Introduction to the Proteomics Toolset for Integrated Data Analysis (Protigy)

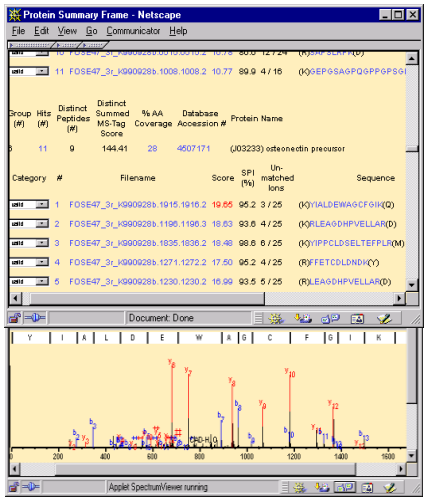
BroadE Proteomics Workshop
February 28, 2018

Proteomics Data Processing Flow

LC-MS/MS

- m/z and intensities of intact peptides
- m/z and intensities of peptide fragment ions

Spectrum Mill



- Peptide and protein identities
- Relative peptide/protein abundances

Downstream analysis

- Protein binding partners (AE-MS)
- Differentially expressed proteins (Discovery MS)

How can we streamline downstream data analysis?

User

Developer

- Easy-to-use (no coding skills required)
- Fast and reproducible analysis
- Interactive exploration of results (≠ static Excel sheets)

- Easy to maintain and extent
- Ability to ‘plug-in’ already developed code/scripts
- Flexible framework for different kinds of projects

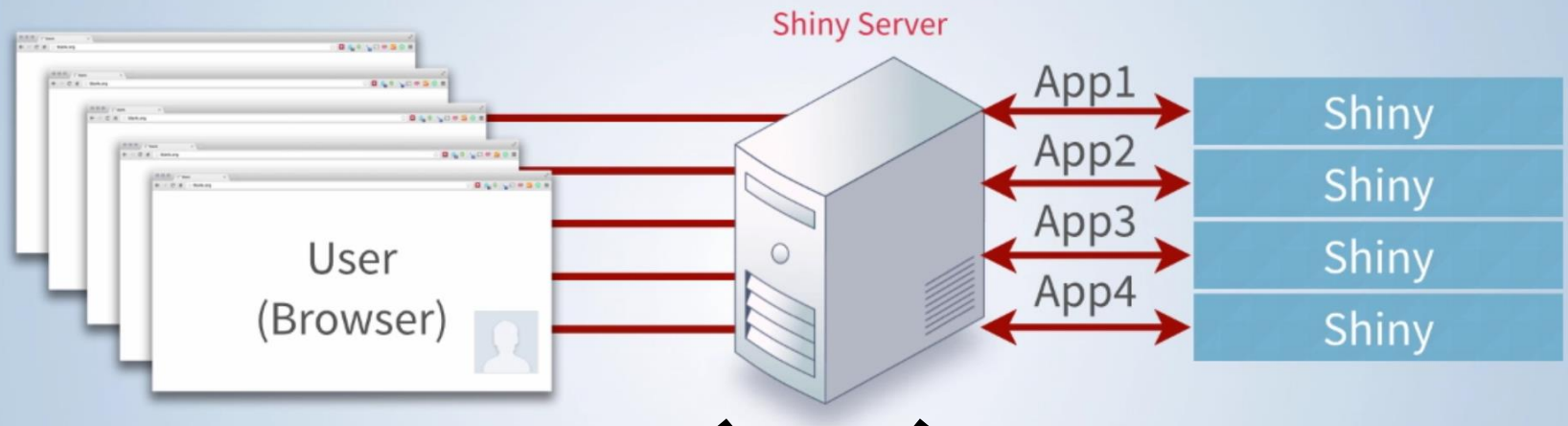
Shiny - Bring R data analysis to life

What is R-Shiny?

- Framework to develop R-powered, interactive web-applications
- Interactivity is a central feature (reactive programming)
- Building web interfaces by writing R-code (Shiny apps)
- Perform interactive data analysis in a web browser

https://www.rstudio.com/

Shiny Server Architecture



Open source version

- Everything to get started

Shiny Server Pro

- User authentication
- Load balancing
- …

https://www.rstudio.com/

Proteomics Toolset for Integrative Data Analysis

Protigy (v0.8.0.4) ≡ karsten@broadinstitute.org Logout Help me! BROAD PROTEOMICS

Upload file (txt, csv, gct):

Browse... No file selected

Saved sessions:

Search sessions

Import Manage sessions

To analyze another data set or to start over hit the F5 button.

PROTigy

PROteomics Toolset for InteGrative Data AnalYsis

This Shiny app facilitates exploratory and interactive analysis of data sets derived from quantitative *proteomics* experiments, *RNA-seq* and gene expression *microarrays*.

The app can run locally on your Desktop computer (Windows/Linux/MAC) or can deployed to Shiny Server environments. To access all implemented features the app has to run on a Shiny Server Pro (SSP) instance, see below for a summary of features only available in SSP.

Supported input formats:

- Any type of text file containing both, expression and annotation columns, can directly be imported into the app.
- Supported file formats:
  - text files (tsv, csv, txt)
  - gct 1.2
  - *NEW* gct 1.3

Data manipulation:

- Transformation
  - log transformation
- Sample-wise Normalization
  - Centering (median)
  - Centering and scaling (median-MAD)
  - Quantile
  - 2-component
- Filtering
  - Reproducibility filter across replicate measurements
  - Standard deviation across samples

Marker selection (based on *limma* package)

- One-sample moderated T-test
- Two-sample moderated T-test
- Moderated F-test

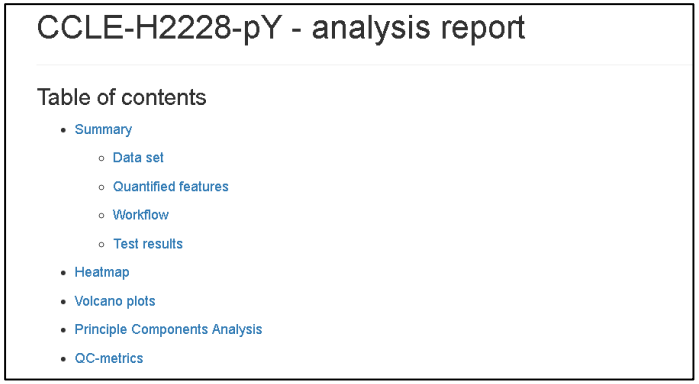
Interactive data analysis and visualization

- Heatmaps and cluster analysis
- Volcano plots
- Scatterplots
- QC-plots
  - Pairs-plots
  - Correlation matrix
  - Distribution of expression values
  - Missing values

http://shiny-proteomics.broadinstitute.org:3838/protigy/

Secure and Reproducible Data Analysis

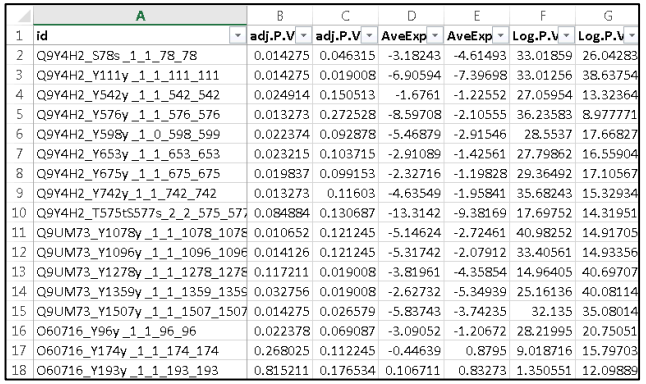
- Secure:
  - Google authentication – log-in with your Broad ID
- Reproducible:
  - R Markdown reports
  - Parameter file and R-session file to document workflow and parameters
- Export of results to HTML, PDF and Excel data formats



CCLE-H2228-pY - analysis report

Table of contents

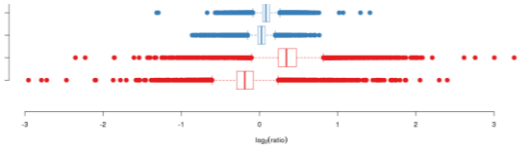
- Summary
  - Data set
  - Quantified features
  - Workflow
  - Test results
- Heatmap
- Volcano plots
- Principle Components Analysis
- QC-metrics



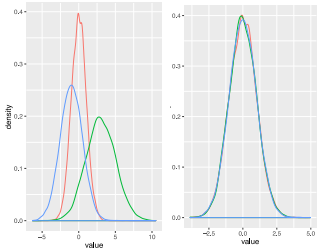
	A	B	C	D	E	F	G
1	id	adj.P.V	adj.P.V	AveExp	AveExp	Log.P.V	Log.P.V
2	Q9Y4H2_S78s_1_1_78_78	0.014275	0.046315	-3.18243	-4.61493	33.01859	26.04283
3	Q9Y4H2_Y111y_1_1_111_111	0.014275	0.019008	-6.90594	-7.39698	33.01256	38.63754
4	Q9Y4H2_Y542y_1_1_542_542	0.024914	0.150513	-1.6761	-1.22552	27.05954	13.32364
5	Q9Y4H2_Y576y_1_1_576_576	0.013273	0.272528	-8.59708	-2.10555	36.23583	8.977771
6	Q9Y4H2_Y598y_1_0_598_599	0.022374	0.092878	-5.46879	-2.91546	28.5537	17.66827
7	Q9Y4H2_Y653y_1_1_653_653	0.023215	0.103715	-2.91089	-1.42561	27.79862	16.55904
8	Q9Y4H2_Y675y_1_1_675_675	0.019837	0.099153	-2.32716	-1.19828	29.36492	17.10567
9	Q9Y4H2_Y742y_1_1_742_742	0.013273	0.11603	-4.63549	-1.95841	35.68243	15.32934
10	Q9Y4H2_T575tS577s_2_2_575_577	0.084884	0.130687	-13.3142	-9.38169	17.69752	14.31951
11	Q9UM73_Y1078y_1_1_1078_1078	0.010652	0.121245	-5.14624	-2.72461	40.98252	14.91705
12	Q9UM73_Y1096y_1_1_1096_1096	0.014126	0.121245	-5.31742	-2.07912	33.40561	14.93356
13	Q9UM73_Y1278y_1_1_1278_1278	0.117211	0.019008	-3.81961	-4.35854	14.96405	40.69707
14	Q9UM73_Y1359y_1_1_1359_1359	0.032756	0.019008	-2.62732	-5.34939	25.16136	40.08114
15	Q9UM73_Y1507y_1_1_1507_1507	0.014275	0.026579	-5.83743	-3.74235	32.135	35.08014
16	O60716_Y96y_1_1_96_96	0.022378	0.069087	-3.09052	-1.20672	28.21995	20.75051
17	O60716_Y174y_1_1_174_174	0.268025	0.112245	-0.44639	0.8795	9.018716	15.79703
18	O60716_Y193y_1_1_193_193	0.815211	0.176534	0.106711	0.83273	1.350551	12.09889

Cover all Aspects of Data Analysis

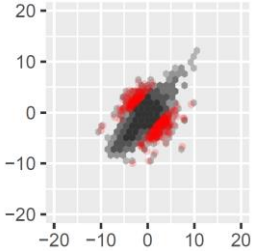
- Quality Control



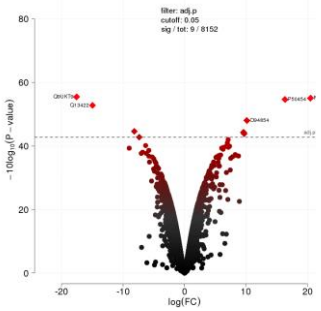
- Data transformation/normalization
  - Centering/scaling



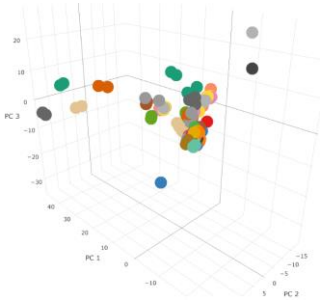
- Data filtering
  - Remove non-reproducible measurements to increase power



- Moderated test statistics
  - One-sample tests, two-sample, F-tests



- Interactive data visualization
  - Heatmaps, volcano plots, PCA …



Protigy Data Analysis Workflow

Import

Data

Exp1_REP1	Exp2_REP1	Exp2_REP2		Id
-0.139	-0.255	0.054	0.033	E9Q634
-0.088	-0.247	0.001	0.039	B1AR69
0.014	-0.234	-0.107	-0.007	P97479
-0.1	-0.33	0.233	0.121	P13541
-0.079	-0.33	0.145	0.057	Q91Z83
0.029	0.137	-0.029	-0.006	Q6PB66
0.064	-0.285	-0.259	-0.161	E9Q9F7
0.093	0.23	-0.188	-0.117	Q61699
-0.441	-0.49	0.075	-0.113	Q3V0K9
-0.251	-0.136	0.041	0.023	Q8K411
-0.008	0.069	0.046	0.072	Q60597
0.171	0.216	0.233	0.301	P22681-2
0.162	0.222	0.225	0.31	P22681-1
-0.008	0.095	0.038	0.029	Q8CAQ8-2

Colum.Name	Experiment
Exp1_REP1	A
Exp1_REP2	A
Exp2_REP1	B
Exp2_REP2	B

Exp design

Gene mapping

RefSeq
UniProt

Human
Mouse
Rat
Zebrafish

Data manipulation

- Log transformation
- Data normalization
- Data filtering

Marker selection

Moderated T and F statistics

Significance Filter

p-value, FDR, top N

QC

- Data distributions
- P-values
- Correlation matrix

Exploratory data analysis

- Cluster analysis
- Protein-protein-interactions
- Principle component analysis

Export

- Markdown (html)
- Excel
- Zip (pdf, txt, xlsx)

Gene symbol mapping

- Protigy tries to automatically map protein accession numbers to gene symbols

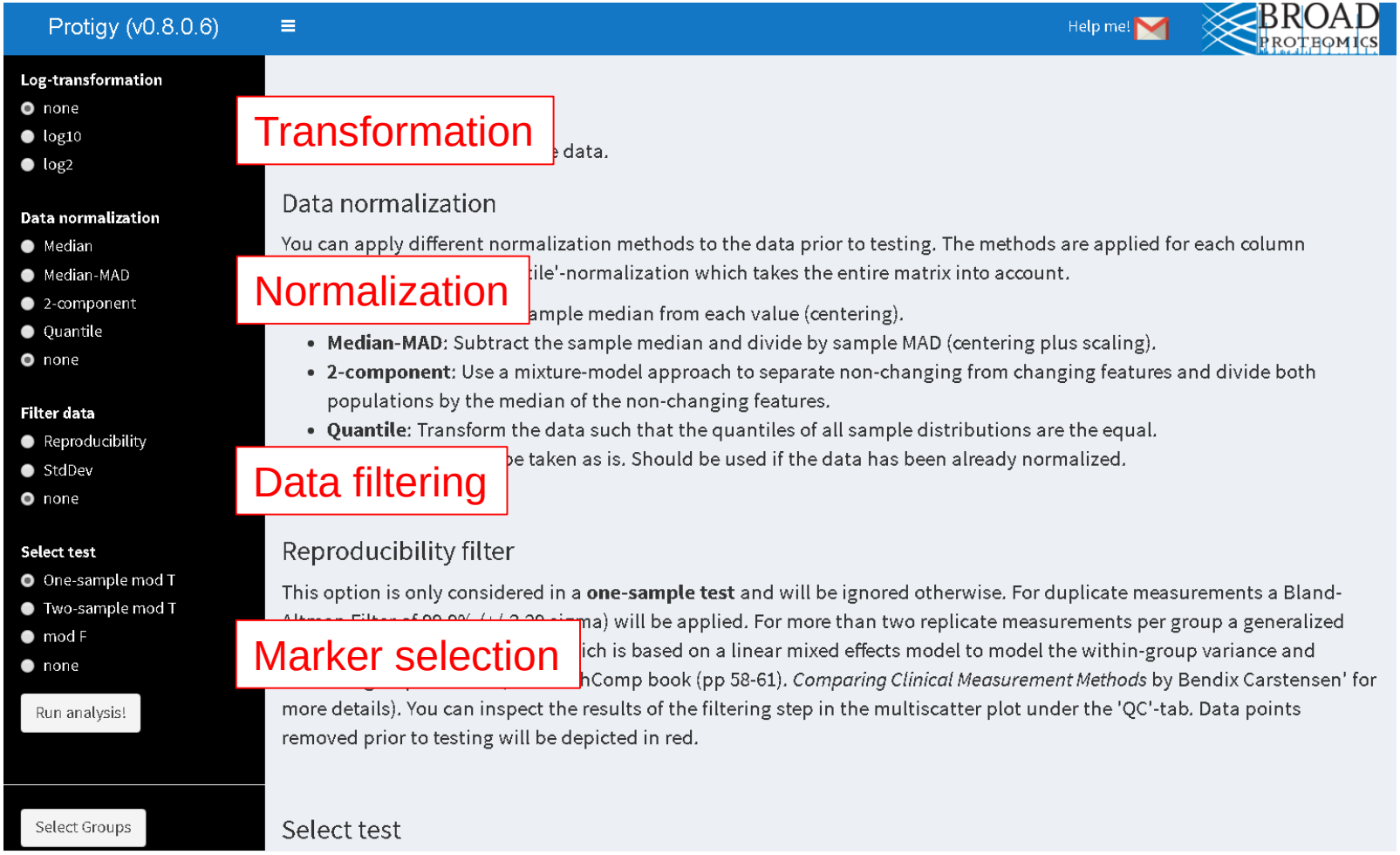
- Mapping based on Bioconductor 3.6 `orgDb` annotation packages

- Supported protein accessions:
  - UniProt `http://www.uniprot.org/`
  - RefSeq `https://www.ncbi.nlm.nih.gov/refseq/`

- Supported organisms (Feb 2018):
  - Human, mouse, rat, zebrafish

- Primary Protigy IDs: *proteinAccession_geneSymbol*

Setting up the Analysis Workflow



Data manipulation

- Transformation
  - log transformation

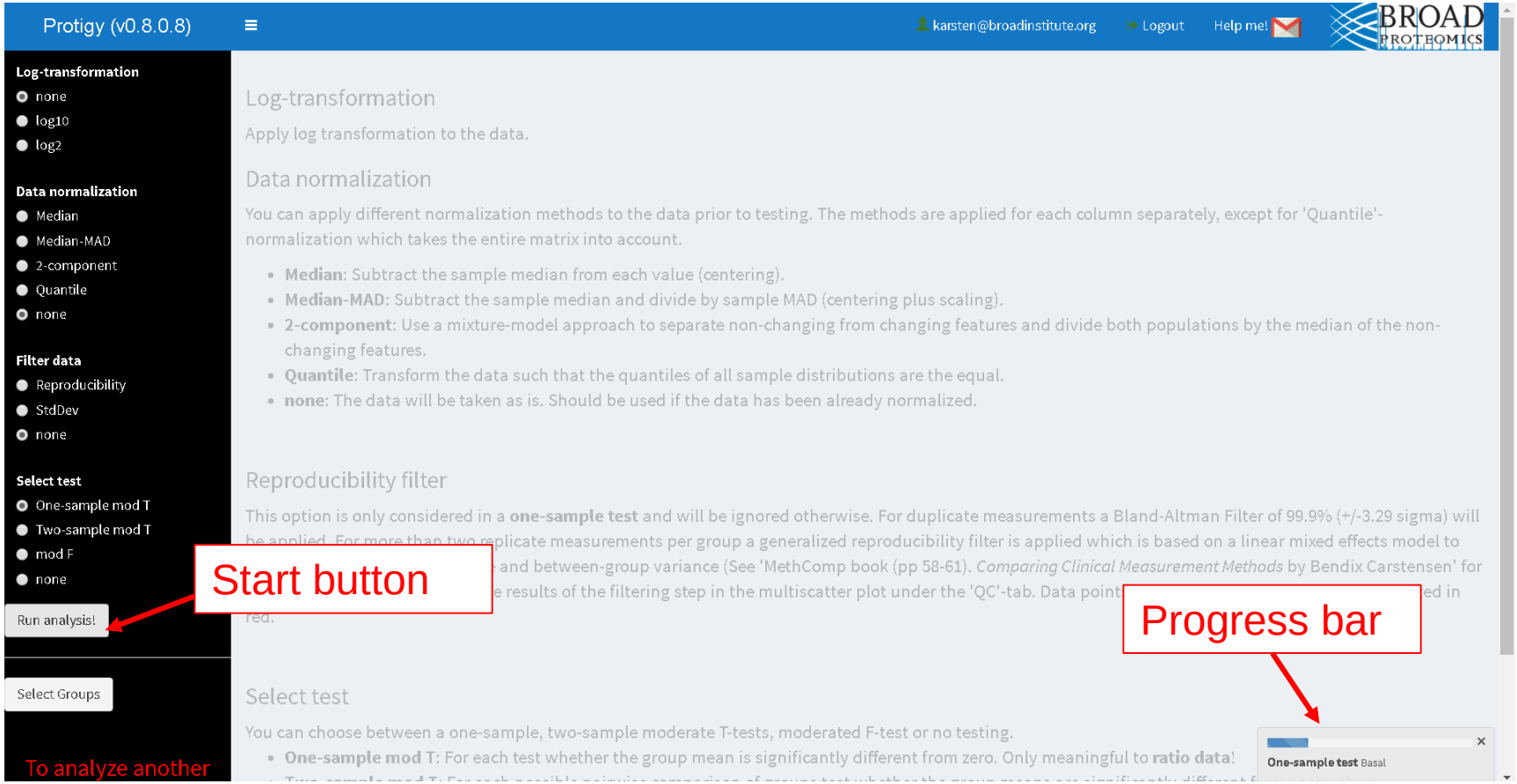
- Sample-wise Normalization
  - Centering (median)
  - Centering and scaling (median-MAD)
  - Quantile normalization
  - 2-component normalization

- Filtering
  - Reproducibility filter across replicate measurements
    - remove non-reproducibly measured features
  - Standard deviation across samples
    - remove features with low variance across samples

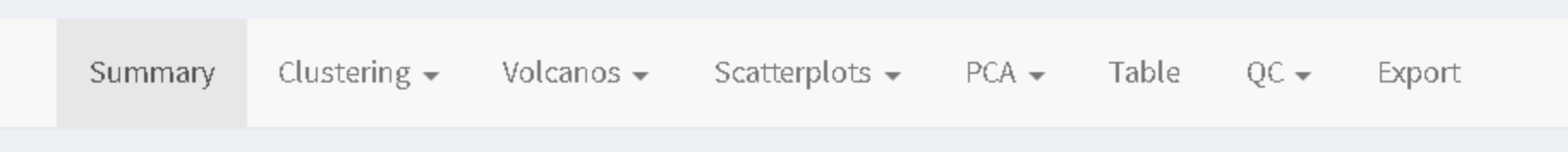
Marker Selection (based on *limma* R-package)

- One-sample moderated T-test
  - Is the log(case/control) ratio statistically different from 0?
- Two-sample moderated T-test
  - Is log (A/control) ratio statistically different from log (B/control)?
- Multiple Group comparison: Moderated F-test
  - Are any of the log (group$_i$ / reference) ratios statistically different from 0?
    - i = 1, 2, …, k
    - k = total number of groups

Running the analysis



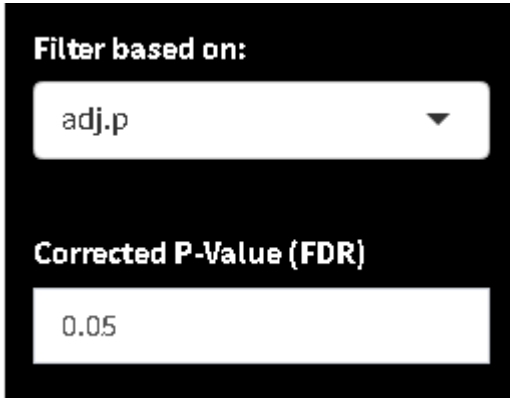
Navigation via Tabs



- Cluster analysis
  - Heatmap
  - Fanplot
- Volcano plots
  - Protein-protein-interactions
- Scatterplots
  - Protein-protein-interactions
- Principle component analysis
- Preview result table
- QC metrics
  - Data distributions
  - Correlation matrix
  - ...

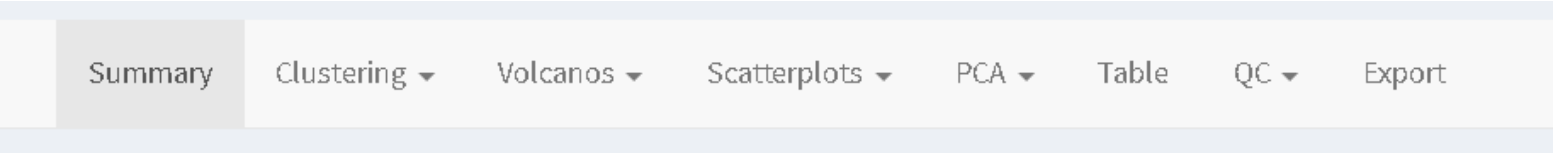
Set Significance Thresholds

- Apply filter to results of the test statistic
  - adjusted p-values (FDR)
  - nominal p-values
  - top N
  - none

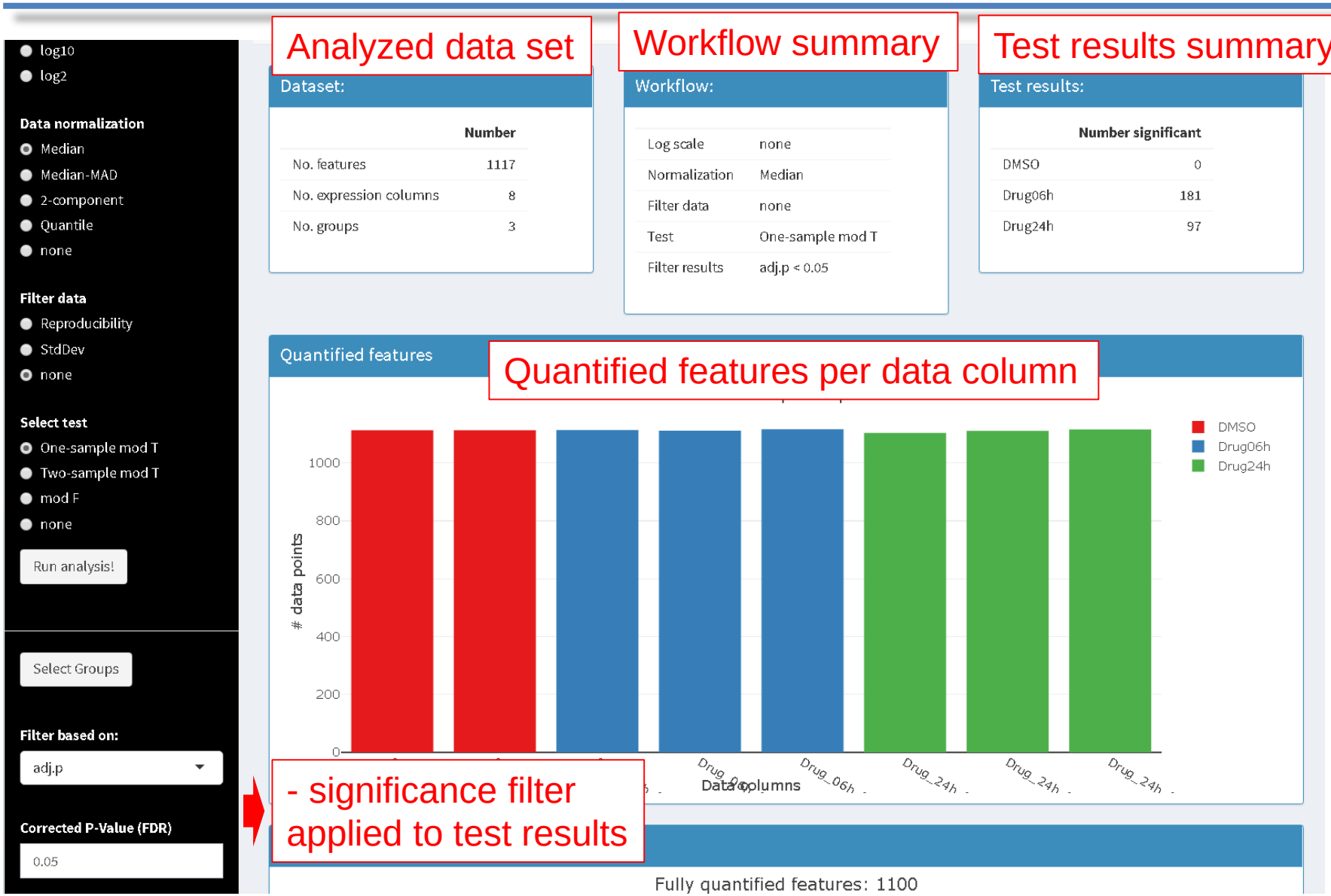


- If a test is applied to **multiple groups** (e.g. multiple one-sample tests) the filter setting will be applied to each test result **separately**.

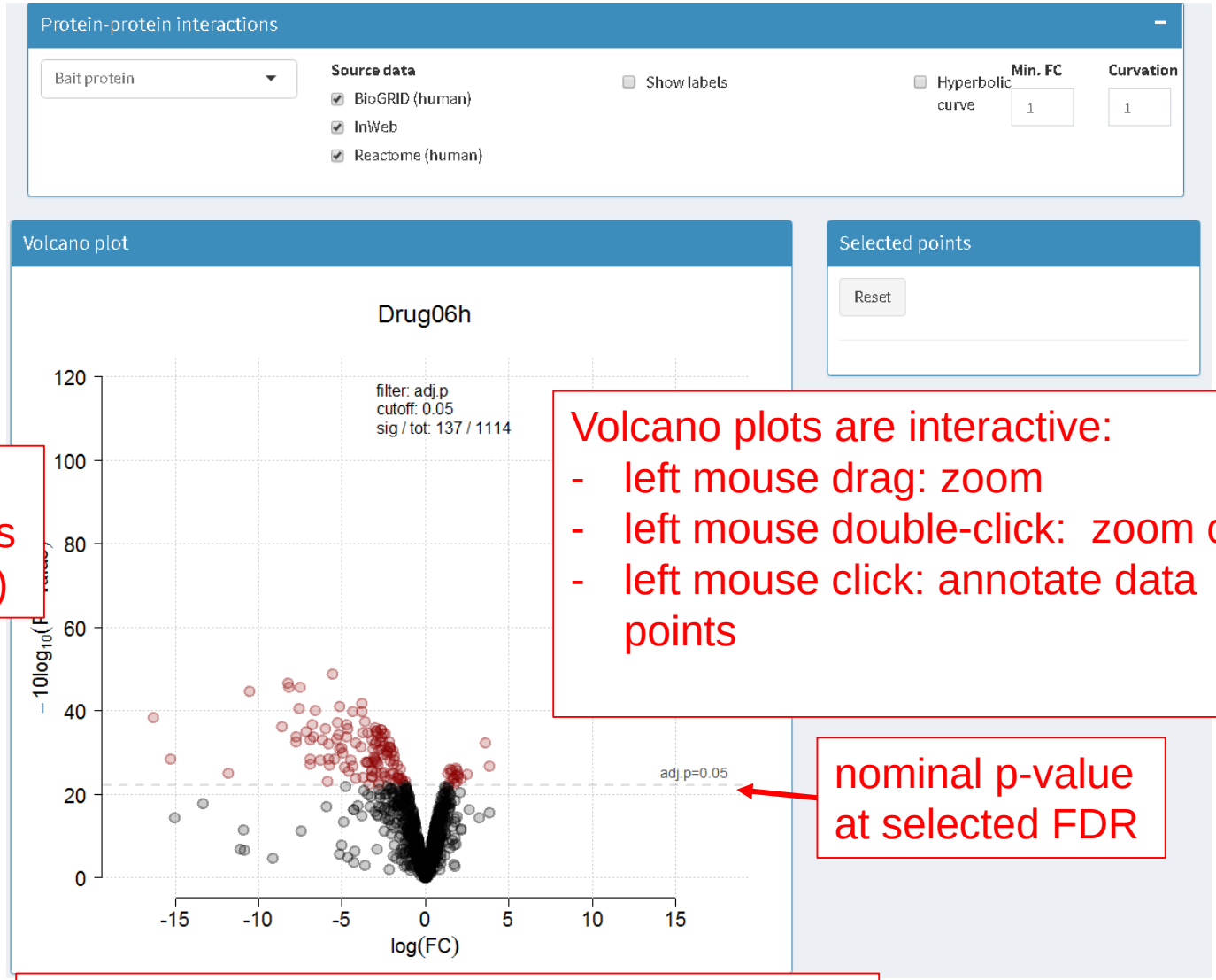
- This setting applies to all Tabs (except QC)



Summary Page



Exploratory Data Analysis – Interactive Volcano Plots



y-axis depicts nominal p-values (-10log10(p-val))

Volcano plots are interactive:

- left mouse drag: zoom
- left mouse double-click: zoom out
- left mouse click: annotate data points

nominal p-value at selected FDR

x-axis depicts log fold change drug/DMSO

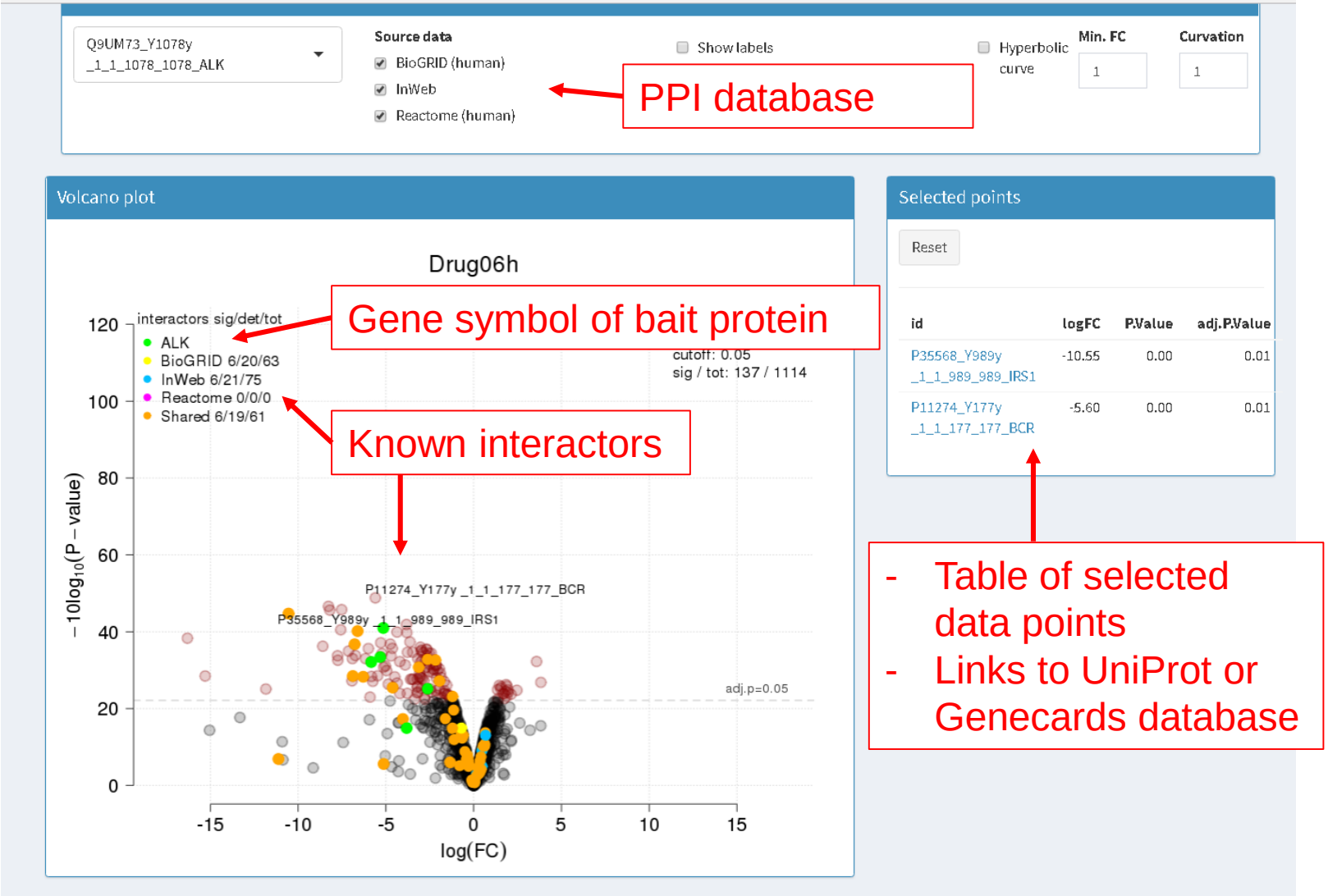
Integration of Protein-Protein Interaction Networks

- Routinely incorporate known protein-protein interactions into analyses of affinity proteomics experiments

- Integration of three major PPI databases
  - InWeb_IM – well curated resource of ~585K human protein-protein interactions
  - BioGRID - curated set of physical and genetic interactions
  - Reactome - computationally generated interactions (not curated and not from experimental data)

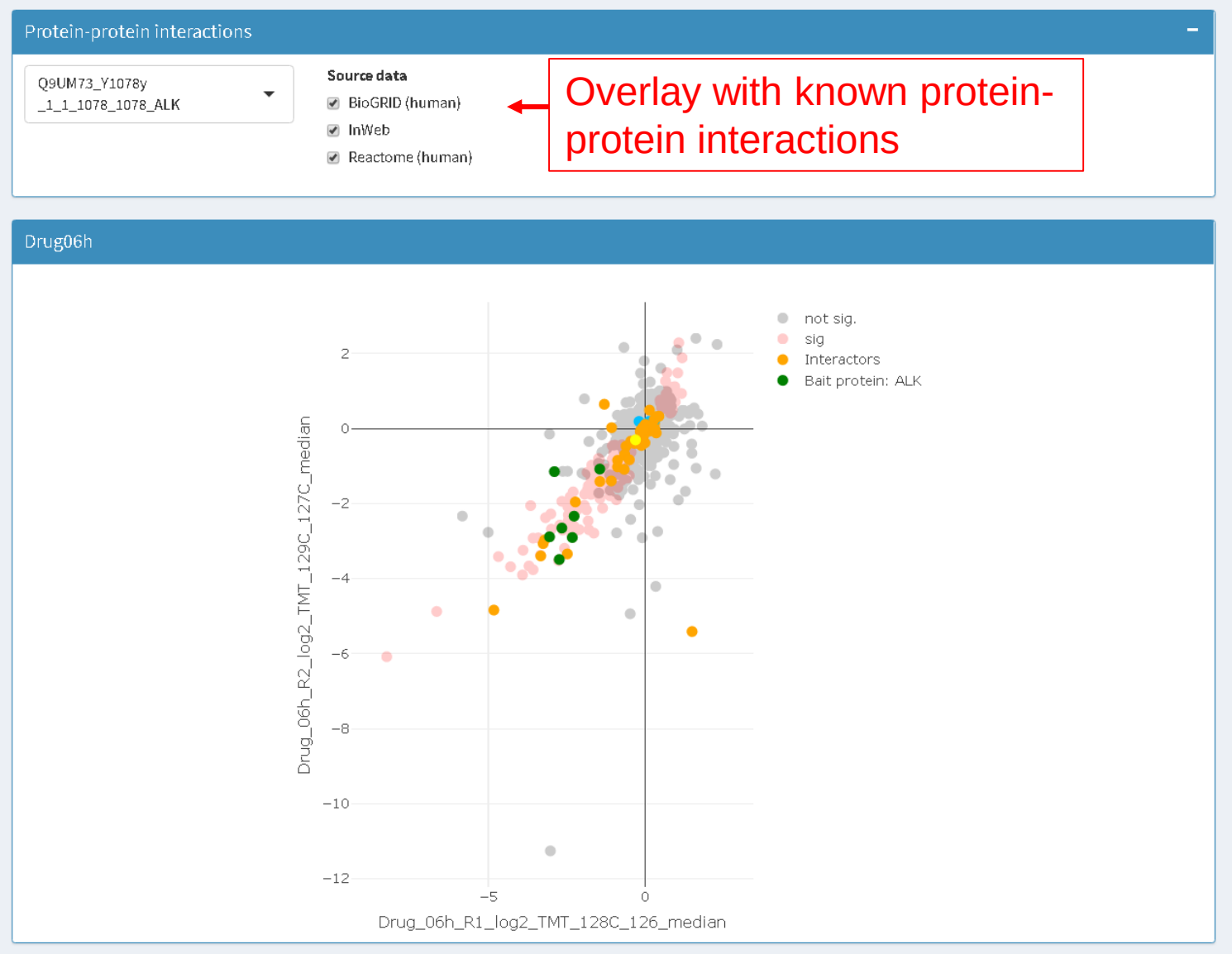
- Overlay known protein-protein interactors of the bait protein on top of proteomics results

Integration of Protein-Protein Interaction Networks

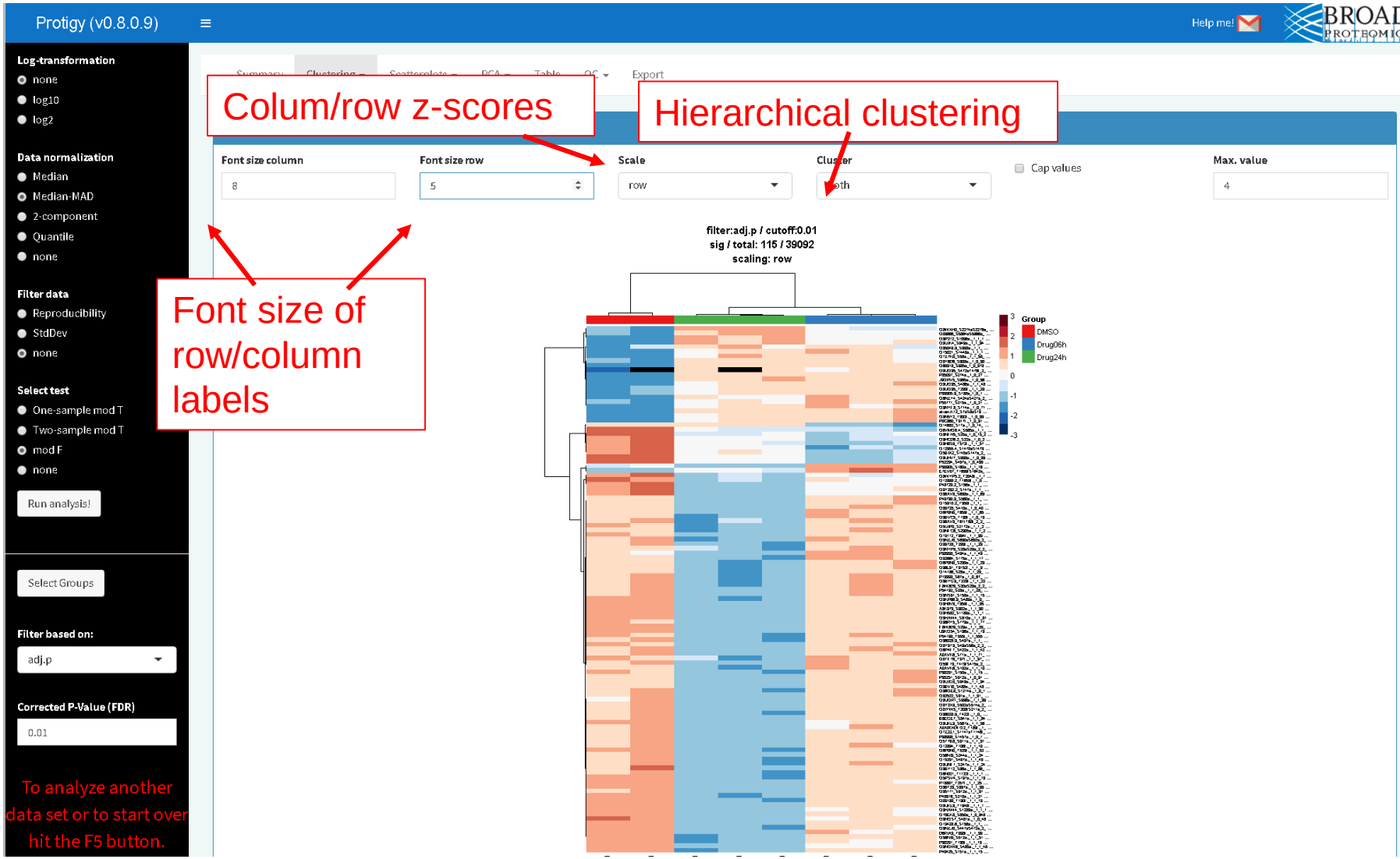


Exploratory Data Analysis – Interactive Volcano Scatterplots

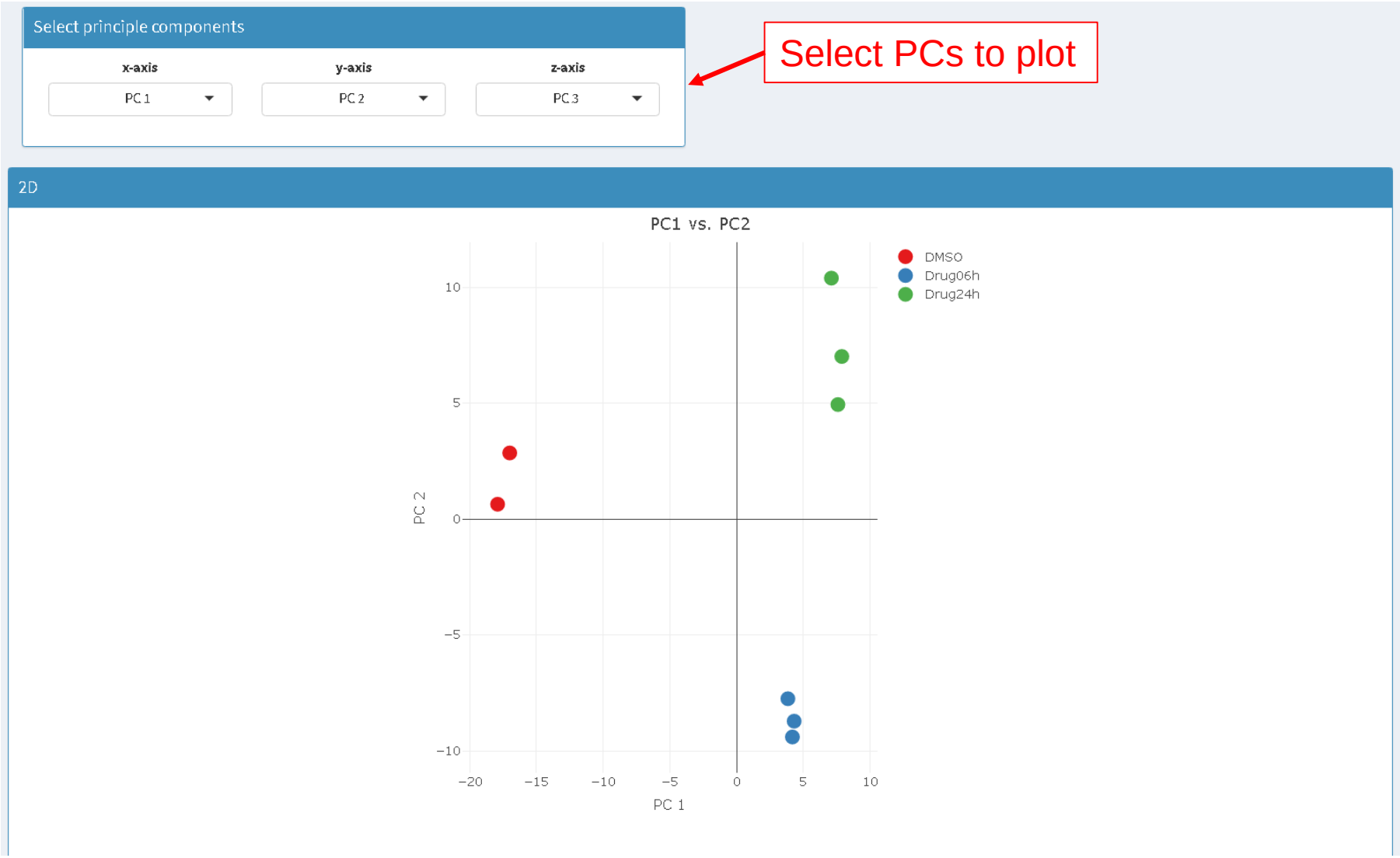
Exploratory Data Analysis – Interactive Volcano Scatterplots



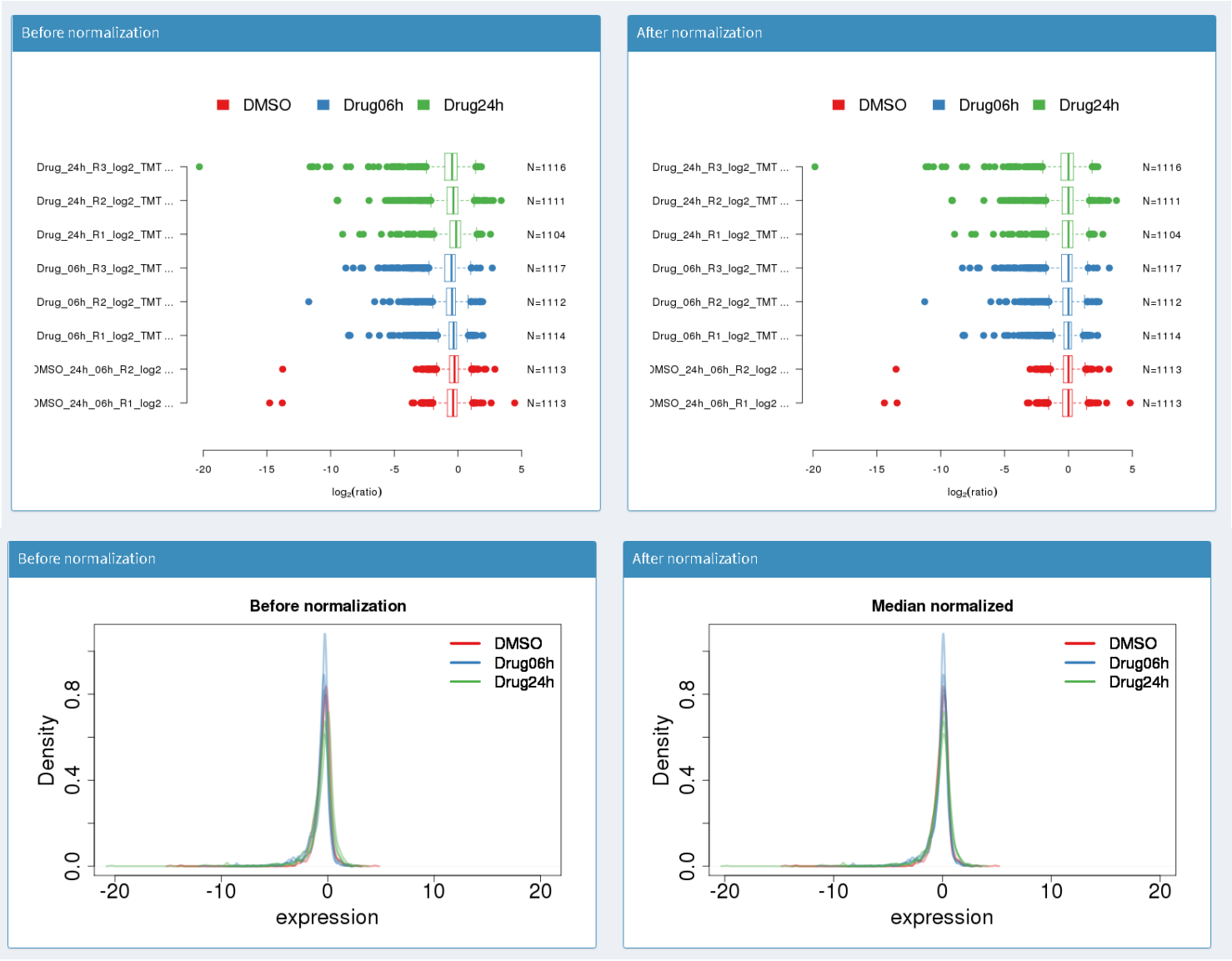
Exploratory Data Analysis– Heatmaps



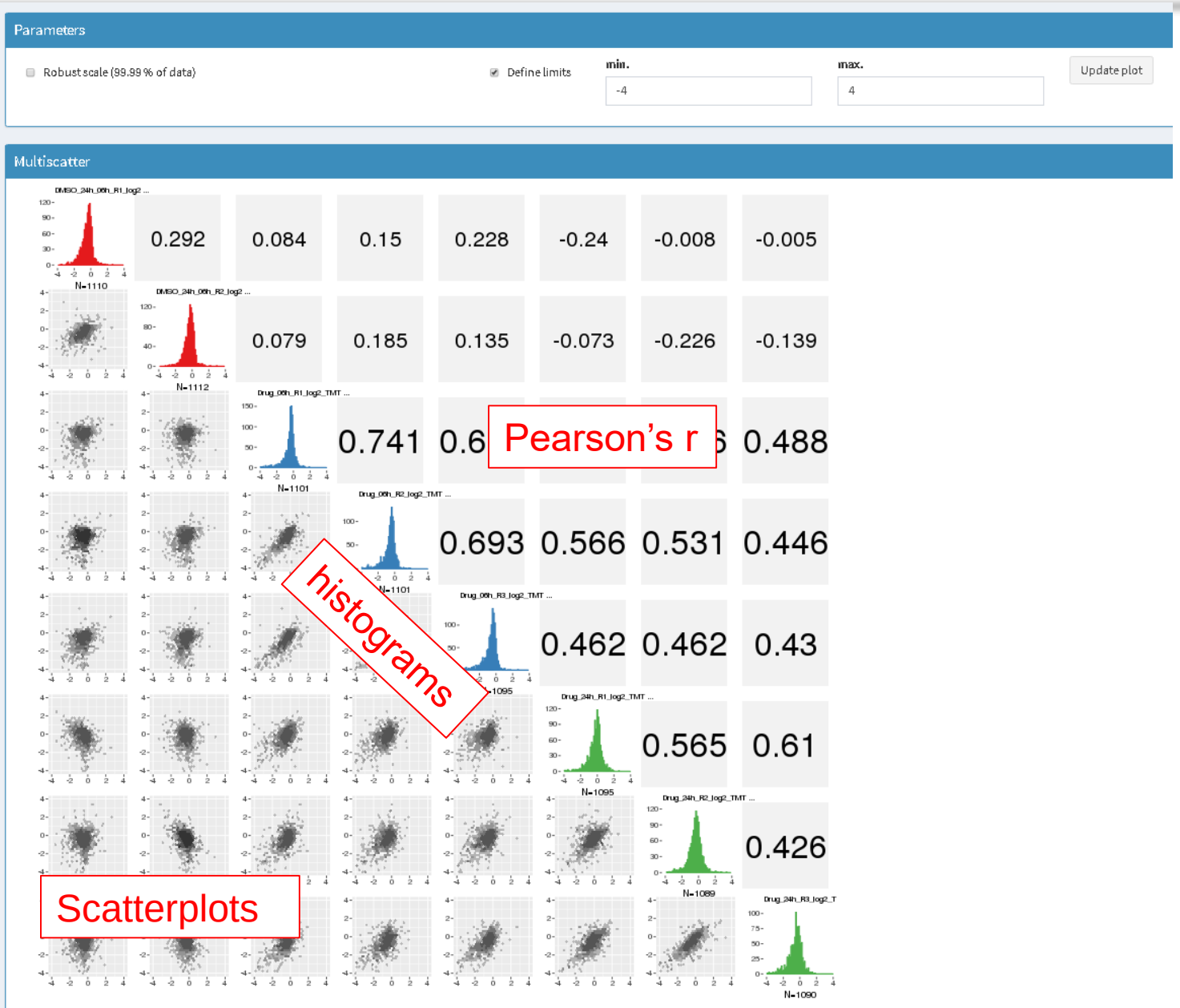
Exploratory Data Analysis - Principle Component Analysis



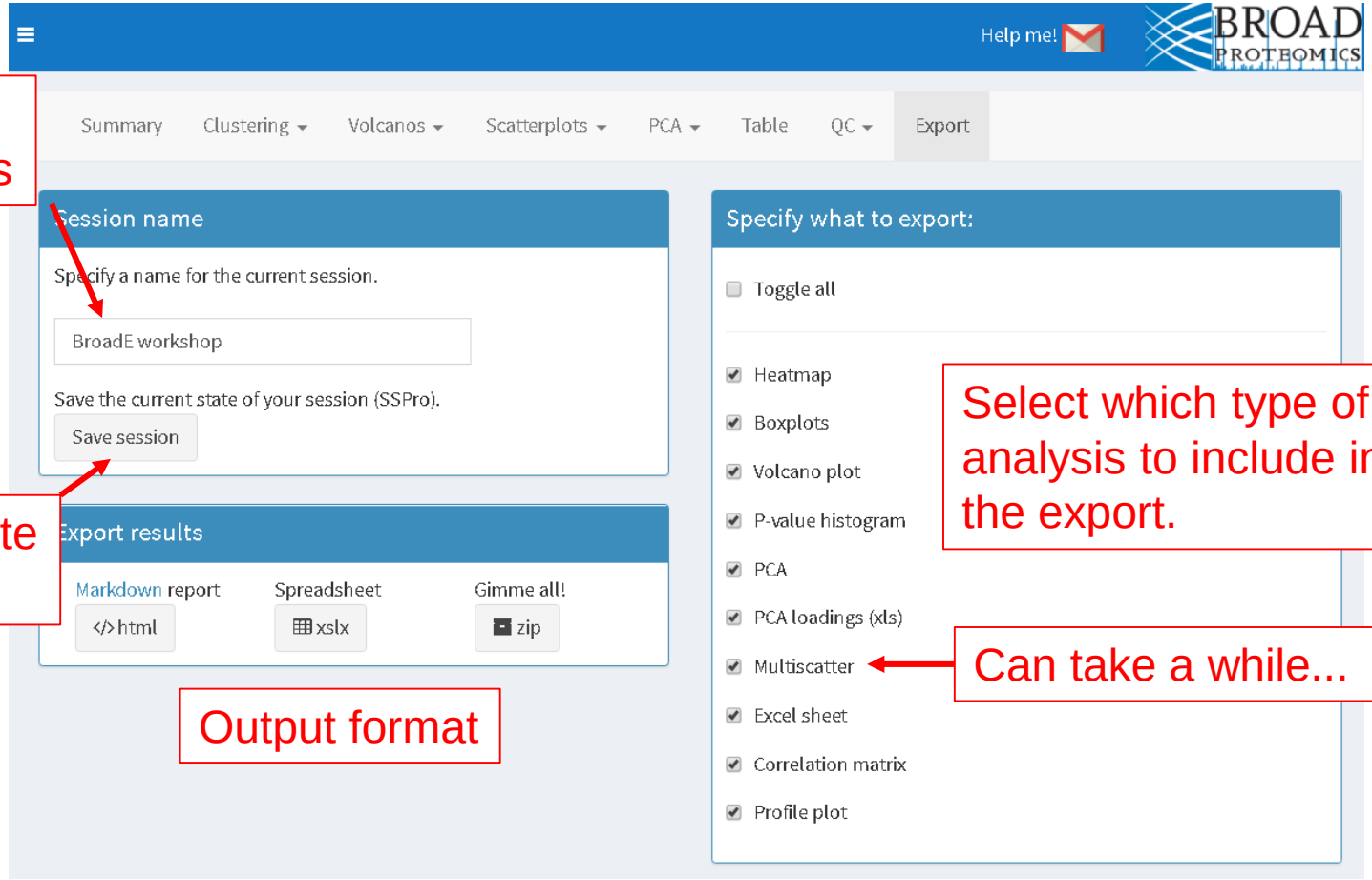
QC metrics – Data Distribution



QC metrics – Pairwise Correlation



Data export



- Export is based on the current session state, i.e. all user customizations will be visible in the export (annotations in volcano plots, heatmap scaling, ...)

R Markdown reports

- R Markdown is a file format for making dynamic documents with R.

- An R Markdown document is written in markdown (an easy-to-write plain text format) and contains chunks of embedded R code.

- Protigy uses R markdown to create interactive html reports that summarize your analysis.

```
## Residuals

To motivate the use of models we're going to start with an interesting pattern from the NYC flights dataset -- the number of flights per day.

```{r}
library(nycflights13)
library(lubridate)
library(dplyr)

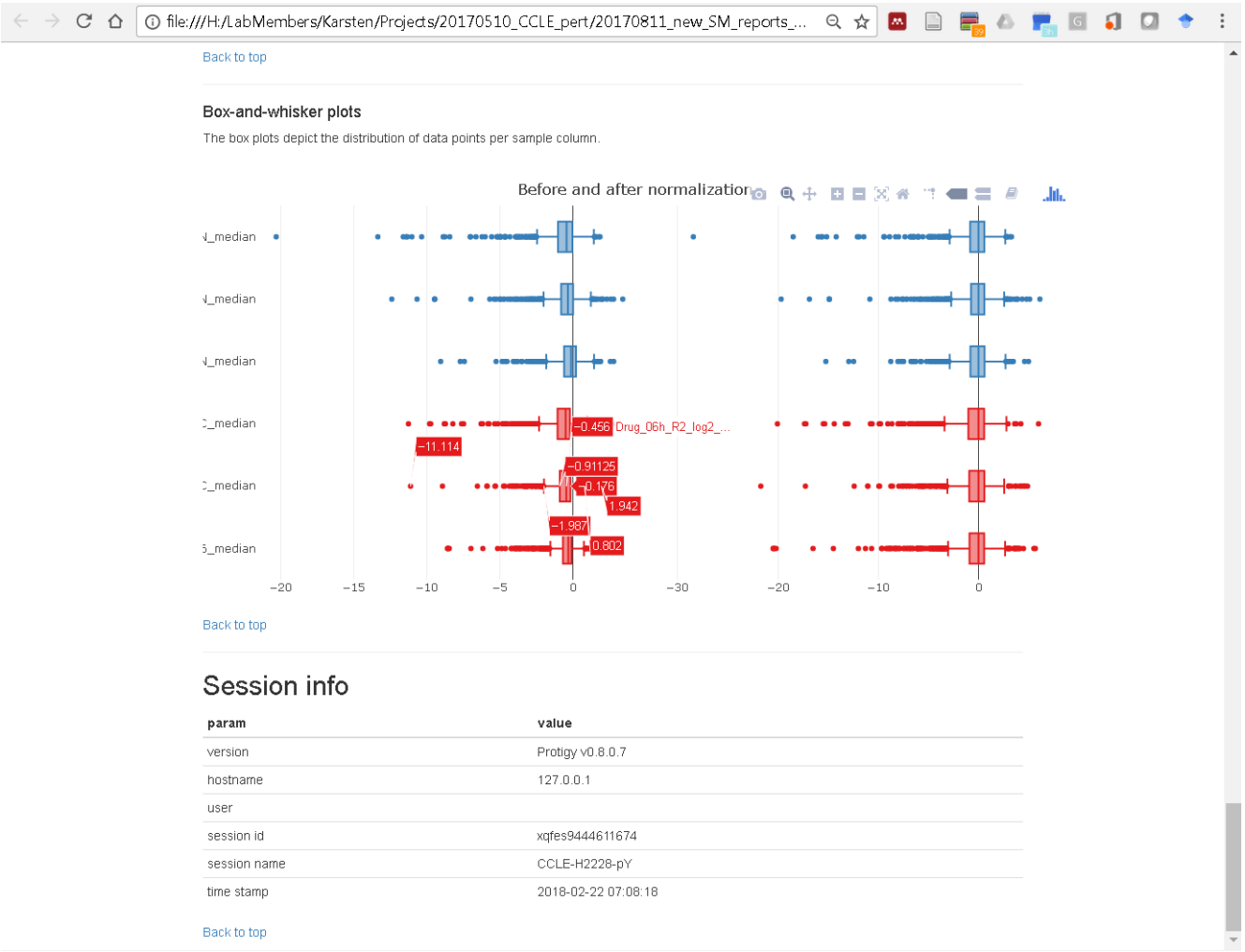
daily <- flights %>%
  mutate(date = make_datetime(year, month, day)) %>%
  group_by(date) %>%
  summarise(n = n())

ggplot(daily, aes(date, n)) +
  geom_line()
```
```

https://rmarkdown.rstudio.com/

R Markdown reports

- R Markdown reports can be downloaded as single html-file and opened in any web browser.



Shiny Server Professional (SSP)

- Running Protigy on SSP provides some exclusive features
  - User authentication (via Google)
  - Save sessions on server
  - Load sessions from server
  - Share sessions with collaborators
- Access to the SSP@ Proteomics Platform is currently limited to our collaborators

Running Protigy on a local PC/Mac

Software requirements:

- R >3.4 (https://cran.r-project.org/)
- Shiny R-package : `install.packages("shiny")`
- Pandoc (optional, required to create R Markdown reports)
  - https://github.com/jgm/pandoc/releases/tag/2.1.1
- Perl (optional, required to create Excel sheets)
  - http://strawberryperl.com (Windows OS)

To run Protigy directly from GitHub open R and run:

```
shiny::runGitHub("protigy", "karstenkrug")
```

- Please follow the instructions to make sure all required R packages will get installed.
- This process might take several minutes when you run Protigy for the first time.

Summary

- R-shiny provides a powerful and flexible framework to streamline data analysis
  - Fast QC
  - Standardized workflows
  - Flexible and versatile – not restricted to affinity proteomics experiments
- Interactivity is a key feature of Shiny
  - Facilitates exploratory data analysis
- Common platform for project managers and collaborators
  - Currently in beta testing phase

Further reading

- RStudio
  https://www.rstudio.com/

- Shiny tutorial
  http://shiny.rstudio.com/tutorial/

- Shiny gallery – small example applications
  http://shiny.rstudio.com/gallery/