Bioinformatical analysis of omics expression data Part 2

Dr. Michael Turewicz[1,2]

[1]Institut für Klinische Biochemie und Pathobiochemie,
Deutsches Diabetes-Zentrum, Leibniz-Zentrum für Diabetesforschung an der Heinrich-Heine-Universität, Düsseldorf, Deutschland
[2]Deutsches Zentrum für Diabetesforschung (DZD), München-Neuherberg, Deutschland

Course schedule

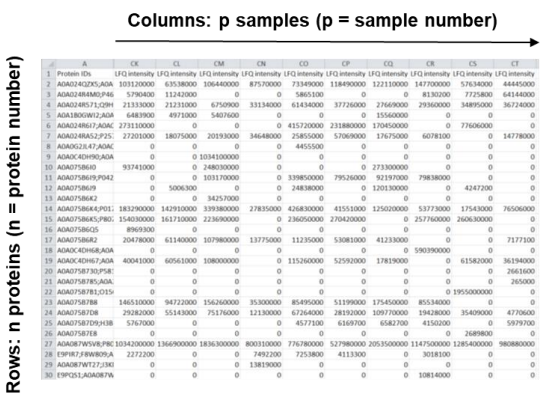
DDZ Deutsches Diabetes-Zentrum

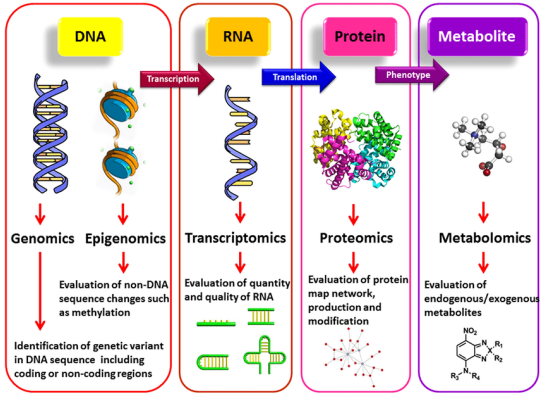
- Part 1 (25.10.23)
  - Introduction (omics, example data, programming)
  - Data preprocessing (data inspection, normalization, missing values)
  - Exercises: R programming tutorial (part 1)
- **Part 2 (08.11.23)**
  - **Differential expression analysis (statistics, volcano plot)**
  - **Exercises: R programming tutorial (part 2)**
- Part 3 (15.11.23)
  - Machine learning I: Clustering (clustering, PCA)
- Part 4 (22.11.23)
  - Overrepresentation analysis (GO, Reactome)
- Part 5 (29.11.23)
  - Network analysis (STRING, Cytoscape)
- Part 6 (06.12.23)
  - Machine learning II: Classification algorithms

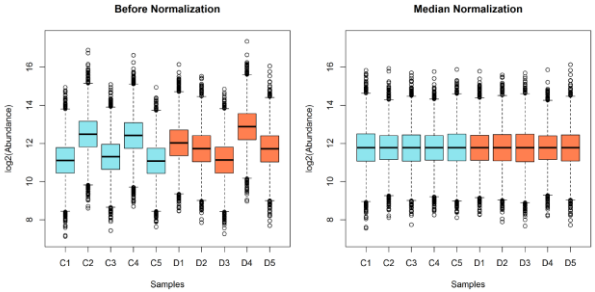
Recap of previous part

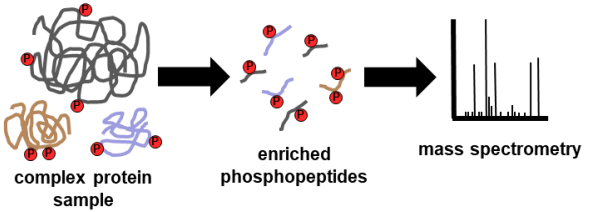
DDZ Deutsches Diabetes-Zentrum

- **Omics technologies**
  - Information at various bimolecular levels
  - High-throughput measurements (proteomics → all proteins)
  - Multiple samples needed
  - Problems: large data, “n << p”-problem, noisy data, missing values
- **Example data**
  - Phosphoproteomics of murine adipocytes
  - 4 samples before & 4 after insulin stimulation
- **Data preprocessing**
  - Inspection / removing meaningless data
  - Normalization
  - Missing value imputation
- **Programming (example: R)**
  - Needed for profound & flexible data analysis
  - R code in Google Colab & Rstudio
  - R tutorial part 1: basic commands & functions, vectors, matrices, data frames









Finding most interesting candidates

DDZ Deutsches Diabetes-Zentrum

- High-throughput measurements of many thousands of expression values of biomolecules across samples…
- Which of them are “interesting”?
- Depends on study design
  - Often comparing target group(s) of samples with control group(s)
  - E.g. search for differences between patients and “healthy” individuals
- Goal: list of candidates for whom we have **clearly proven differential group-specific values**

Columns: p samples (p = sample number)

Rows: n proteins (n = protein number)

	A	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT
1	Protein IDs	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity	LFQ intensity
2	A0A024QZX5;A0A	103120000	63538000	106440000	87570000	73349000	118490000	122110000	147700000	57634000	44445000
3	A0A024R4M0;P46	5790400	11242000	0	0	5865100	0	0	8130200	7725800	64144000
4	A0A024R571;Q9H	21333000	21231000	6750900	33134000	61434000	37726000	27669000	29360000	34895000	36724000
5	A0A1B0GWI2;A0A	6483900	4971000	5407600	0	0	0	15560000	0	0	0
6	A0A024R6I7;A0A	273110000	0	0	0	415720000	231880000	170450000	0	77606000	0
7	A0A024RA52;P25	27201000	18075000	20193000	34648000	25855000	57069000	17675000	6078100	0	14778000
8	A0A0C4DJL47;A0A	0	0	0	0	4455500	0	0	0	0	0
9	A0A0C4DH90;A0A	0	0	1034100000	0	0	0	0	0	0	0
10	A0A075B6I0	93741000	0	248030000	0	0	0	273300000	0	0	0
11	A0A075B6I9;P042	0	0	103170000	0	339850000	79526000	92197000	79838000	0	0
12	A0A075B6J	0	5006300	0	0	24838000	0	120130000	0	4247200	0
13	A0A075B6K7	0	0	34257000	0	0	0	0	0	0	0
14	A0A075B6	183290000	142910000	339380000	27835000	426830000	41551000	125020000	53773000	17543000	76506000
15	A0A075B6	154030000	161710000	223690000	0	236050000	270420000	0	257760000	260630000	0
16	A0A075B6Q5	8969300	0	0	0	0	0	0	0	0	0
17	A0A0	20478000	61140000	107980000	13775000	11235000	53081000	41233000	0	0	7177100
18	A0A0C4DH68;A0A	0	0	0	0	0	0	0	590390000	0	0
19	A0A	40041000	60561000	108000000	0	115260000	52592000	17819000	0	61582000	36194000
20	A0A075B7	0	0	0	0	0	0	0	0	0	2661600
21	A0A075B7	0	0	0	0	0	0	0	0	0	265000
22	A0A075B7	0	0	0	0	0	0	0	0	1955000000	0
23	A0A075B7B8	146510000	94722000	156260000	35300000	85495000	51199000	175450000	85534000	0	0
24	A0A075B7D8	29282000	55143000	75176000	12130000	67264000	28192000	109770000	19428000	35409000	4770600
25	A0A075B7D9	5767000	0	0	0	4577100	6169700	6582700	4150200	0	5979700
26	A0A075B7E8	0	0	0	0	0	0	0	0	2689800	0
27	A0A087WSV8	1034200000	1366900000	1836300000	800310000	776780000	527980000	2053500000	1147500000	1285400000	980880000
28	E9PIR7;F8W8	2272200	0	0	7492200	7253800	4113300	0	3018100	0	0
29	A0A087WT27;	0	0	0	13819000	0	0	0	0	0	0
30	E9PQS1;A0A087W	0	0	0	0	0	0	0	10814000	0	0

Finding most interesting candidates

DDZ Deutsches Diabetes-Zentrum

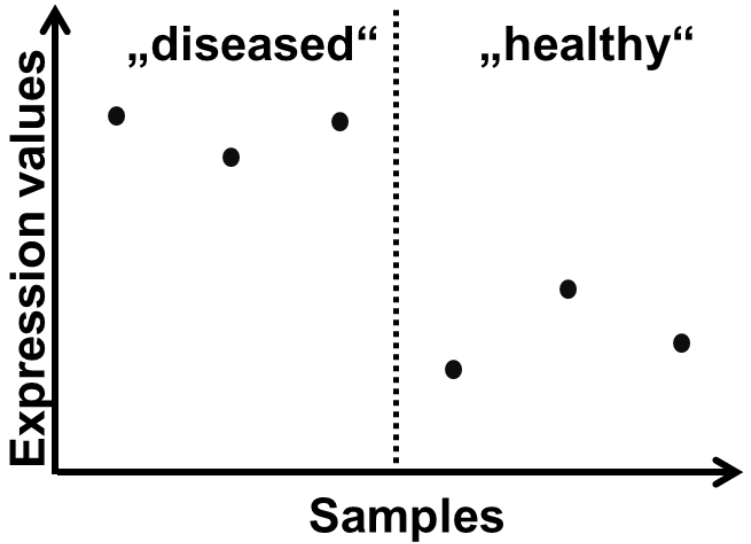
Protein expression values

Diseased samples

Healthy samples

	D1	D2	D3	H1	H2	H3
P1	50	100	50	100	50	100
P2	300	250	270	90	120	110
P3	500	600	550	590	490	610
…	…	…	…	…	…	…

Proteins



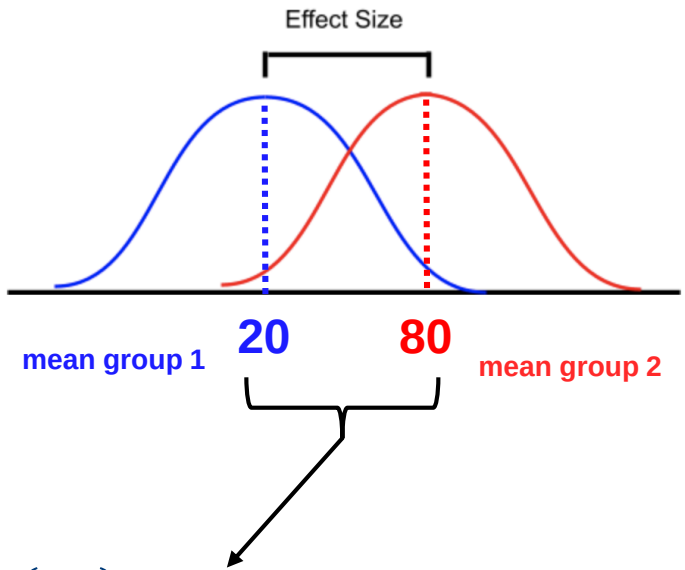
Criteria for interesting candidates

- High difference between group means → **fold change**
- Clear / statistically significant difference between group means → **p-value**

Fold change (FC)

DDZ Deutsches Diabetes-Zentrum

- Measure for effect size (biological relevance)
- Describes how much a quantity changes from an initial value ($\mathbf{A}$ = mean of group 1) to a target value ($\mathbf{B}$ = mean of group 2)
- **Definition:** $FC = B/A$
- Calculation on non-log-transformed values (in contrast to p-values)
- For analysis & visualization often log2-fold changes used



Examples:

$\mathrm{A} = 20, \mathrm{B} = 80 \Rightarrow FC = 4 \Rightarrow \log_2(FC) = 2$

$\mathrm{A} = 20, \mathrm{B} = 5 \Rightarrow FC = 1/4 \Rightarrow \log_2(FC) = -2$

log2(FC) is symmetrical!

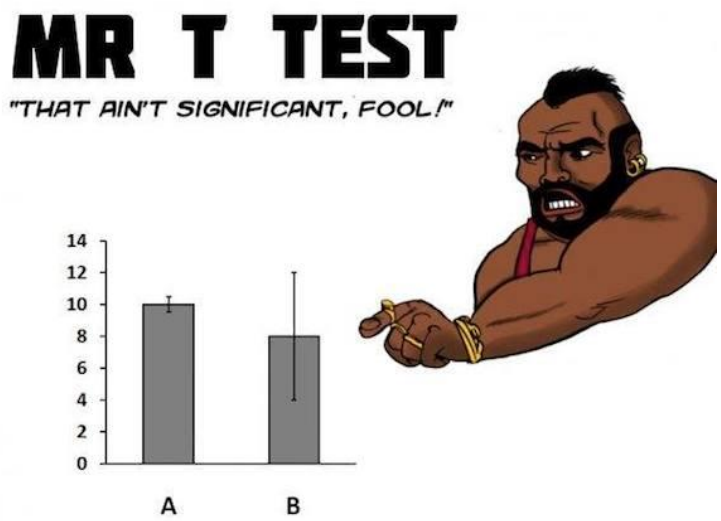
t-test

DDZ Deutsches Diabetes-Zentrum

In general:
Statistical tests make a justified decision about the validity or invalidity of a hypothesis (the so called **null hypothesis H_0**) using available observations.

Most important example: **t-test**

„Do means of two experimental groups differ significantly (considering sample size & variance)?“



Variations:

- **Two sample** / one sample
- **Unpaired** / paired design
- **unequal** / equal variances
- **Two-sided** / one-sided

Default: unpaired, two-sample & two-sided t-test with unequal variances

t-test

DDZ Deutsches Diabetes-Zentrum

- n **samples for group 1:** $x_1, \dots, x_n$
- m **samples for group 2:** $y_1, \dots, y_m$

Test theory:

Null hypothesis H_0: there is no difference between means

Alternative hypothesis H_1 : there is a difference between means

$\mu_x = \mu_y$

$\mu_x \neq \mu_y$

Type I error: reject H_0 when it is true (probability α)

Type II error: do not reject H_0 when it is false (probability β)

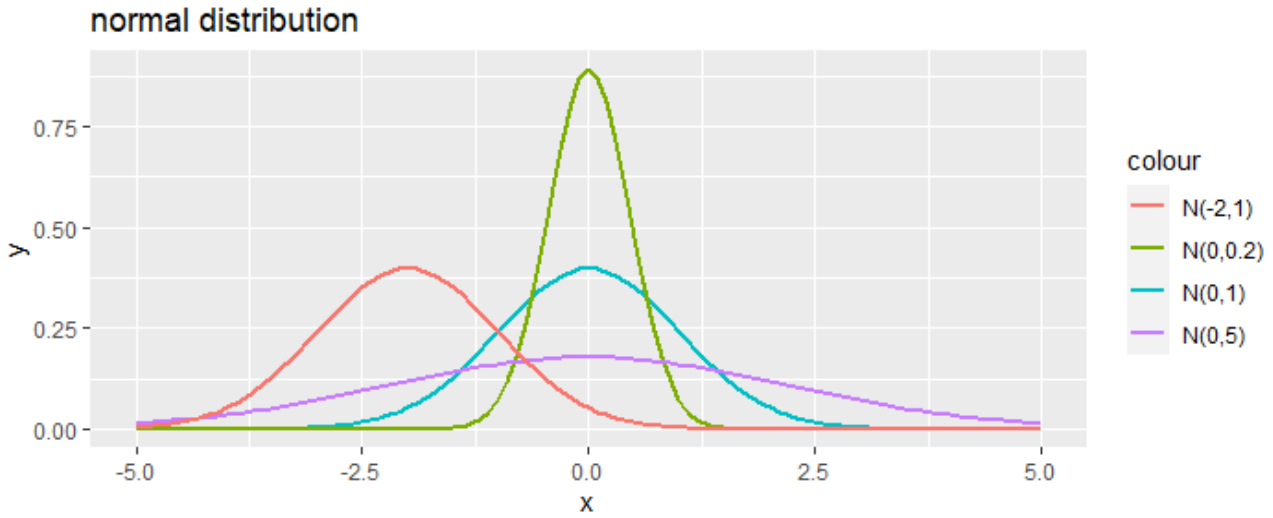
Power: Probability to reject H_0 when it is false, i.e. ability to detect true differences (probability 1- β)

Truth	Test decision	
	do not reject H_0	reject H_0
H_0 true	☺	☹ type I error (α)
H_0 false (H_1 true)	☹ type II error (β)	☺ power (1-β)

t-test

DDZ
Deutsches Diabetes-Zentrum

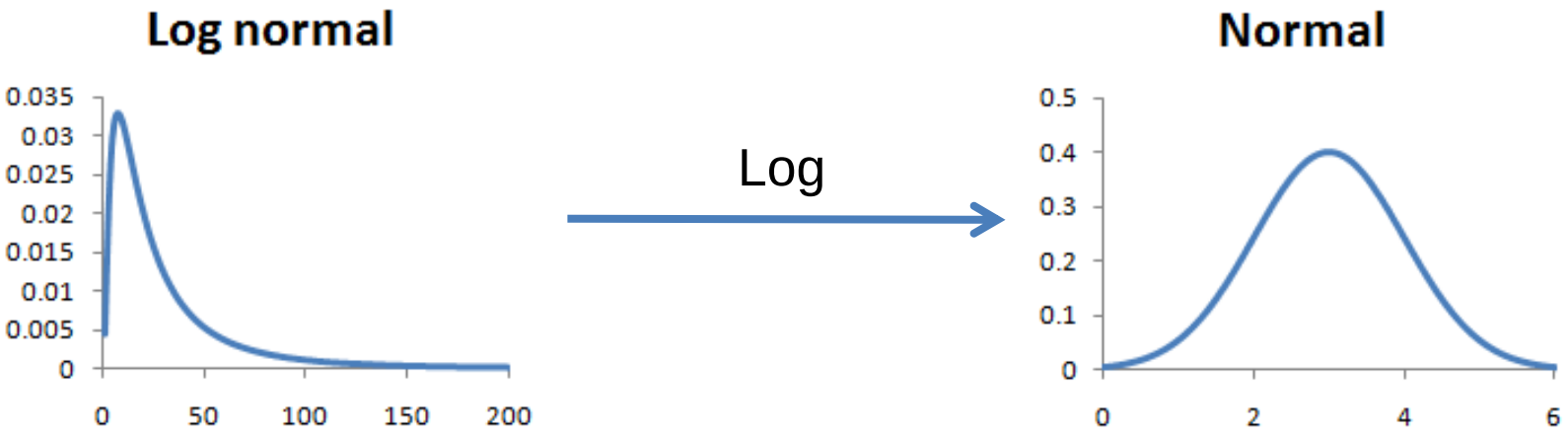
- t-test assumes that the data follow a normal distribution (in each group)
- $f(x) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \cdot \exp(-\frac{(x-\mu)^2}{2\sigma^2})$



t-test

DDZ Deutsches Diabetes-Zentrum

- Omics values are often not normal, but **log-normal distributed**
- → log-transformed data follows a normal distribution
- → omics data are usually log-transformed before t-test calculation



t-test

DDZ Deutsches Diabetes-Zentrum

- t-test statistic (for the special case $m = n$, i.e. equal group sizes):

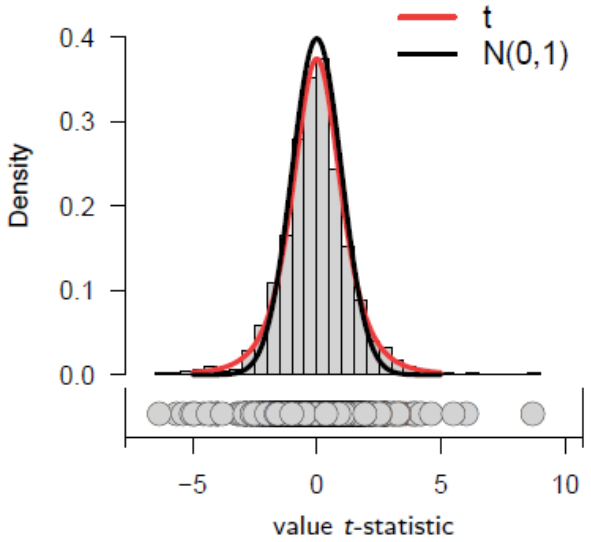
$$t = \frac{\delta}{\sqrt{\frac{1}{n}\left(\sigma_X^2 + \sigma_Y^2\right)}}$$

- $\delta = \bar{x} - \bar{y}$: difference of mean values of the two groups
- n: number of samples in each group
- σ_X^2, σ_Y^2: variances in the two groups

- Question: If H_0 is true, what values for t are typical?
- $\rightarrow$ t follows a t-distribution

	x_1	x_2	x_3	y_1	y_2	y_3	t(x,y)$t
1	0.33	-0.19	-2.23	-0.18	1.95	-0.19	-1.16
2	1.51	0.50	0.48	1.41	-0.28	-0.42	0.87
3	0.46	1.22	0.24	1.45	-0.65	1.08	0.02

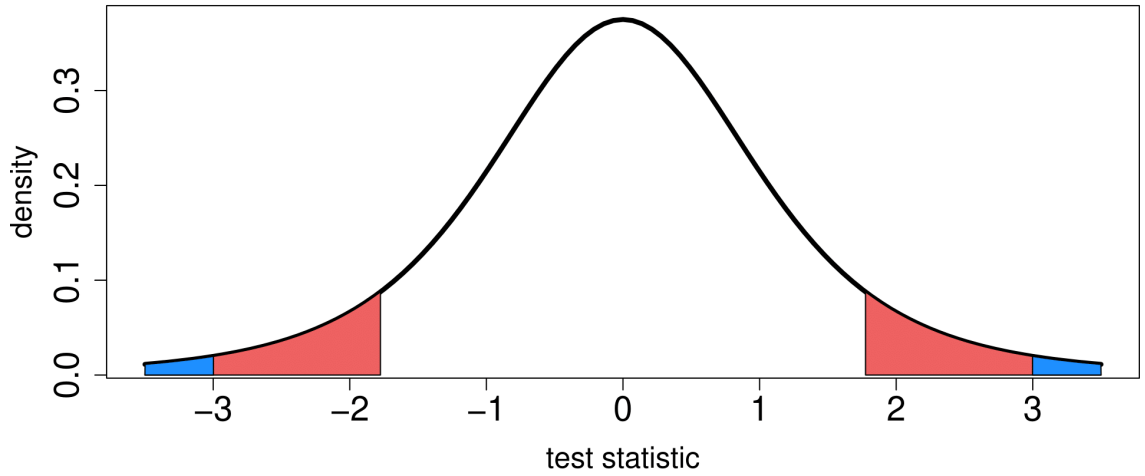
Random numbers from a normal distribution for which H_0 is true



t-test: p-value

DDZ Deutsches Diabetes-Zentrum

- Reject H_0, if observed value of t is 'too extreme'
- Area under the density curve equals a probability of 1
- ***The p-value is the probability that under H_0 the value of the test statistic is at least as extreme as the one calculated from the data***
- H_0 is rejected, if the p-value is less than a predefined (!) significance level α (usually $\alpha = 0.05$)



t = 1.77
→ p = 0.15

t = 3.00
→ p = 0.04

t-test: power

DDZ Deutsches Diabetes-Zentrum

$$t = \frac{\delta}{\sqrt{\frac{1}{n}\left(\sigma_X^2 + \sigma_Y^2\right)}}$$

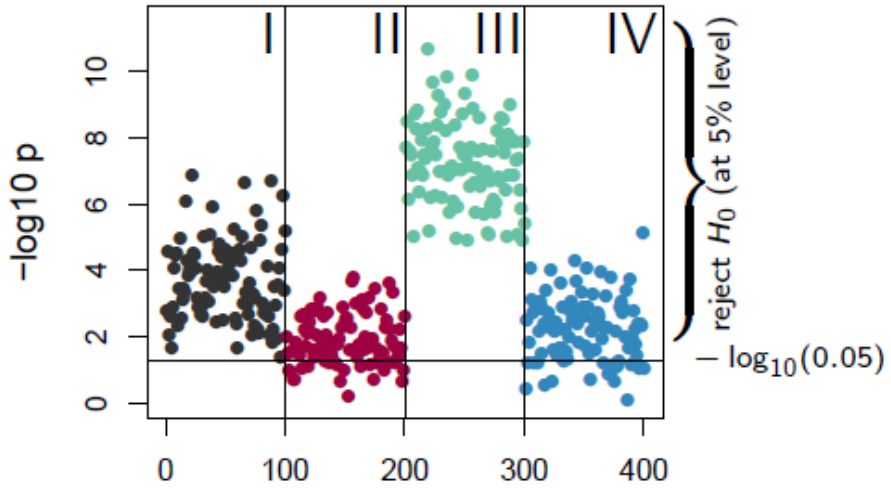
Different situations, where H_0 is false:

	Setting		draw random numbers from n	δ	σ	Power *est*	Power *theo*
I			**10**	**2**	**1**	1.00	0.99
II	$n \downarrow$		5	2	1	0.77	0.79
III	$\delta \uparrow$		10	4	1	1.00	1.00
IV	$\sigma \uparrow$		10	2	1.5	0.79	0.80

Power increases for:

- Higher sample sizes
- Higher group mean differences
- Smaller standard deviation (variance)

Truth	Test decision	
	do not reject H_0	reject H_0
H_0 true	☺	☹ type I error (α)
H_0 false (H_1 true)	☹ type II error (β)	☺ power (1-β)



Multiple testing problem

DDZ Deutsches Diabetes-Zentrum

- Usually thousands of tests are performed per omics data set
- Problem: Multiple tests → inflation of false positives
- E.g., in 100 tests, at least 1 test has $p < 0.05$, although there is **no** difference between the means !!!

$$\pi = 1 - (1 - \alpha)^m$$

Number of tests (m)	Significance level α per test	Cumulated type I error (= π)
1	0.05	0.05
2	0.05	0.0975
3	0.05	0.142625
5	0.05	0.2262191
10	0.05	0.4012631
100	**0.05**	**0.9940795**
1,000	0.05	1
10,000	0.05	1

P-value adjustment

DDZ Deutsches Diabetes-Zentrum

- m: number of conducted tests

Family-wise error rate (FWER):

- Probability of obtaining at least one false positive, if H_0 is true for all tests
- P-values can be corrected by **Bonferroni method** (multiplication of p-values with m)
  ⇒ FWER ≤ 0.05

False discovery rate (FDR):

- FDR = $\frac{\#\ false\ positives}{\#\ all\ significant\ proteins}$
- Correct p-values by **Benjamini-Hochberg method**: (⇒ FDR ≤ 0.05)

$$\tilde{p}_{(m)} = p_{(m)}$$

$$\tilde{p}_{(i)} = \min\left(p_{(i)} \cdot \frac{m}{i}, \tilde{p}_{(i+1)}\right)$$

P-value adjustment

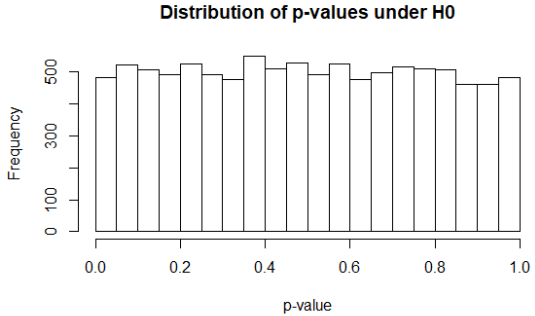
DDZ Deutsches Diabetes-Zentrum

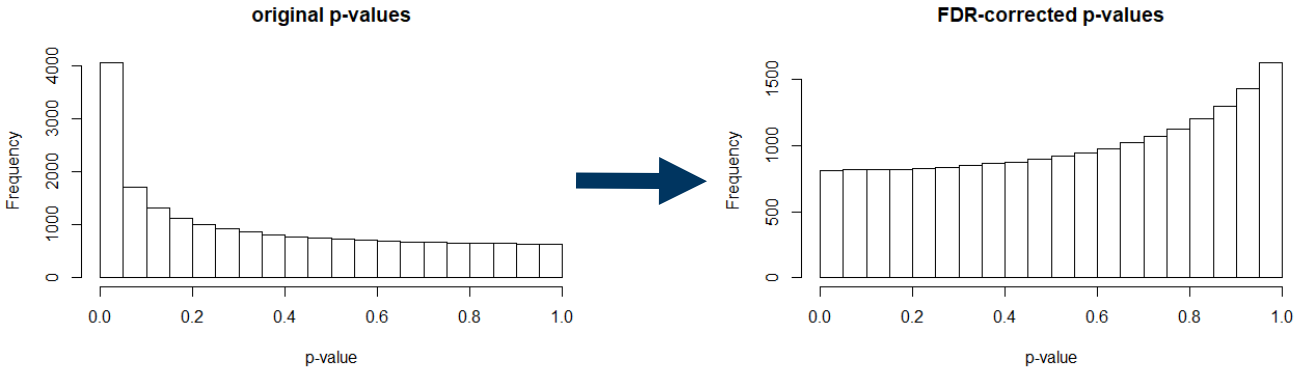
- p-values are enlarged

→ less proteins are significant, i.e. have a p-value below 0.05

→ remaining significant proteins are more reliable

Under H0, p-values follow a continuous uniform distribution between 0 and 1



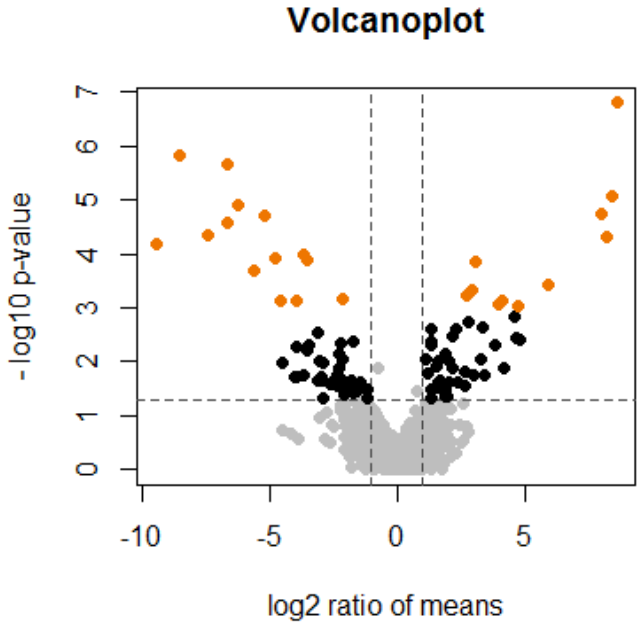


Volcano plot

Deutsches Diabetes-Zentrum

- For each protein, you have now calculated a p-value and a fold change
- Volcano plot shows $-\log_{10}(\mathrm{p-value})$ vs. $\log_2(\mathrm{FC})$

- Promising biomarker candidates usually have a small p-value (≤ 0.05) and a large fold change ($|FC| \geq 2$)
- i.e. these proteins show up in the left and right upper corners of the volcano plot
- Usually, the original p-values are used, proteins that are significant after FDR-correction can be coloured



Goal: list of candidates

DDZ Deutsches Diabetes-Zentrum

- E.g., top 20 list of phosphopeptides sorted according to p-value and FC

Peptide ID	Modifications	FC	p-value	Adj. p-value	Basal1	Basal2	Basal3	Basal4	Insulin1	Insulin2	Insulin3	Insulin4
P31750_peptide2	P31750 2xPhospho	87.759	1.00E-17	1.80E-16	630424	NA	NA	NA	50854105.7	56748133.2	53937850.6	66060762.5
Q91V92_peptide5	Q91V92 1xPhospho	76.389	1.00E-17	1.80E-16	365200.875	NA	NA	NA	18304132.6	28179874.2	27617663.7	30173615.6
Q60823_peptide2	Q60823 2xPhospho	61.726	1.00E-17	1.80E-16	391435.188	5295027.65	2064519.27	NA	98298839.2	122817396	132223911	135915099
Q9Z120_peptide1	Q9Z120 1xPhospho	59.368	1.00E-17	1.80E-16	NA	NA	368242.827	NA	25278608.6	16221798.1	22371727.7	21363576.2
P62754_peptide1	P62754 3xPhospho	58.029	1.00E-17	1.80E-16	132964376	115745575	88336907.5	70498239.6	5273357626	5482982170	6529024054	5500987554
Q3UPF5_peptide1	Q3UPF5 3xPhospho	55.58	1.00E-17	1.80E-16	1321675.5	NA	NA	1216819.49	46805297	64310257.9	78871339.2	77251959.9
Q61409_peptide11		51.747	1.00E-17	1.80E-16	1054045.75	994530.441	1094527.15	NA	50056270.5	42749707.9	71931937.2	58231950.5
Q60876_peptide6	Q60876 3xPhospho	48.543	1.00E-17	1.80E-16	2450149.52	3887475.35	1218929.51	1192042.64	90657018.4	106706158	77068720	65951887.9
Q8BSK8_peptide5	Q8BSK8 1xPhospho	47.575	1.00E-17	1.80E-16	438437.25	NA	NA	NA	19652045.1	19467464.5	23411671.4	22139707.3
Q9QZQ1_peptide2	Q9QZQ1 1xPhosph	45.662	1.00E-17	1.80E-16	1779093	2028343.57	2630561.25	2326267.68	90538602.2	107507126	96409359.5	104953921
Q61409_peptide13	Q61409 1xPhospho	45.139	1.00E-17	1.80E-16	3462868.75	5801882.22	4316337.31	3943713.43	218861197	212524391	170486066	163215416
Q8K3A9_peptide3	Q8K3A9 1xPhospho	44.651	1.00E-17	1.80E-16	5703.89014	NA	NA	NA	52094.7211	379647.555	246306.627	263341.943
O70405_peptide15	O70405 2xPhospho	41.604	1.00E-17	1.80E-16	443341.813	NA	NA	NA	18736060.7	18185062.1	18708450.2	18145959.1
Q69ZS7_peptide1	Q69ZS7 1xPhospho	41.419	1.00E-17	1.80E-16	441748.656	NA	NA	680786.613	19914047.7	21432643.3	25907033	25382133.9
Q3UR85_peptide1	Q3UR85 1xPhospho	38.857	1.00E-17	1.80E-16	NA	NA	551828.909	NA	28871425.6	24805324.8	18535045.7	15695373.7
Q9CW46_peptide5	Q9CW46 1xPhosph	38.717	1.00E-17	1.80E-16	NA	NA	88985.1395	77512.1016	2512392.47	3042619.82	3398112.3	3642486.71
P42128_peptide4	P42128 2xPhospho	37.731	1.00E-17	1.80E-16	845668.688	495542.055	NA	NA	17804434.3	29783485.8	28000820.3	33508440.4
P62754_peptide5	P62754 3xPhospho	36.156	1.00E-17	1.80E-16	10185576	9222357.74	8766475.33	3376205.11	471642161	325093580	252502586	153500204
E9PYH6_peptide2	E9PYH6 2xPhospho	35.877	1.00E-17	1.80E-16	1019608.13	3314505.19	1914659	2002022.6	64315163.8	72497853.2	76069086.9	68056702.1
Q6P5E6_peptide2	Q6P5E6 1xPhospho	34.864	1.00E-17	1.80E-16	363606.469	1143499.06	581036.554	397762.632	12694787.1	25750223.3	20228799.1	17225119.7

Hands on part!

Exercises

DDZ
Deutsches Diabetes-Zentrum

- **Exercise 2**
  - https://drive.google.com/drive/folders/1vmewprs0gkpakU8idbgtexDIwmGVUJz3?usp=sharing
  - Work through part 2 of the given R tutorial (video & slides)
  - Solve tutorial exercises 2.2 – 2.10
  - Please send me your solutions as an ".R"-file

Thank you!