Bioinformatical analysis of omics expression data Part 4

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

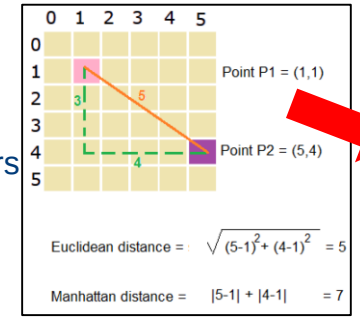
DDZ Deutsches Diabetes-Zentrum

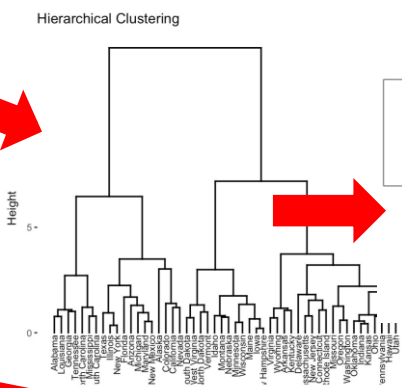
- Part 1 (25.10.23)
  - Introduction (omics, example data, programming)
  - Data preprocessing (data inspection, normalization, missing values)
  - Exercises: R programming tutorial (part 1)
- Part 2 (08.11.23)
  - Differential expression analysis (statistics, volcano plot)
  - Exercises: R programming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

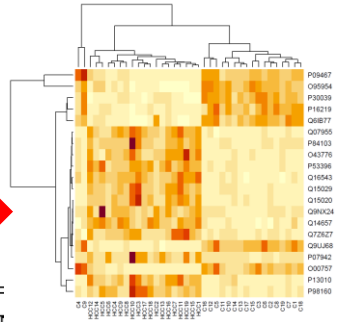
Recap of previous part

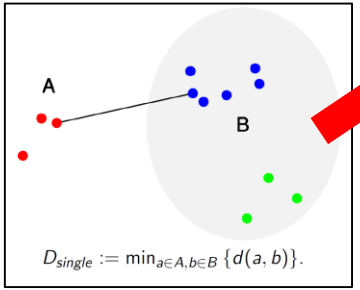
Deutsches Diabetes-Zentrum

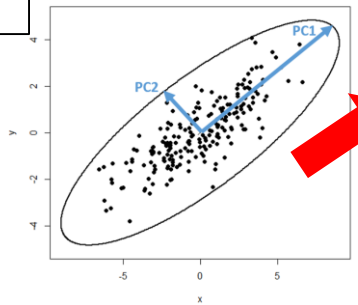
- **Hierarchical clustering**
  - Basic idea:
    1. Initialization: each data point (=vector) own cluster
    2. Iteration: compute distance between current clusters & merge most nearby clusters
    3. End: all data points in one “root” cluster
  - Visualization: dendrogram & heat map
  - Distance functions: Euclidean, Manhattan, correlation-based, (…)
  - Linkage functions: single, complete, average, (…)
- **Principal Component Analysis (PCA)**
  - Developed for data reduction
  - Principal components (PCs): linear combinations of original variables (= genes, proteins, …)
  - PC1: vector with largest share of variance in data
  - PCn: orthogonal vector to previous n-1 PCs & maximum “remaining” part of variance
  - Visualization: scatter plot of PC1 and PC2
  - Usage: data inspection, finding clusters, quality control
- **Programming (example: R)**
  - Own hierarchical clustering & PCA + visualization in R

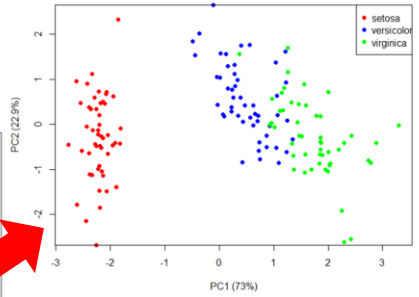








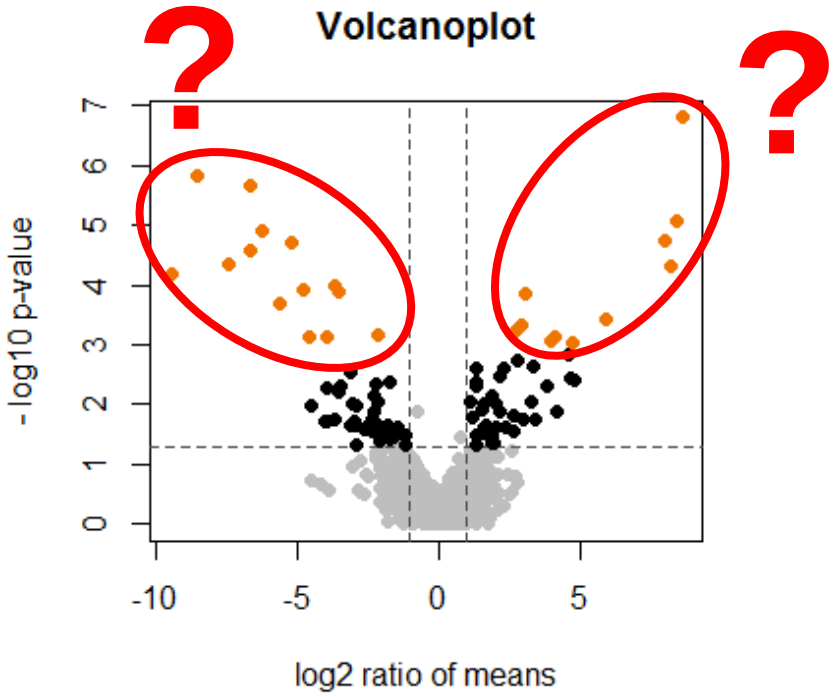




List of candidates: what's next?

DDZ Deutsches Diabetes-Zentrum

- We have learned to find a list of statistically significant differential candidates with p-values and fold changes
- How can we interpret these biologically?
- Are there biological connections that could explain a common occurrence?
- Can a common function be assumed?



Overrepresentation analysis (ORA)

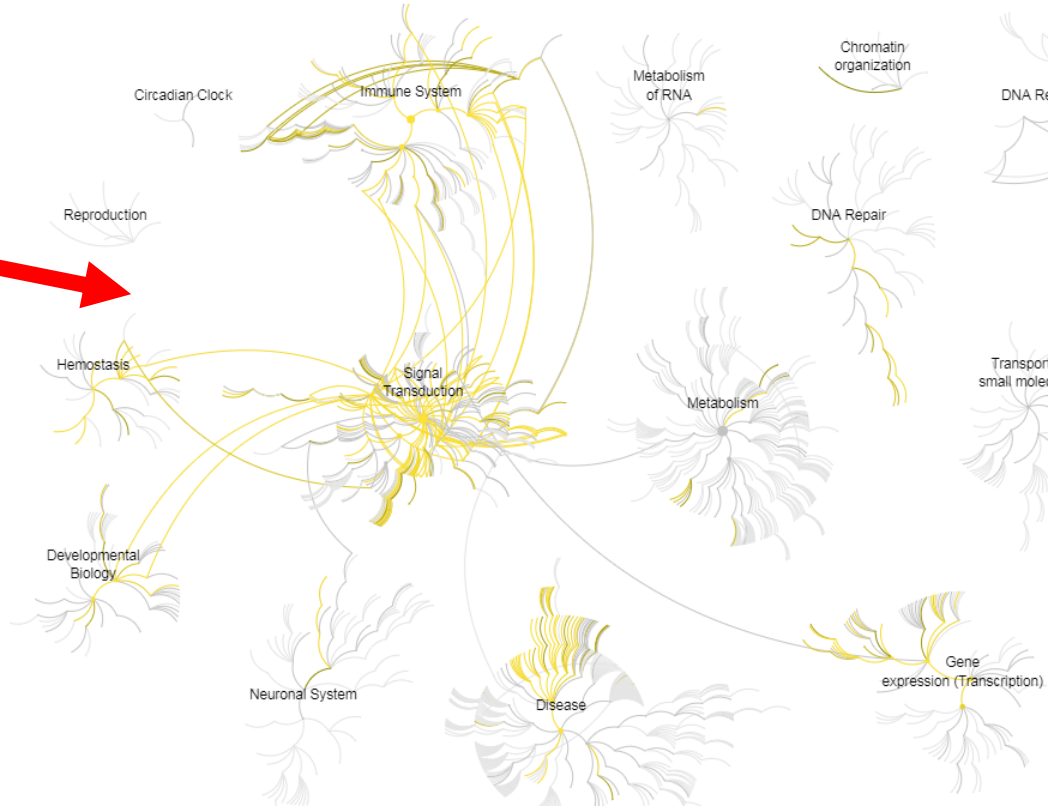
DDZ Deutsches Diabetes-Zentrum

- **Basic principle:** Annotate input genes/proteins with biological terms (existing knowledge from databases) to assess whether e.g. specific pathways, kinase substrates or cellular compartments are significantly overrepresented among them. → **Functional analysis of input genes/proteins.**
- **E.g., input genes/proteins may be:**
  - Differentially expressed genes/proteins.
  - Co-expressed genes/proteins.
  - A gene/protein cluster from hierarchical clustering.
  - Set of interesting genes/proteins from literature / a database.
- **Aim:** Statistical score (p-value) to decide objectively whether an annotated term is overrepresented among the input gene/proteins more than would be expected by chance.

DBs containing relevant knowledge

DDZ
Deutsches Diabetes-Zentrum

- **Gene Ontology (GO):** Controlled vocabulary for genes & proteins
- **Reactome:** Biochemical pathway DB & annotation tool
- **KEGG:** Biochemical pathways
- **PhosphoSitePlus:** Kinases and their substrates
- And many more…



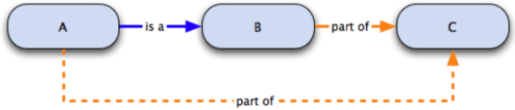
Gene Ontology (GO): a hierarchy of terms

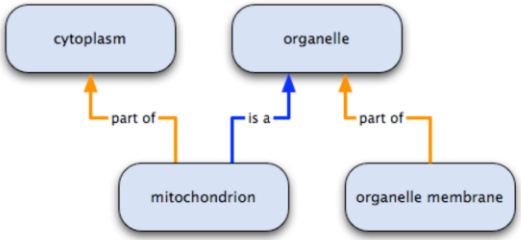
DDZ Deutsches Diabetes-Zentrum

- GO: organism-specific hierarchy of controlled biological terms (→ "controlled vocabulary"), which are well defined, labelled with unique ID, always related to at least one parent/child term
- Visualization: directed acyclic graph (DAG) of terms (= nodes). Edges: "is a"- & "part of"-relationships. → general terms close to root of DAG & terminal nodes most specific
- Organized in 3 seperate GO "domains" (= separate DAGs).
  - Biological process: 29,687 terms
  - Molecular function: 11,110 terms
  - Cellular component: 4,206 terms

  (numbers as of January 2019)

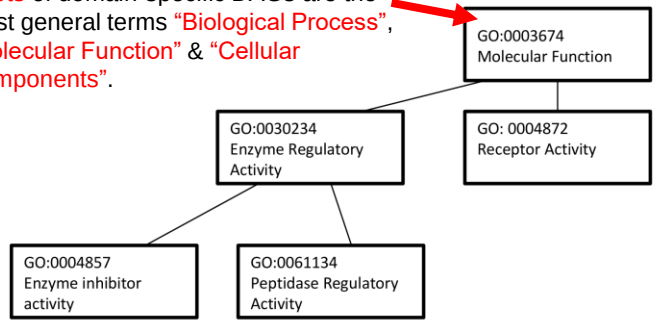
- Based on "is a" or "part of" relationship

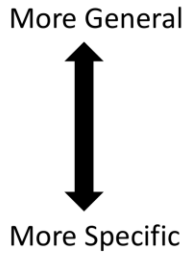




- Hierarchical relationship

Roots of domain-specific DAGs are the most general terms "Biological Process", "Molecular Function" & "Cellular Components".

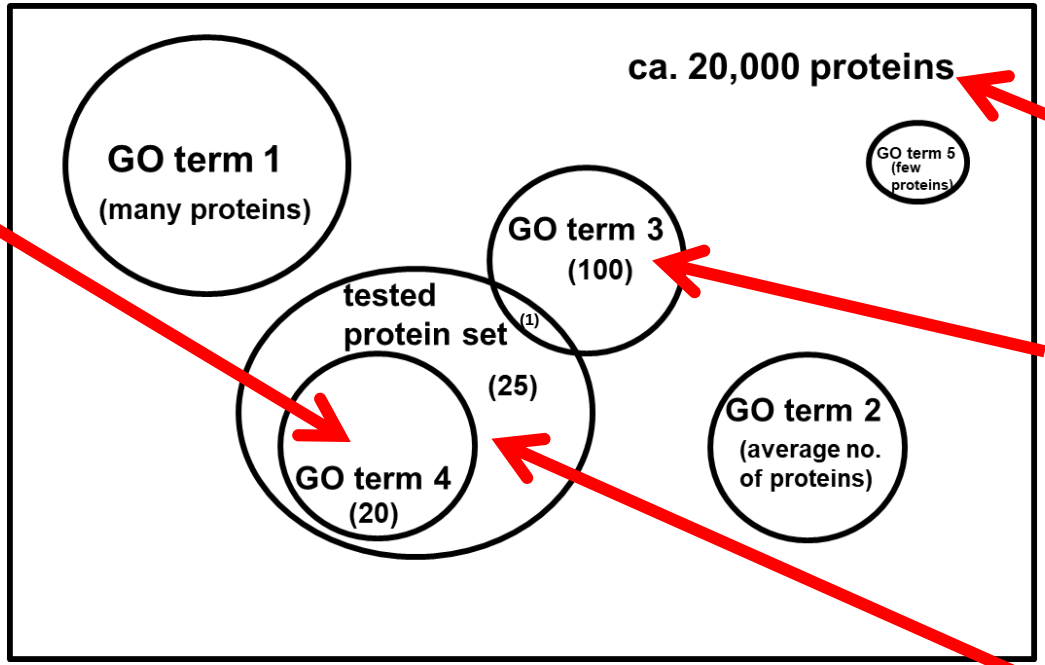




Basic concept of GO-based ORA

DDZ Deutsches Diabetes-Zentrum

GO term 4 is overrepresented in tested protein set (20 proteins)?



Background proteome (should be as exact as possible, e.g. proteome of analyzed tissue)

GO term 3 is NOT overrepresented (1 protein)?

The tested protein set (25 proteins)

- Here, 80% of the tested proteins & 0.1% of proteome are annotated with **"GO term 4"**. → **Overrepresented?!**
- 4% of tested proteins & 0.5% of proteome are annotated with **"GO term 3"**. → **Not overrepresented?!**
- We need a statistical method to make this decision more objective. → We need a statistical test!

→ 20/25*100 = 80% of tested proteins annotated with "GO term 4" & 20/20,000*100 = 0.1% of complete proteome

→ 1/25*100 = 4% of tested proteins annotated with "GO term 3" & 100/20,000*100 = 0.5% of complete proteome

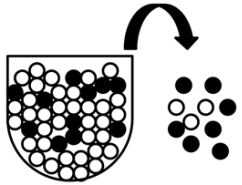
ORA: statistical tests

DDZ Deutsches Diabetes-Zentrum

- Specific statistical tests return **p-values** that are used as scores for term overrepresentation in a given protein set.
- Tests used for ORA are:
  - Fisher's exact test / hypergeometric test
  - Kolmogorov Smirnov test
  - And others…

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

➢ Fisher's Exact Test (Hypergeometric Test)



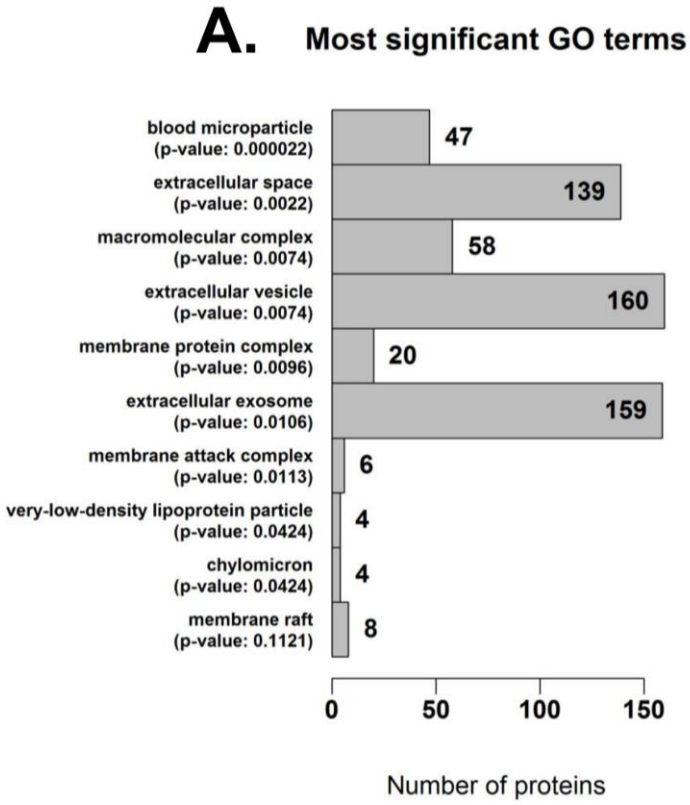
The test implements the urn model.

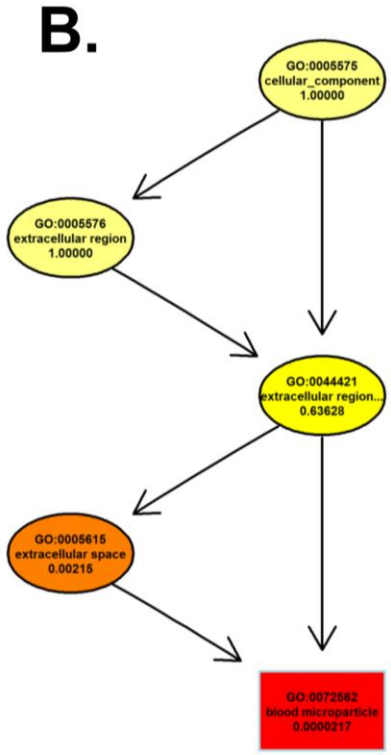
What is the probability of getting 7 or more black balls?

- **N:** total number of proteins
- **M:** total number of proteins annotated with this term
- **n:** number of proteins in the set (all balls drawn)
- **k:** number of proteins in tested set annotated with this term (black balls drawn)

GO-based ORA

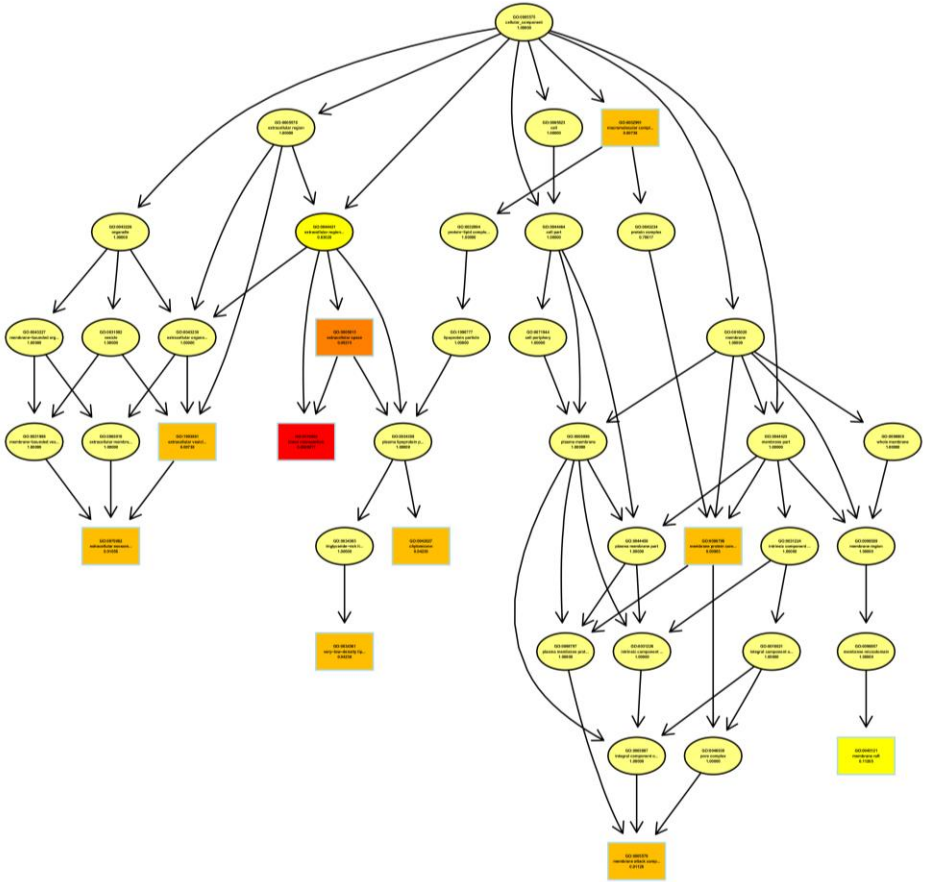
Deutsches Diabetes-Zentrum





GO-based ORA

DDZ
Deutsches Diabetes-Zentrum



ORA: tools (examples)

DDZ Deutsches Diabetes-Zentrum

- ORA-Tools in R
  - topGO
  - ReactomePA
  - (…)
- Online GO-based ORA
  - Panther: https://pantherdb.org/
  - (…)
- Online Reactome-based ORA
  - Reactome: https://reactome.org/
  - (…)
- Tools performing ORA with multiple databases
  - g:Profiler: https://biit.cs.ut.ee/gprofiler/gost
  - CPDB
  - STRING
  - (…)

Hands on part!

Exercises

DDZ Deutsches Diabetes-Zentrum

- **Exercise 4**
  - https://drive.google.com/drive/folders/1vmewprs0gkpakU8idbgtexDIwmGVUJz3?usp=sharing
  - Use our example dataset from GitHub for the following exercises
  - **Exercise 4.1:** Perform an own GO-based ORA in R using differential candidates and all three GO-domains (BP, CC and MF). Visualize the resulting top 10 terms as bar plots.
  - **Exercise 4.2:** Perform an own Reactome-based ORA in R using differential candidates. Visualize the resulting top 10 terms as bar plots.
  - **Exercise 4.3 (optional):** Can you reproduce your R-based results with respective GO (https://biit.cs.ut.ee/gprofiler/gost) & Reactome (https://reactome.org/) online tools?
  - Please send me your solutions as an ".R"-file

Thank you!