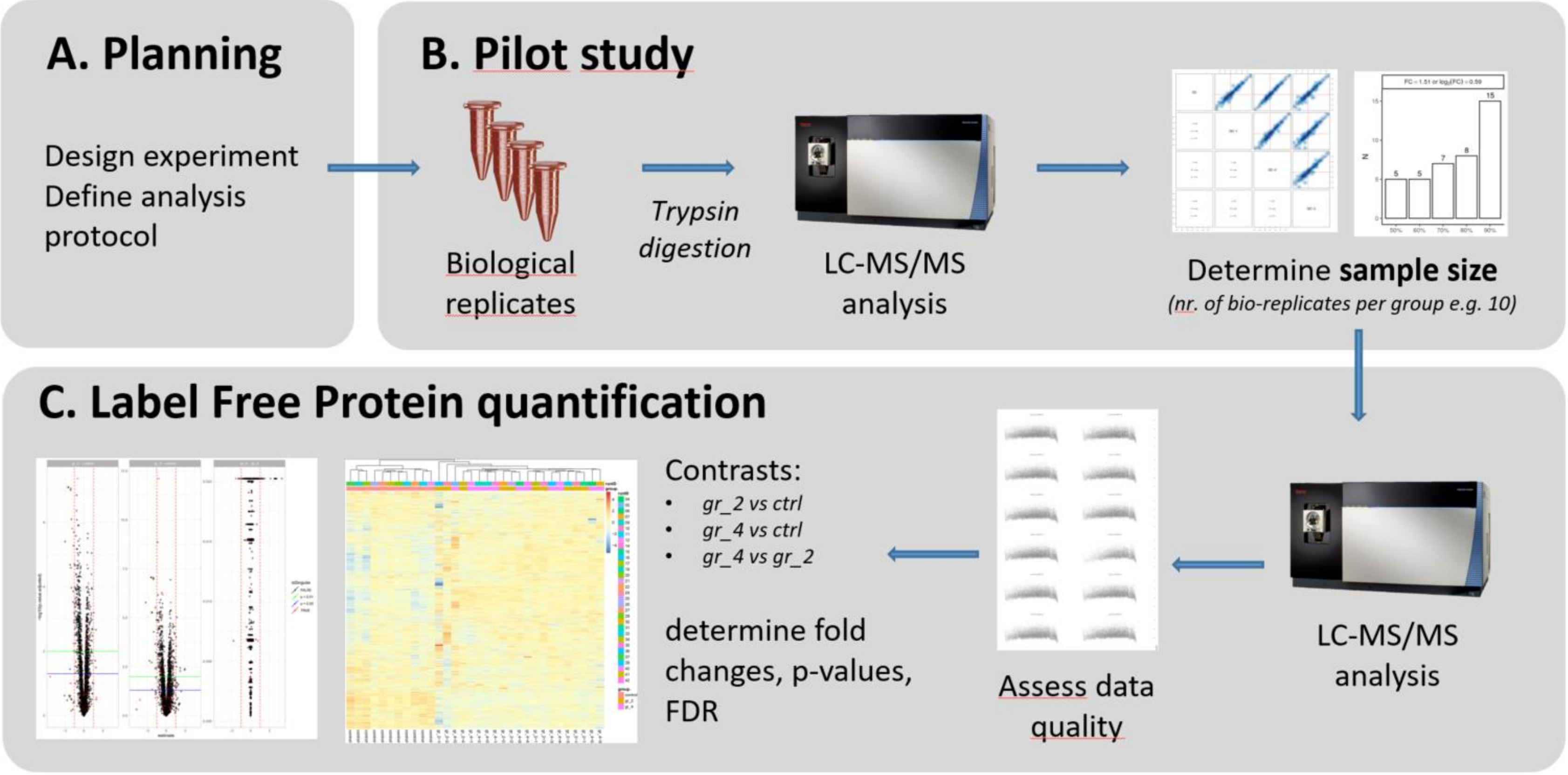
The **FGCZ-GSEA-ORA** Tool for Pathway Analysis in Proteomics

Nanni P.[1], Wolski W.[1], Kook L. , Panse C.[1], Grossmann J.[1], Fortes C.[1], Kunz L.[1], Schlapbach R.[1]

[1]Functional Genomics Center Zurich (FGCZ), University/ETH Zurich

Quantify thousands of proteins

Mass spectrometry (LC-MS) can quantify thousands of proteins in a biological sample, revealing up- or downregulation of the proteome in response to a stimulus. The result is a list of proteins, ranked by fold-change, t-statistic or p-value, which can be used for pathway analysis.



A. Planning

Design experiment
Define analysis protocol

B. Pilot study

Biological replicates → *Trypsin digestion* → LC-MS/MS analysis → Determine **sample size** *(nr. of bio-replicates per group e.g. 10)*

C. Label Free Protein quantification

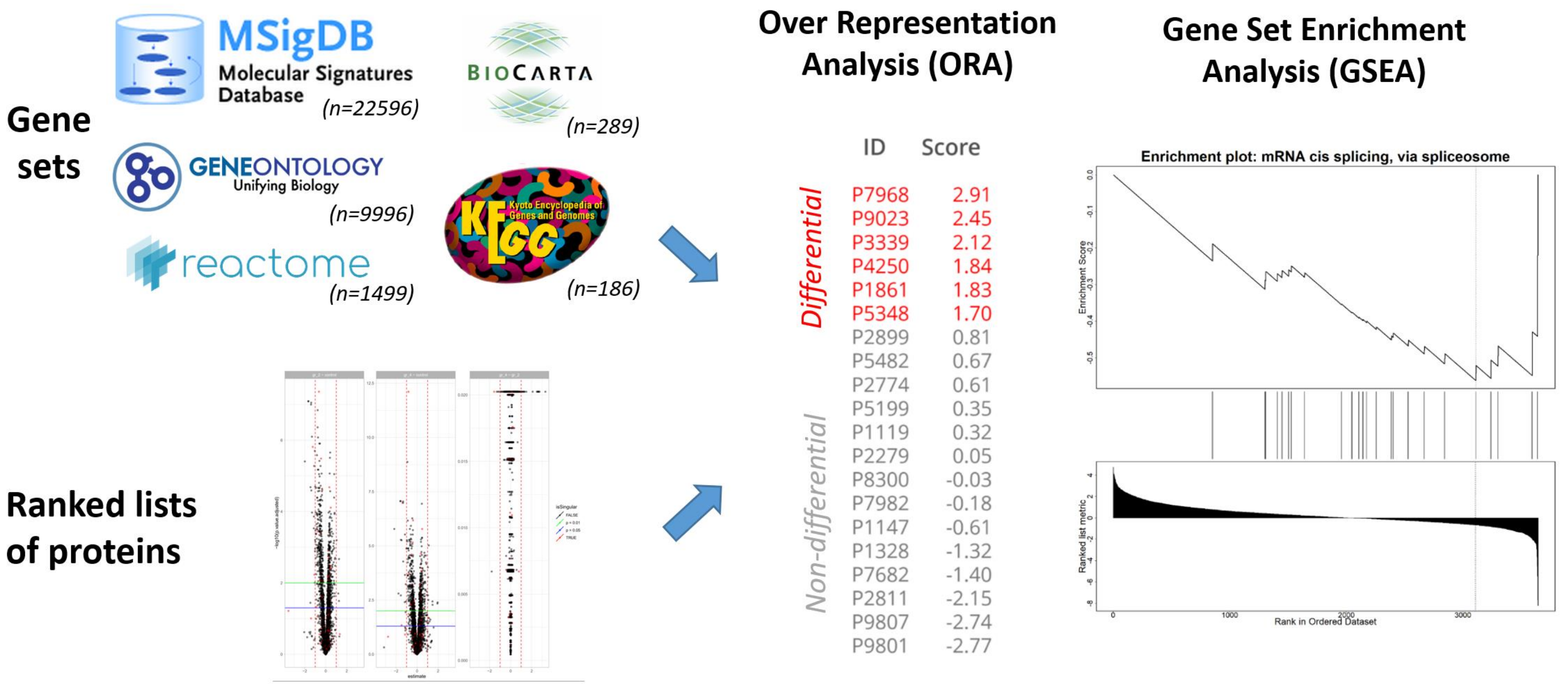
LC-MS/MS analysis → Assess data quality → determine fold changes, p-values, FDR

Contrasts:
- *gr_2 vs ctrl*
- *gr_4 vs ctrl*
- *gr_4 vs gr_2*

Figure 1. *A standard label-free protein quantification workflow*

Analyze which pathways are affected

Analyzing the pathways in your experiment adds biological context and is often more sensitive and specific than monitoring single proteins. Pathway analysis [1] starts with the ranked proteins from your quantification experiment and with pre-defined "gene sets". These sets contain genes which are part of the same pathway or are located close to each other on a chromosome. The analysis subsequently identifies gene sets which are differentially regulated.



ID	Score
P7968	2.91
P9023	2.45
P3339	2.12
P4250	1.84
P1861	1.83
P5348	1.70
P2899	0.81
P5482	0.67
P2774	0.61
P5199	0.35
P1119	0.32
P2279	0.05
P8300	-0.03
P7982	-0.18
P1147	-0.61
P1328	-1.32
P7682	-1.40
P2811	-2.15
P9807	-2.74
P9801	-2.77

Figure 2. *Pathway analysis requires gene sets and a ranked list of protein identifiers, like the one you get from a quantitative proteomics experiment [2].*

Two pathway analysis methods:

Overrepresentation Analysis (ORA)

- Pick all proteins exceeding a threshold (e.g. log2|FC|>1)
- Test if a gene set is over-represented among them (e.g. Fisher's Exact Test).
- How to choose the threshold? E.g. biologically relevant effect size

Gene Set Enrichment Analysis (GSEA)

- Proteins ranked by score (log2(FC), t-statistic)
- For each gene set: locate its members among the proteins and compute enrichment scores
- No thresholds needed

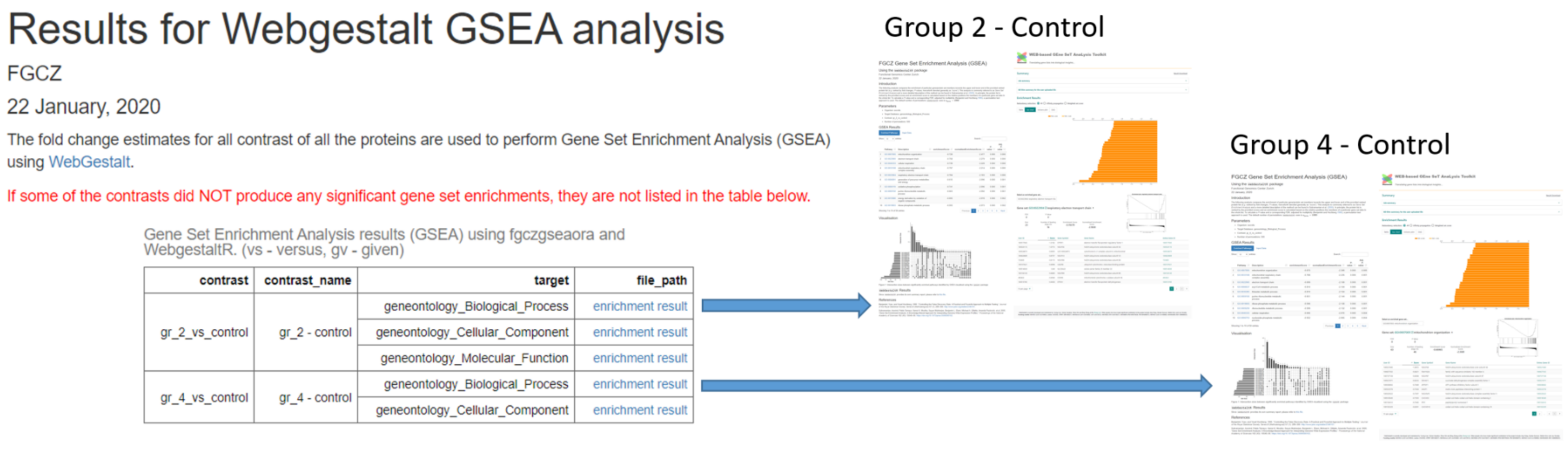
The FGCZ-GSEA-ORA Tool

A plethora of R packages for ORA or GSEA are available on the Comprehensive R Archive Network [3] and Bioconductor [4]. We developed the **FGCZ-GSEA-ORA** package [5] to make pathway analysis even easier.

FGCZ-GSEA-ORA is an R package for pathway analysis which:

- **unifies ID mapping and enrichment analysis**
- **ensures reproducible results**
- **is user-friendly**

Our R package generates HTML reports with visualizations and explains pathway analysis results. As of now, we integrated the pathway analysis using WebgestaltR [6] and sigora (Signature Overrepresentation Analysis [7]).



Results for Webgestalt GSEA analysis

FGCZ

22 January, 2020

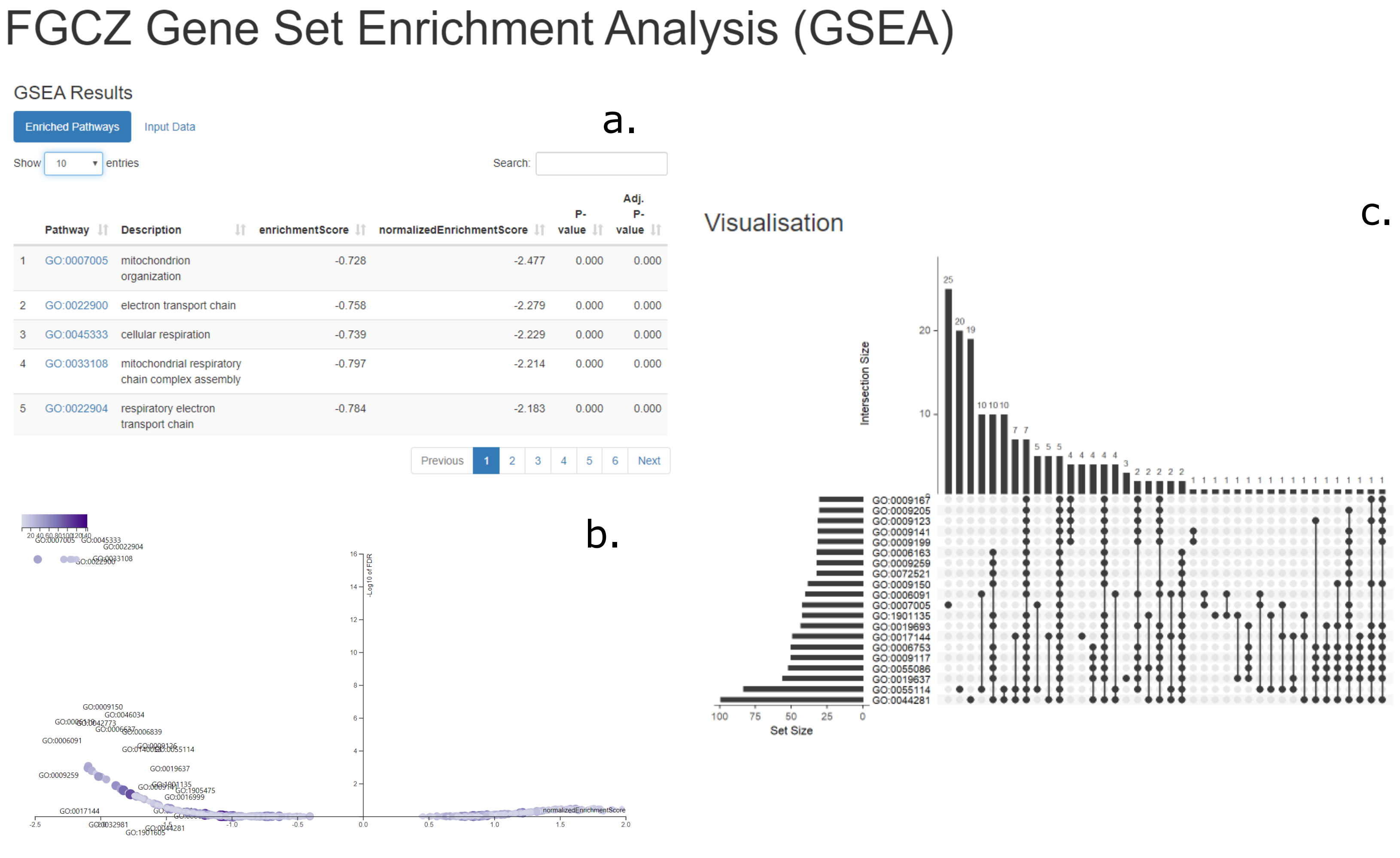
The fold change estimates for all contrast of all the proteins are used to perform Gene Set Enrichment Analysis (GSEA) using WebGestalt.

If some of the contrasts did NOT produce any significant gene set enrichments, they are not listed in the table below.

Gene Set Enrichment Analysis results (GSEA) using fgczgseaora and WebgestaltR. (vs - versus, gv - given)

contrast	contrast_name	target	file_path
gr_2_vs_control	gr_2 - control	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result
		geneontology_Molecular_Function	enrichment result
gr_4_vs_control	gr_4 - control	geneontology_Biological_Process	enrichment result
		geneontology_Cellular_Component	enrichment result

Figure 3. *FGCZ-GSEA-ORA creates a folder structure with HTML files visualizing the ORA and GSEA results for all contrasts (e.g. gr_2 vs ctrl, gr_4 vs ctrl) and all selected signature databases (e.g. GO biological process, …).*



FGCZ Gene Set Enrichment Analysis (GSEA)

GSEA Results

	Pathway	Description	enrichmentScore	normalizedEnrichmentScore	P-value	Adj. P-value
1	GO:0007005	mitochondrion organization	-0.728	-2.477	0.000	0.000
2	GO:0022900	electron transport chain	-0.758	-2.279	0.000	0.000
3	GO:0045333	cellular respiration	-0.739	-2.229	0.000	0.000
4	GO:0033108	mitochondrial respiratory chain complex assembly	-0.797	-2.214	0.000	0.000
5	GO:0022904	respiratory electron transport chain	-0.784	-2.183	0.000	0.000

Figure 4. *Example of GSEA plots included in the report: list of enriched pathways (a); volcano plots (b); intersection sizes between significantly enriched pathways (c).*

FGCZ-GSEA-ORA creates automatic reports and is provided as part the label-free proteomics quantification analysis performed at the FGCZ.

Upcoming Features

- Creating reports with enrichment results for different treatments
- Standardized R-API interface to compare different ORA GSEA algorithms
- Generating algorithm-independent reports.
- Additional analysis tools (e.g. enrichR, topGO).

Want to quantify your proteins? Contact us at proteomics@fgcz.ethz.ch!

Label-free quantification is one of our most popular services and we're happy to discuss your project.

● ● ● **www.fgcz.ch/applications/proteomics**

[1] https://en.wikipedia.org/wiki/Pathway_analysis
[2] MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp)
[3] CRAN (https://cran.r-project.org/)
[4] Bioconductor (http://bioconductor.org/)
[5] fgcz.gsea.ora (https://github.com/protViz/fgczgseaora)
[6] WebgestaltR (http://www.webgestalt.org/)
[7] sigora (https://CRAN.R-roject.org/package=sigora)