



prolfquapp - Streamlining Protein Differential Expression Analysis in Core Facilities

Witold Wolski^{1; 2}, Bernd Roshitzki¹, Jonas Grossmann^{1; 2}, Claudia Fortes¹, Paolo Nanni¹, Christian Panse^{1; 2}, Ralph Schlapbach¹

¹ Functional Genomics Center Zurich - ETH Zurich/University of Zurich (<https://www.fgcz.ch/>); ² Swiss Institute of Bioinformatics (<https://www.sib.swiss/>)

- Protein differential expression analysis (DEA)
 - DIANN
 - FragPipe DDA
 - FragPipe TMT
 - MaxQuant
- Uses preprocessing and statistical models implemented in the R package `prolfqua`
- Generates dynamic HTML reports
- Exports results as XLSX files, `.rnk` and `.txt` files for GSEA and ORA

How To

Install R and prolfquapp

```
install.packages('remotes')
remotes::install_github('wolski/prolfquapp', dependencies = TRUE)
```

Create a directory with :

- `config.yaml` (parameter file)
- `dataset.csv` (experimental design)
- the FASTA file
- DIANN, FragPipe or MaxQuant results

Copy the R code into the working directory by running one of the functions:

```
copy_DEA_DIANN
copy_DEA_FragPipe_DDA
copy_DEA_FragPipe_TMT
copy_DEA_MaxQuant
```

The content of the working directory is:

```
..
_DiffExpQC.Rmd
_Grp2Analysis.Rmd
bibliography.bib
C3000WU289521
config.yaml
dataset.csv
diann-output.tsv
fgcz_tripleProteome_MSV000090837_20221214.fasta
FP_DIA.R
```

Finally, from R console `source("FP_DIA.R")`, or execute `Rscript FP_DIA.R`. This creates a subfolder with the DEA results.

```
..
DE_Groups_vs_Controls.html
DE_Groups_vs_Controls.xlsx
GSEA_B_vs_A.rnk
Ora_B_vs_A.txt
ORA_background.txt
QC_Groups_vs_Controls.html
```

- `DE_Groups_vs_Controls.html` report describing the main steps of the analysis and shows the results.
- `DE_Groups_vs_Controls.xlsx` contains the raw and transformed abundances, annotations, results of the differential expression analysis.
- `.rnk`, and `.txt` files for GSEA and ORA analysis

The entire working directory is archived. It contains all the data and R code and data to replicate the analysis.

Analysis parameters

The `config.yaml` file specifies the parameters of the analysis:

- project related information e.g. `projectID`, is shown in the HTML report
- aggregation method (`medpolish`, `rlm`, `top_3`)
- abundance transformation (`robscale`, `vsn`, `none`),
- FDR and effect size thresholds

```
Bfabric:
  projectID: 3000
  projectName: ''
  orderID: 3000
  workunitID: 289521
  inputID: 2286617
  inputURL: https://fgcz-bfabric.uzh.ch
pop:
  transform: vsn
  aggregate: medpolish
  Diffthreshold: 1.0
  FDRthreshold: 0.1
  removeCon: no
  removeDecoys: no
  revpattern: ^REV
  contpattern: ^CON1^zz
  Software: FragPipeTMT
  zipdir: C3000WU289521
```

Sample annotation

The `dataset.csv` file contains the information about the measured samples:

- `Relative.Path/raw.file/channel (unique)`
- `name` - used in plots and figures (unique)
- `group` - main factor
- `subject/biorePLICATE (optional)` - blocking factor
- `control` - used to specify the control condition (C) (optional)

dataset	RelativePath	Name	GroupingVar	CONTROL	Subject
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_B_03.raw	TripleProteome_B_DIABenchmark_1	B	T	id1
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_B_02.raw	TripleProteome_B_DIABenchmark_2	B	T	id2
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_B_01.raw	TripleProteome_B_DIABenchmark_3	B	T	id3
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_A_04.raw	TripleProteome_A_DIABenchmark_1	A	C	id1
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_B_04.raw	TripleProteome_B_DIABenchmark_4	B	T	id4
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_A_02.raw	TripleProteome_A_DIABenchmark_2	A	C	id2
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_A_01.raw	TripleProteome_A_DIABenchmark_3	A	C	id3
/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30x150minGradients_01.raw	Grad000_LPF_A_05.raw	TripleProteome_A_DIABenchmark_4	A	C	id4

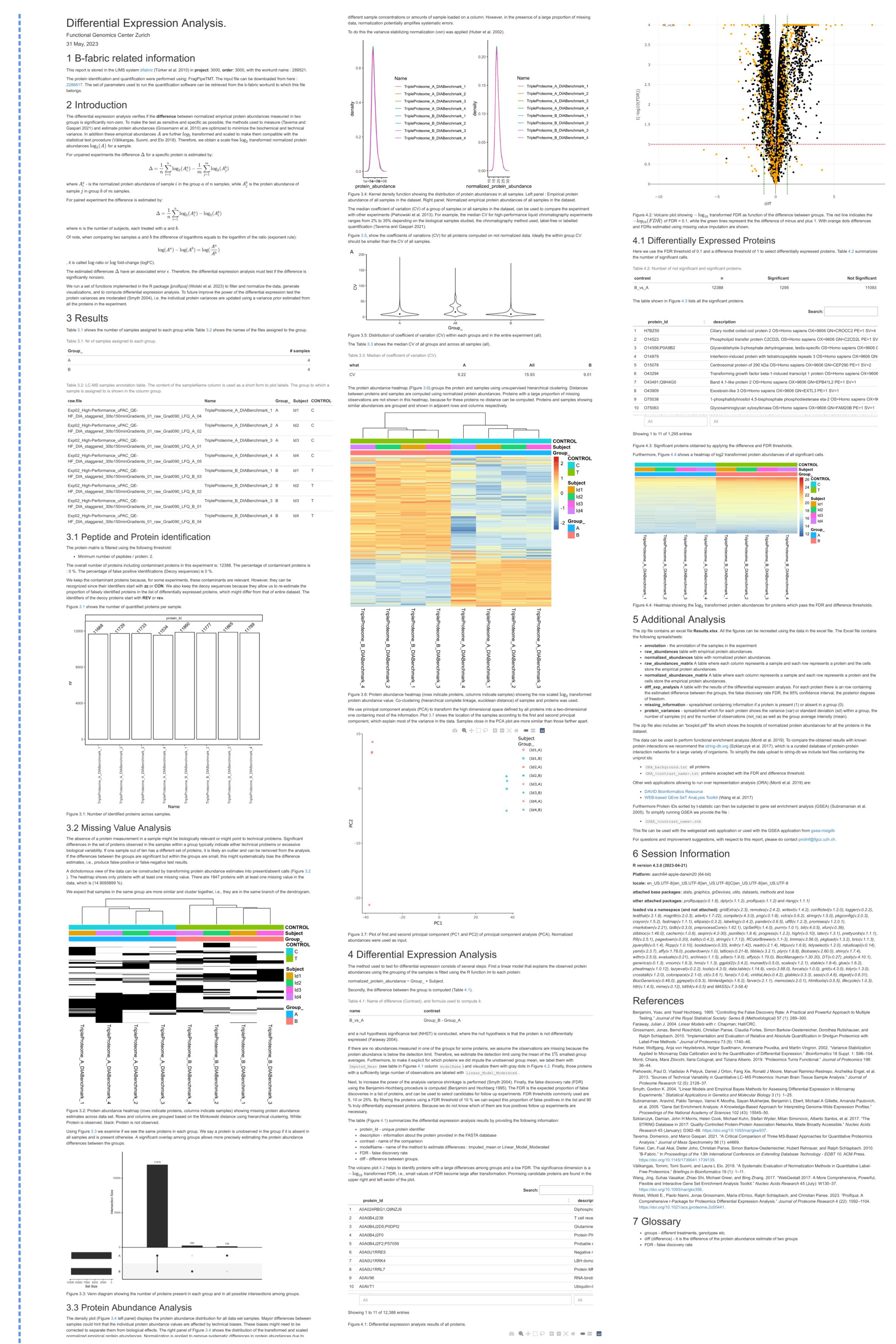
If `subject` is specified then the model is `abundance ~ group + subject`, otherwise `abundance ~ group`. The group differences to compute are determined from the `group` column and the `control` column.

HTML report

- Project related information (project ID etc)
- Primary introduction to DEA
- Sums up the design of the experiment
- Summarizes of protein ident. and quant.: missigness, CV, clustering, PCA
- DEA results with volcano plots and tables (they interact using `crosslink`)
- Explains outputs, give pointers to GSEA and ORA
- Additional QC report

Conclusion

- Integrates into LIMS system
- Archived directory contains all information needed to replicate analysis you can rerun the analysis at your PC
- our users know Excel and like XLSX files
- Shiny app in development



Download

<https://github.com/wolski/prolfquapp>
<https://doi.org/10.1021/acs.jproteome.2c00441>



SIB
Swiss Institute of Bioinformatics