

## Introduction

- Protein differential expression analysis (DEA) for DIANN, FragPipe DDA, FragPipe TMT, MaxQuant outputs, or MSstats inputs.
- Uses preprocessing and statistical models implemented in the R package [prolfqua doi.org/10.1021/acs.jproteome.2c00441](https://doi.org/10.1021/acs.jproteome.2c00441)
- Generates dynamic HTML reports
- Exports results as XLSX files, .rnk and .txt files for GSEA and ORA
- Archived analysis can easily be replicate on any system running R (>= 4.1)

## 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
- Diagnostic plots for each proteins (boxplots, lineplots for peptide abundances)

The entire working directory including input data, R code and results is archived. You can unzip it later and replicate the analysis using your R installation.

### 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: ^CON!^zz
Software: FragPipeTMT
zipdir: C3000WU289521
```

## Sample annotation

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

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

The column names are not case sensitive.

dataset					
	Relative.Path	Name	GroupingVar	CONTROL	Subject
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_03.rnw	TripleProteome_B_DIABenchmark_1	B	T	l01	
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_02.rnw	TripleProteome_B_DIABenchmark_2	B	T	l02	
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_04.rnw	TripleProteome_B_DIABenchmark_3	B	T	l03	
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_05.rnw	TripleProteome_A_DIABenchmark_1	A	C	l01	
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_06.rnw	TripleProteome_A_DIABenchmark_2	A	C	l02	
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_07.rnw	TripleProteome_A_DIABenchmark_3	A	C	l03	
I/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to10minGradients_01_raw_Grad900_LQF_B_08.rnw	TripleProteome_A_DIABenchmark_4	A	C	l04	

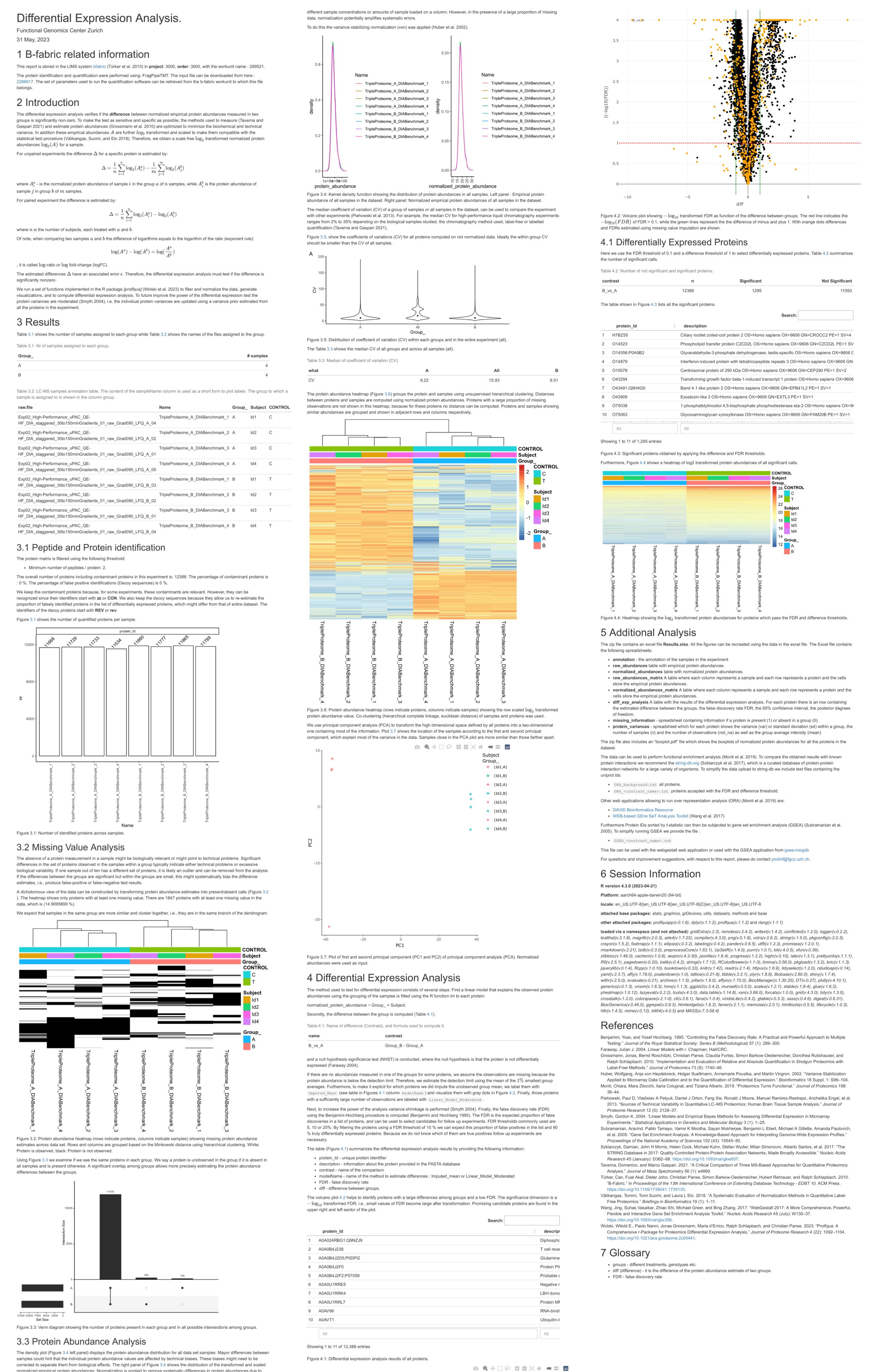
If subject is specified then the model is `abundance ~ group + subject`, otherwise `abundance ~ group`. The group differences to compute are determined from the group and control columns. MSstats annotation.csv and dataset.csv are similar.

## 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 output formats, gives pointers to follow up analysis (GSEA, ORA)

## Summary

- Integrates into LIMS system [doi.org/10.1515/jib-2022-0031](https://doi.org/10.1515/jib-2022-0031)
- Archived working directory contains the results and all the data needed to replicate analysis on your PC
- User-friendly data formats (XLSX, txt, rnk)



## Download

<https://github.com/fgcz/prolfqua>  
<https://github.com/wolski/prolfquapp>

prolfqua

prolfquapp

