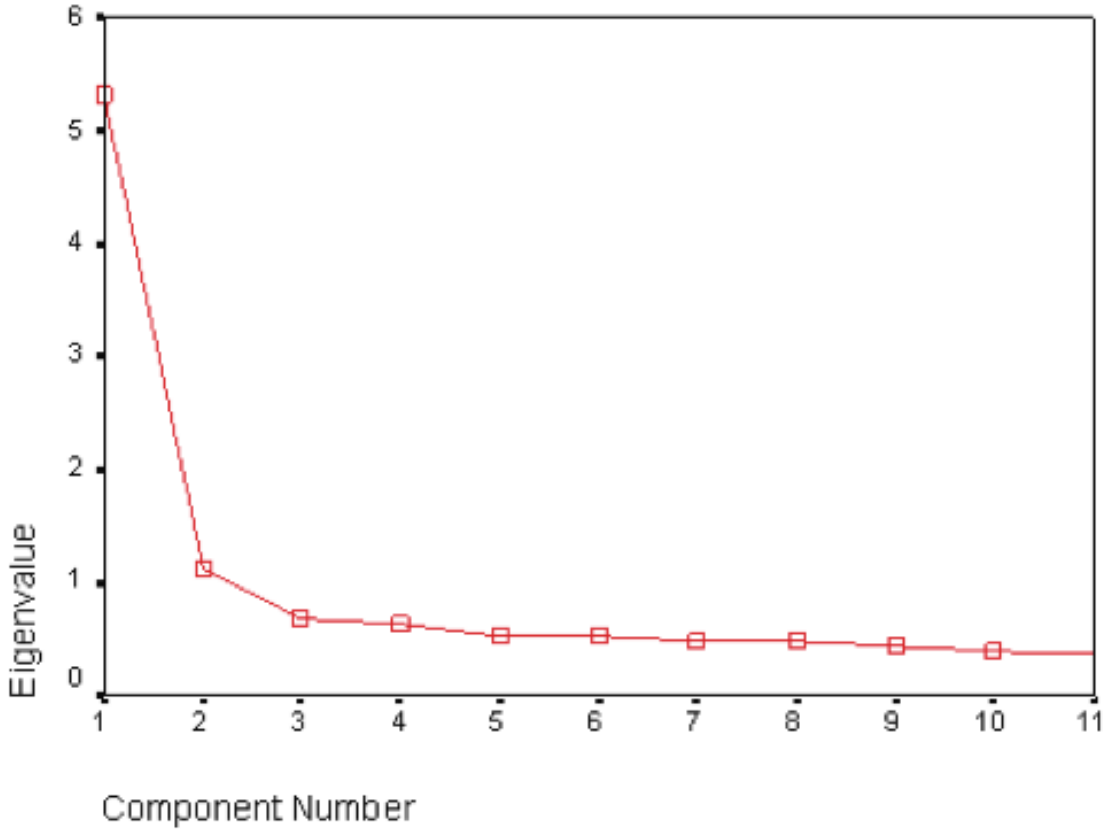
Loadings plot for the top 25 varaibles

PCA (VIII)



PCA (IX)

Scree plot: variance vs. principle component number



PLS-DA (I)

- A supervised method to find a predictive model that describes the direction of maximum covariance between a dataset (X) and the class membership (Y)
- Similar to PCA, the original variables are summarized into much fewer new variables using their weighted averages.
- The new variables are called scores.
- The weighting profiles are called loadings.
- PLS-DA can perform both classification and feature selection.
- Feature importance measure: VIP (Variable Importance in Projection)

PLS-DA (II)

- Interpretation of the model
  - R^2X and R^2Y
    - fraction of the variance that the model explains in the independent (*X*) and dependent variables (*Y*)
    - Range: 0-1
  - Q^2Y
    - measure of the predictive accuracy of the model
    - usually estimated by cross validation or permutation testing
    - Range: 0-1
    - > 0.5 is considered good while > 0.9 is outstanding

PLS-DA (III)

- Note of caution
  - Supervised classification methods are powerful.
  - BUT, they can overfit your data, severely.

Machine Learning

Clustering

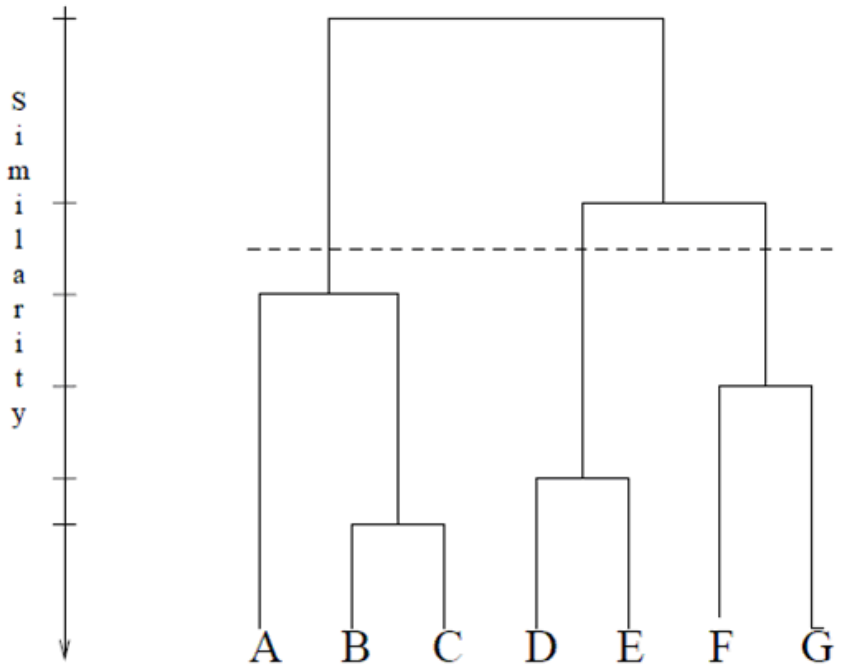
Classification

Clustering

- Group similar objects together
- Any clustering method requires
  - A method to measure similarity/dissimilarity between objects
  - A threshold to decide whether an object belongs to a cluster
  - A way to measure the distance between two clusters
- Common clustering algorithms
  - *K*-means
  - Hierarchical
  - Self-organizing map
- Unsupervised machine learning techniques

Hierarchical clustering (I)

1. Find the two closest objects and merge them into a cluster
2. Find and merge the next two closest objects (or an object and a cluster, or two clusters)
3. Repeat step 2 until all objects have been clustered

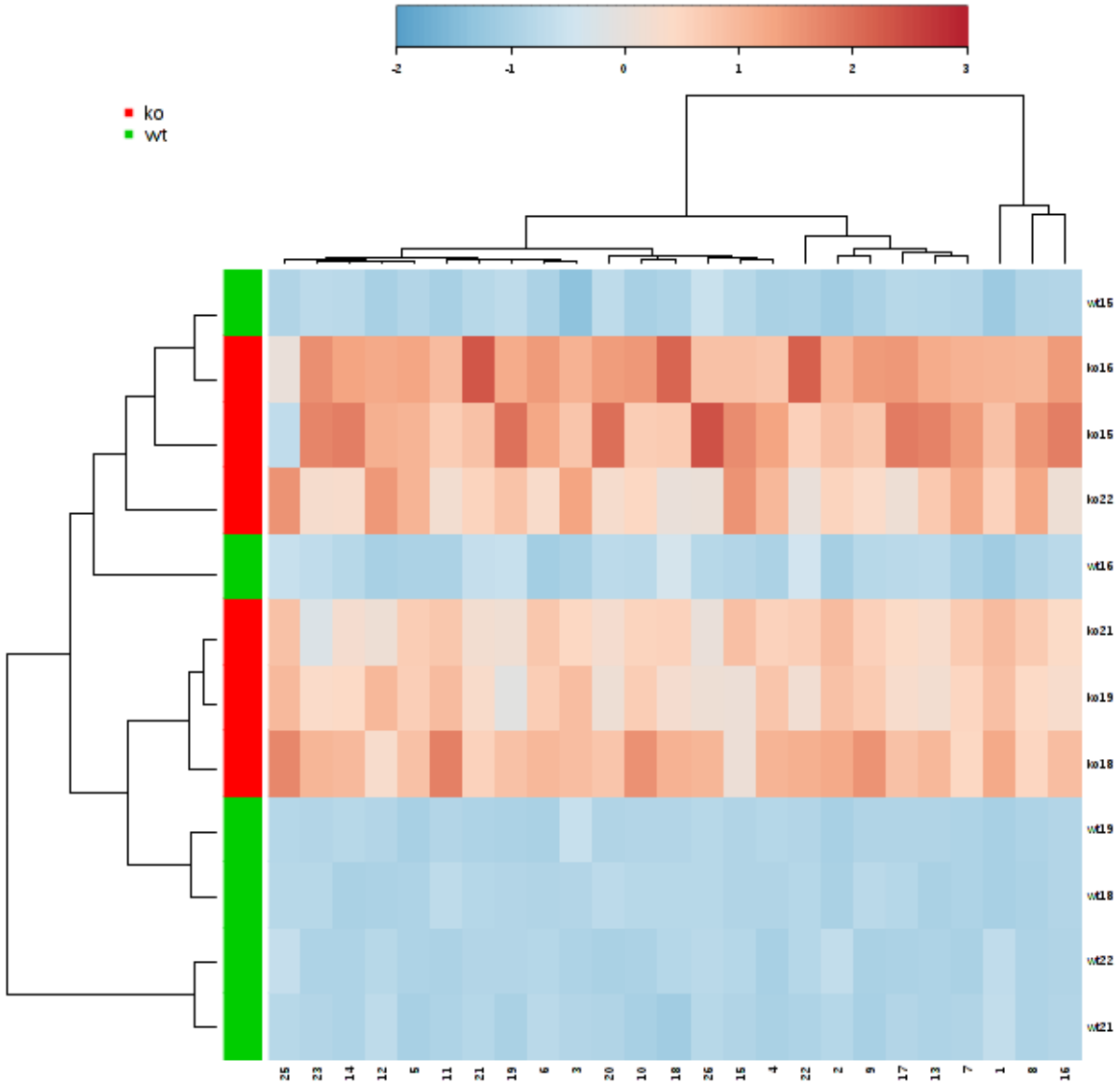


Hierarchical clustering (II)

- Methods to measure similarity between objects
  - Euclidean, Manhattan
  - Pearson correlation
  - Cosine similarity
- Linkage: ways to measure the distance between two clusters

single complete centroid average

Hierarchical clustering (III)



Classification

- Use a training set of correctly-identified observations to build a predictive model
- Predict to which of a set of categories a new observation belongs
- Supervised machine learning
- Methods
  - Linear discriminant analysis
  - Support vector machine (SVM)
  - Artificial neural network (ANN)
  - *k*-nearest neighbor
  - Random forest
  - PLS-DA

Software Packages

MetaboAnalyst

XCMS

For in-depth statistical analysis and data interpretation, please make an appointment with a biostatistician.

Thank you!