prolfquapp - Streamlining Protein Differential Expression Analysis in Core Facilities

Witold Wolski[1; 2], Bernd Roshitzki[1], Jonas Grossmann[1; 2], Claudia Fortes[1], Paolo Nanni[1], Christian Panse[1; 2], Ralph Schlapbach[1]

[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 doi.org/10.1021/acs.jproteome.2c00441
- 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 Markdown and runscript for DIANN diann-output.tsv
copy_DEA_FragPipe_DDA	copy Markdown and runscript for FragPipe combined_protein.tsv
copy_DEA_FragPipe_TMT	copy Markdown and runscript for Fragpipe TMT psm.txt
copy_DEA_MaxQuant	copy Markdown and runscript for MaxQuant peptide.txt proteinGroups.txt

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: ^CON|^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

Relative.Path	Name	Grouping/Var	CONTROL	Subject
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_B_03.raw	TripleProteome_B_DIABenchmark_1	B	T	ld1
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_B_02.raw	TripleProteome_B_DIABenchmark_2	B	T	ld2
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_B_01.raw	TripleProteome_B_DIABenchmark_3	B	T	ld3
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_A_04.raw	TripleProteome_A_DIABenchmark_1	A	C	ld1
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_B_04.raw	TripleProteome_A_DIABenchmark_4	B	T	ld4
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_A_02.raw	TripleProteome_A_DIABenchmark_2	A	C	ld2
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_A_01.raw	TripleProteome_A_DIABenchmark_3	A	C	ld3
F:/Exp02_High-Performance_uPAC_QE-HF_DIA_staggered_30to150minGradients_01_raw_Grad090_LFQ_A_05.raw	TripleProteome_A_DIABenchmark_4	A	C	ld4

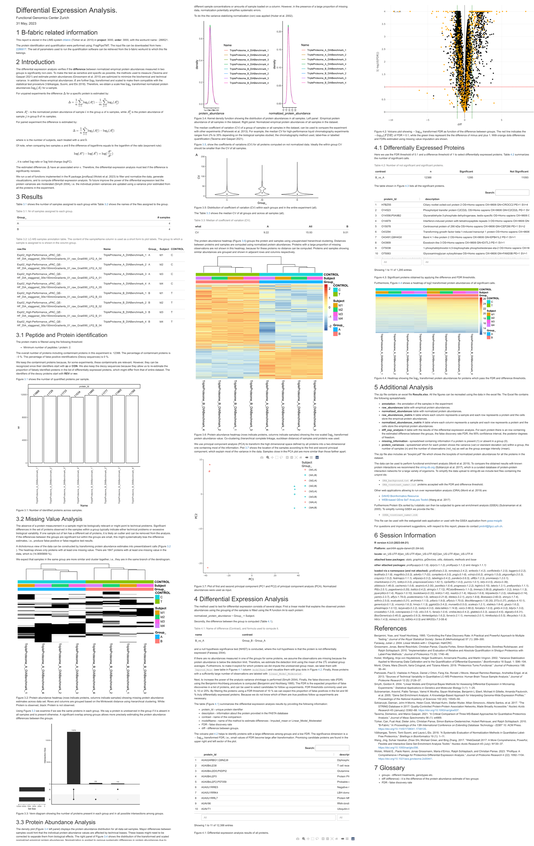
If subject is specified then the model is abundance ~ group + subject, otherwise abundance ~ group. The group differences to compute are determined from the group column and the control column.

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.: missingness, 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

Conlusion

- Integrates into LIMS system doi.org/10.1515/jib-2022-0031
- Archived directory contains all information needed to replicate analysis rerun the analysis on your PC
- Our users know Excel and like XLSX files
- Shiny app in development



Download

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

prolfqua prolfquapp

SIB Swiss Institute of Bioinformatics