Bioinformatical analysis of omics expression data Part 1

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

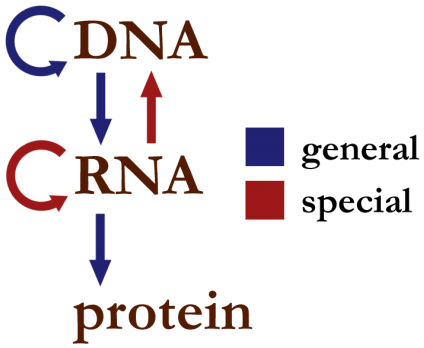
Course schedule

DDZ Deutsches Diabetes-Zentrum

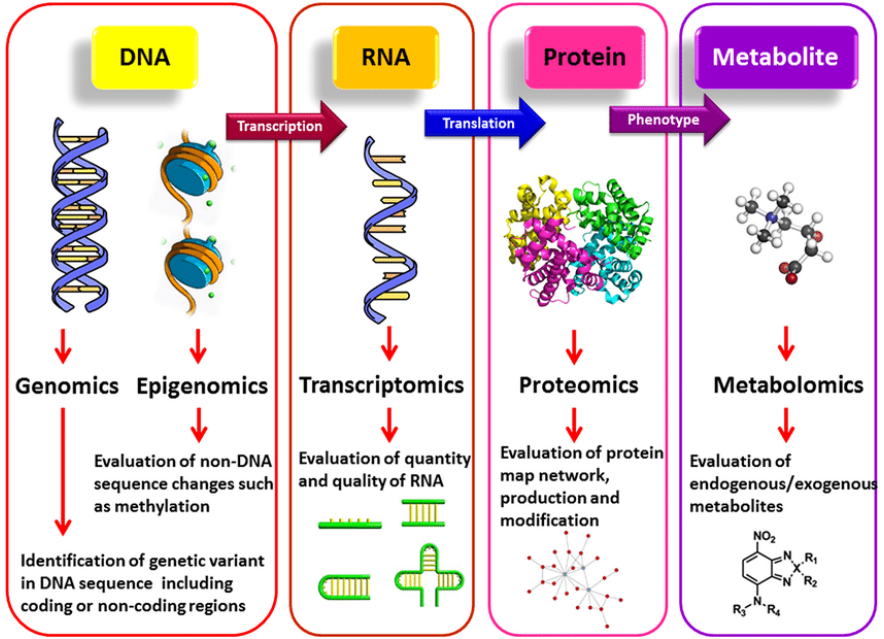
- **Part 1 (25.10.23)**
  - **Introduction (omics, example data, programming)**
  - **Data preprocessing (data inspection, normalization, missing values)**
  - **Exercises: R programming tutorial (part 1)**
- Part 2 (08.11.23)
  - Differential expression analysis (statistics, volcano plot)
  - Exercises: 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

Omics technologies

DDZ
Deutsches Diabetes-Zentrum



Central dogma of molecular biology:
Information flow



Different omics technologies: investigation of information at different levels

Common features of omics data

DDZ Deutsches Diabetes-Zentrum

- High-throughput measurements
  - Ideally, measurement of all genes / transcripts / proteins / etc. in sample
  - → many thousands of molecules per sample
- Multiple samples
  - Ideally, multiple various samples needed to cope with biological variability
  - Needed for robust statistics and algorithms
  - Limitation: sample availability & costs
- Main problems:
  - “n << p” problem: too few samples for too many molecules
  - Noisy data
  - Missing molecules & missing values
  - Data interpretation
  - Large data → computational problems

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;A0A0	273110000	0	0	0	415720000	231880000	170450000	0	77606000	0
7	A0A024RA52;P25	27201000	18075000	20193000	34648000	25855000	57069000	17675000	6078100	0	14778000
8	A0A0G2JL47;A0A0	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	A0A075B6J9	0	5006300	0	0	24838000	0	120130000	0	4247200	0
13	A0A075B6K2	0	0	34257000	0	0	0	0	0	0	0
14	A0A075B6K4;P017	183290000	142910000	339380000	27835000	426830000	41551000	125020000	53773000	17543000	76506000
15	A0A075B6K5;P807	154030000	161710000	223690000	0	236050000	270420000	0	257760000	260630000	0
16	A0A075B6Q5	8969300	0	0	0	0	0	0	0	0	0
17	A0A075B6R2	20478000	61140000	107980000	13775000	11235000	53081000	41233000	0	0	7177100
18	A0A0C4DH68;A0A	0	0	0	0	0	0	0	590390000	0	0
19	A0A0C4DH67;A0A	40041000	60561000	108000000	0	115260000	52592000	17819000	0	61582000	36194000
20	A0A075B730;P581	0	0	0	0	0	0	0	0	0	2661600
21	A0A075B785;A0A2	0	0	0	0	0	0	0	0	0	265000
22	A0A075B7B1;O15(	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;H3B	5767000	0	0	0	4577100	6169700	6582700	4150200	0	5979700
26	A0A075B7E8	0	0	0	0	0	0	0	0	2689800	0
27	A0A087WSV8;P80	1034200000	1366900000	1836300000	800310000	776780000	527980000	2053500000	1147500000	1285400000	980880000
28	E9PIR7;F8W809;A	2272200	0	0	7492200	7253800	4113300	0	3018100	0	0
29	A0A087WT27;J3KI	0	0	0	13819000	0	0	0	0	0	0
30	E9PQS1;A0A087W	0	0	0	0	0	0	0	10814000	0	0

Example data

DDZ Deutsches Diabetes-Zentrum

Third party data from a publication

ARTICLE

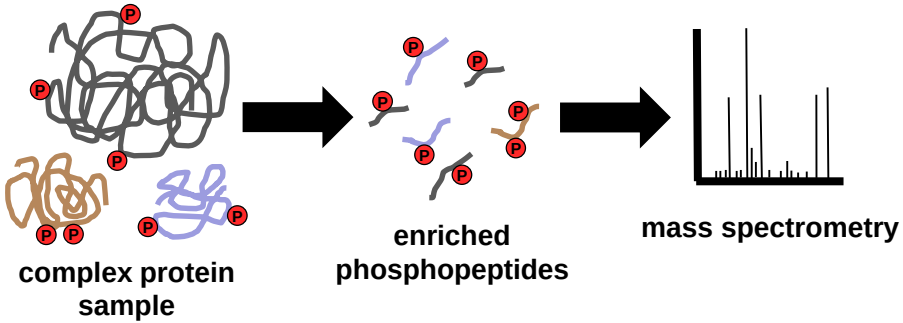
https://doi.org/10.1038/s41467-019-13114-4 OPEN

Global redox proteome and phosphoproteome analysis reveals redox switch in Akt

Zhiduan Su[1,2,13], James G. Burchfield[1,2,13], Pengyi Yang[1,3,13], Sean J. Humphrey[1,2], Guang Yang[1,2], Deanne Francis[1,2], Sabina Yasmin[4], Sung-Young Shin[5,6], Dougall M. Norris[1,2], Alison L. Kearney[1,2], Miro A. Astore[4], Jonathan Scavuzzo[1,2], Kelsey H. Fisher-Wellman[7,8], Qiao-Ping Wang[1,2,9], Benjamin L. Parker[1,2], G. Gregory Neely[1,2,9], Fatemeh Vafaee[1,3], Joyce Chiu[10,11], Reichelle Yeo[10,11], Philip J. Hogg[10,11], Daniel J. Fazakerley[1,2], Lan K. Nguyen[5,6], Serdar Kuyucak[4] & David E. James[1,2,12]*

- Samples: mouse cell line (3T3-L1 adipocytes)
- 4 samples before & 4 samples after insulin stimulation
- Omics technology: **phosphoproteomics**
- Downloaded from raw data repository (PRIDE) & re-analyzed in our lab
- Details: see paper in GitHub repository

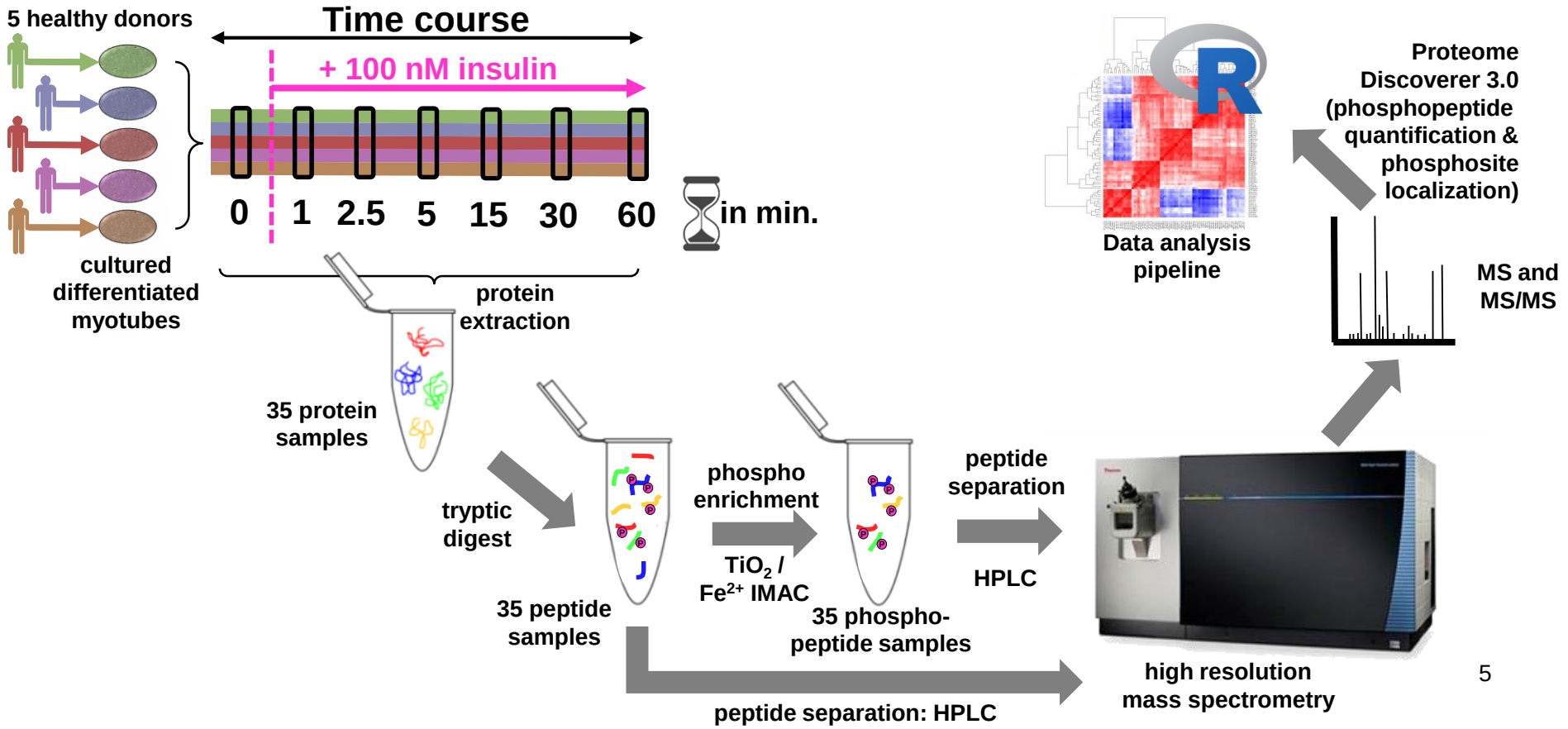
Phosphoproteomics



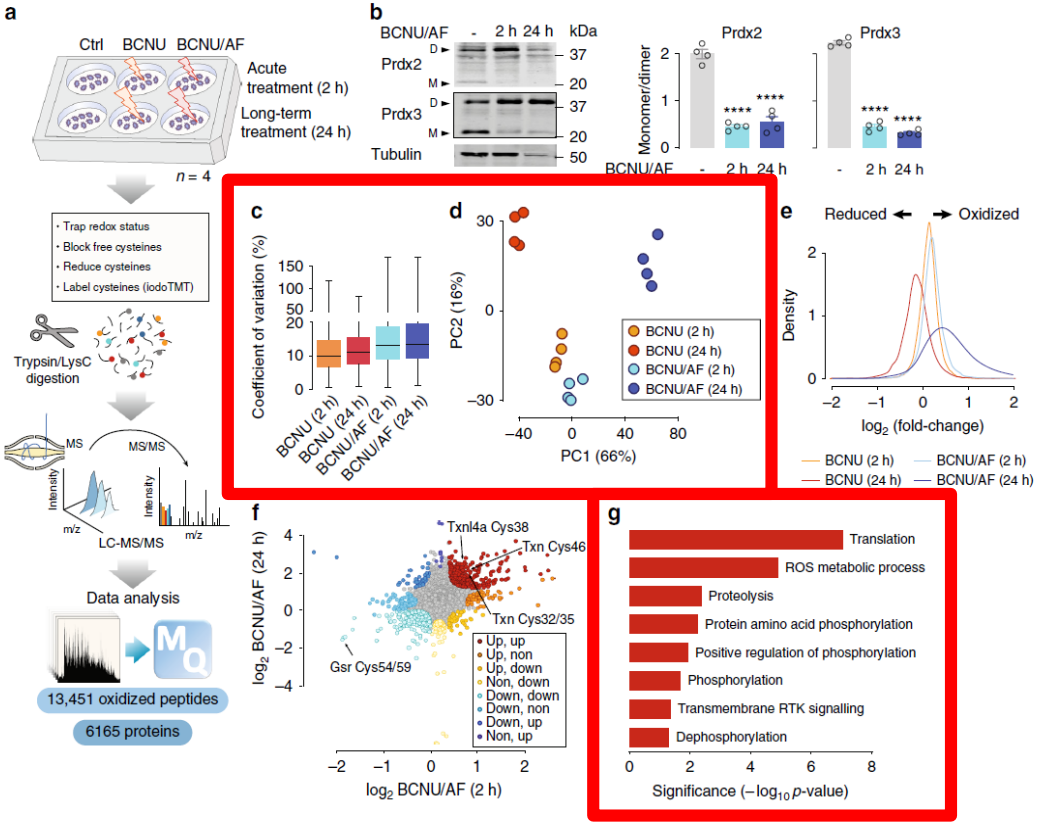
- Completeness: thousands of phosphopeptides (not all) can be identified & quantified in one sample
- Their phosphorylated sites can be localized
- Unbiased view of signaling pathways (at specific time point)

A complex study design

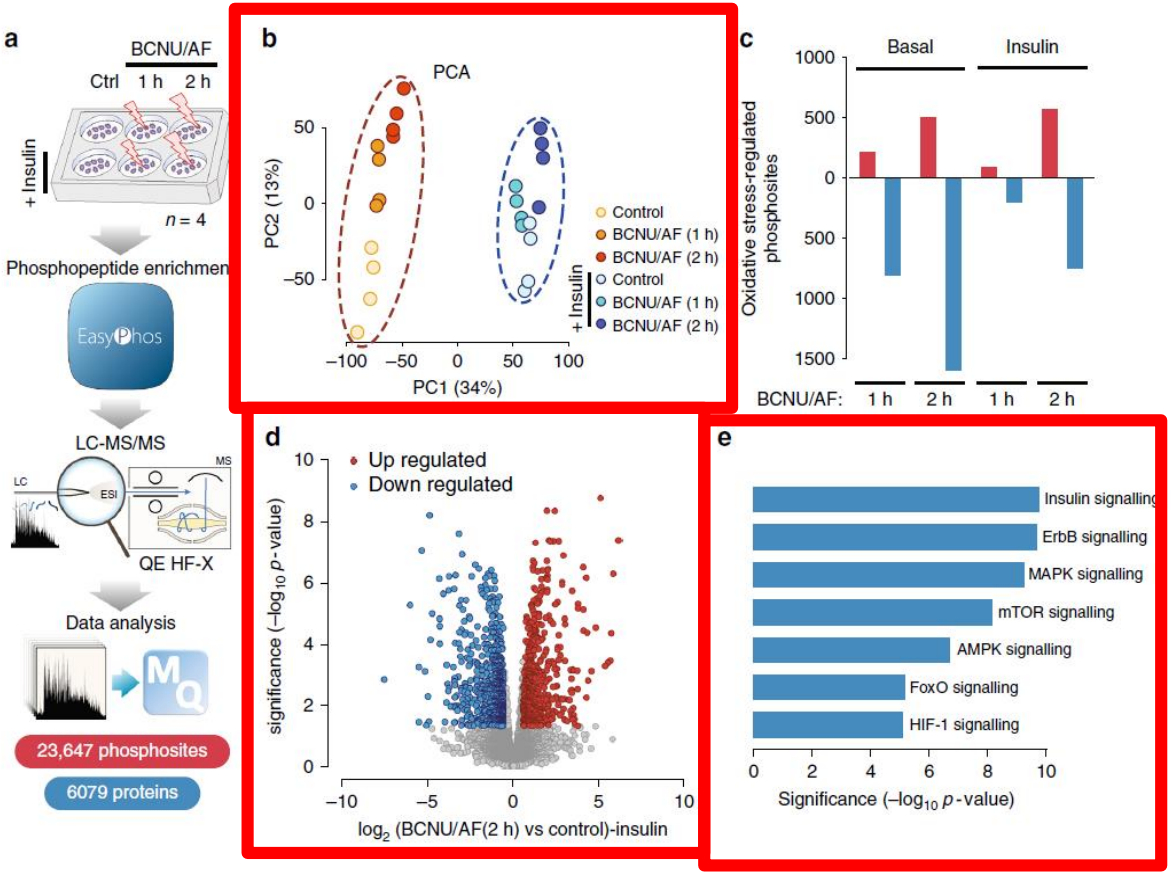
DDZ Deutsches Diabetes-Zentrum



Example data: results



Example data: results



Example data: results

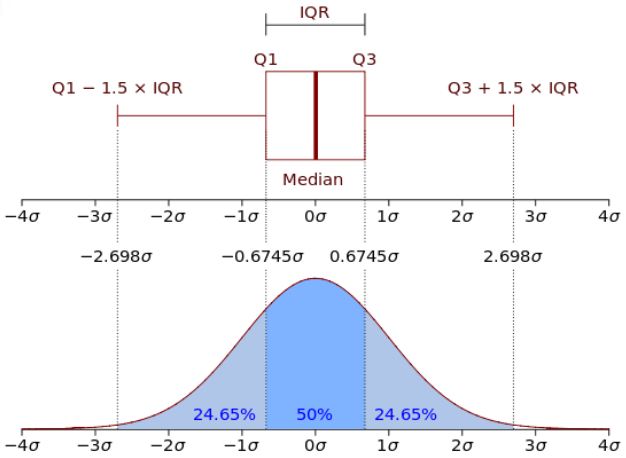
DDZ
Deutsches Diabetes-Zentrum

Data preprocessing: data inspection

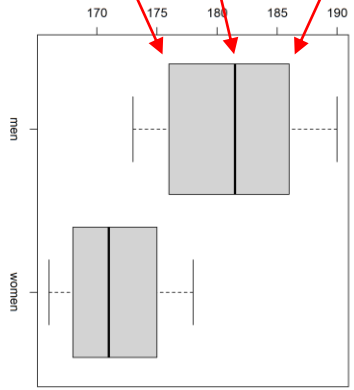
Deutsches Diabetes-Zentrum

	D1	D2	D3	D4	D5	C1	C2
[1,]	56.22302	250.23056	116.351115	60.43167	2.858646	45.68594	16.64358
[2,]	46.04548	898.77619	77.288900	44.52311	74.491848	22.89094	20.66956
[3,]	177.39347	29.87137	211.799695	130.09407	48.632016	25.46085	410.42190
[4,]	74.08954	65.19188	9.852425	167.25783	123.486250	117.47252	34.64786
[5,]	79.11307	95.39491	80.270471	57.08211	73.378632	32.51832	61.40129
[6,]	182.37052	77.76701	71.987875	155.26117	17.509283	34.20217	64.88913

	C3	C4	C5
[1,]	98.67317	111.61690	79.98854
[2,]	142.10068	53.73327	57.16795
[3,]	7821.05165	94.34441	402.20339
[4,]	114.48980	49.67095	95.10256
[5,]	13.12541	340.16019	10.54165
[6,]	117.50140	39.46954	66.53234



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
64.86	74.24	81.37	233.50	340.50	606.52



Data preprocessing: normalization

DDZ
Deutsches Diabetes-Zentrum

Data contains technical and biological variation (but we are only interested in biological differences)

Reasons for technical bias:

- small variations in experimental conditions and sample handling (temperature, age of column, pipetting)
- often exact reasons for bias are unknown

Aims of normalization

- reduce/remove technical bias while keeping biological differences
- make samples more comparable
- make following statistical analysis more reliable

Data preprocessing: normalization

DDZ Deutsches Diabetes-Zentrum

Assumptions:

- high-troughput data
- "true" intensity distribution is similar over all samples
- most proteins are not differentially expressed between groups

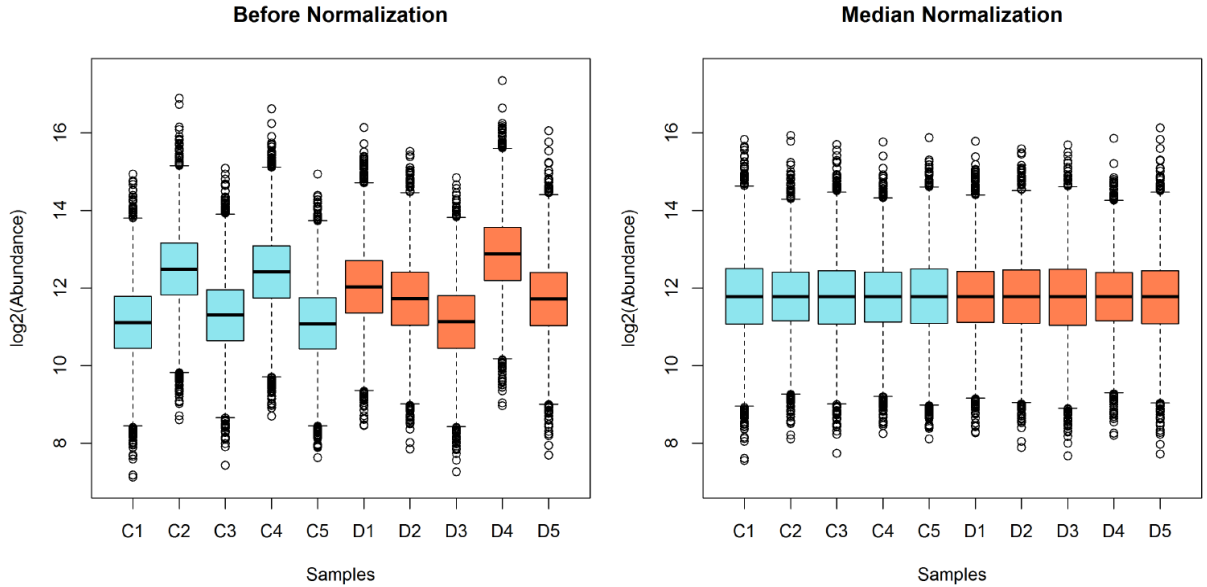
- most normalization methods were developed for genomics and later adapted to proteomics data
- often, data are log-transformed before normalization

Data preprocessing: normalization

DDZ Deutsches Diabetes-Zentrum

Median normalization

shift or scale samples to have the same median



Data preprocessing: normalization

DDZ Deutsches Diabetes-Zentrum

Quantile Normalization

Original dataset

	S1	S2	S3
Prot1	100	50	115
Prot2	85	140	45
Prot3	150	70	80
Prot4	95	65	160

1) Sort Values in each column

S1	S2	S3
85	50	45
95	65	80
100	70	115
150	140	160

2) Replace values with row mean

S1	S2	S3
60	60	60
80	80	80
95	95	95
150	150	150

3) Reconstruct original order

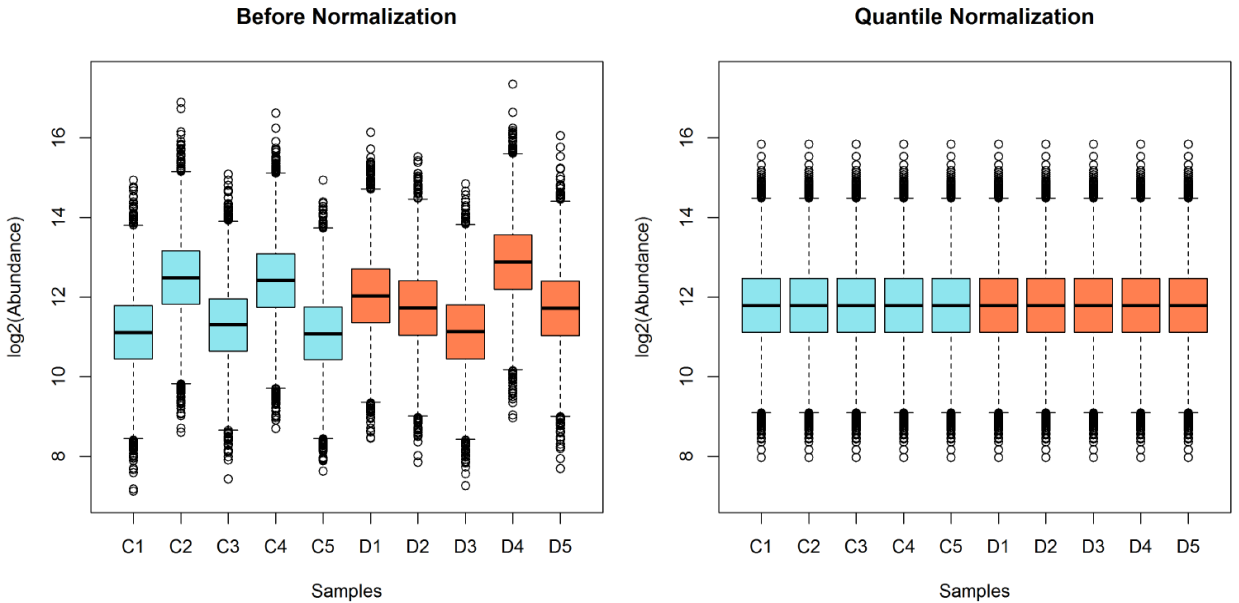
	S1	S2	S3
Prot1	95	60	95
Prot2	60	150	60
Prot3	150	95	80
Prot4	80	80	150

Data preprocessing: normalization

DDZ Deutsches Diabetes-Zentrum

Quantile normalization

normalize all samples to the same distribution



Data preprocessing: missing values

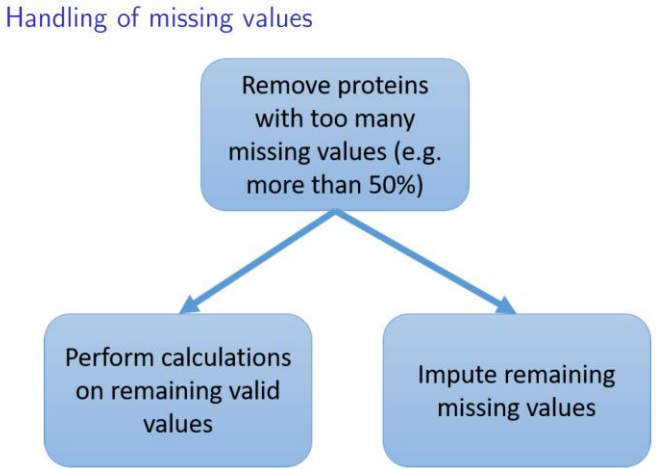
DDZ Deutsches Diabetes-Zentrum

Missing values

- Different codings for missing values depending on software and output settings (NA, NaN, 0, Filtered, ?, empty cell)
- Number of missing values often very high
- Proteins with missing values might be interesting (on/off-proteins)

Handling of missing values

- remove proteins with missing values
- perform analysis only on valid values
- impute missing values



Data preprocessing: missing values

DDZ Deutsches Diabetes-Zentrum

Imputation of missing values

Data Imputation = replace missing values with valid values

Imputation methods

- mean or median of the protein
- random value based on distribution of non-missing values
- small values (e.g. 0 or LOD/2, LOD = limit of detection)
- machine learning based

Disadvantages

- imputation can have a huge impact on result
- imputation of constant value can lead to underestimated variance $\rightarrow$ risk of false positives
- biomarker candidates with too many imputed values may be worthless

Data preprocessing: missing values

DDZ Deutsches Diabetes-Zentrum

On/off-proteins

- proteins that are present in one group and absent in the other group
- $\approx$ proteins that have valid values in one group and missing values in the other group
- higher confidence for found on/off proteins with high sample size
- on/off proteins are often forgotten or filtered out by the software
- t-test not possible $\rightarrow$ not p-value
- fold change $= \infty$?
- cannot be displayed in volcano plot $\rightarrow$ separate list

Hands on part!

Exercises

DDZ Deutsches Diabetes-Zentrum

- **Exercise 0**
  - Install R
  - Install RStudio Desktop
  - Test RStudio
  - Test Google Colab Notebook

- **Exercise 1**
  - https://drive.google.com/drive/folders/1vmewprs0gkpakU8idbgtexDIwmGVUJz3?usp=sharing
  - Work through part 1 of the given R tutorial (video & slides)
  - Solve tutorial exercises 1.1 – 2.1
  - Please send me your solutions as an ".R"-file

Thank you!