

Statistical analysis with MSstats

Design and analysis of quantitative proteomic experiment

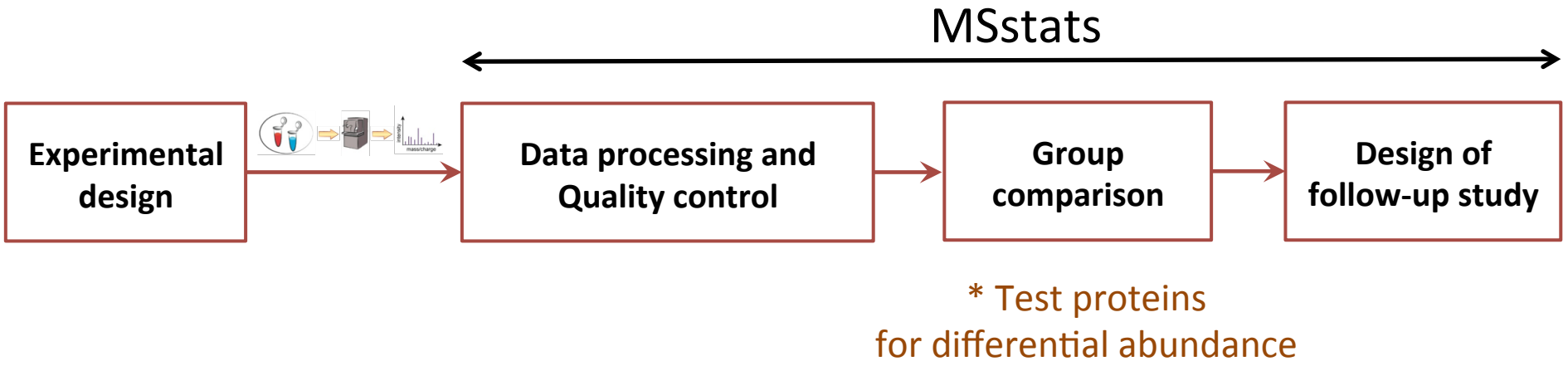
US HUPO short course

Meena Choi
Department of Statistics, Purdue University
Group of Prof. Olga Vitek

PURDUE
UNIVERSITY®

MSstats : statistical tool for quantitative MS proteomics

Open-source R-based package for **statistical relative quantification** of peptides and proteins in mass spectrometry-based proteomic experiments.



I. MSstats compatible experiments
II. MSstats input

I. Data transformation
II. Intensity normalization
III. QC Plot
IV. Feature selection
V. Missing values imputation
VI. Run-level summarization
VII. Profile Plot
VIII. Condition Plot

I. Statistical model
II. Results table
III. Volcano plot
IV. Heatmap
V. Comparison plot

I. Power analysis
II. Statistical power
III. Sample size

MSstats compatible experiments

MS acquisition

- DDA or shotgun
- SRM
- DIA or SWATH

Workflows

- label-free
- label-based
  - With isotope-labeled reference peptides

Experimental designs

- Case-control
- Time course
- Paired analysis
- More than 2 conditions
- Linear combination of conditions

Not supported

- multi-labelling
- iTRAQ

MSstats input : 10 columns are required



- Results from any spectral processing tool:
  e.g. Skyline, MaxQuant, Progenesis, OpenSWATH, Spectronaut
- Tabular .csv format

	Protein	Peptide	Precursor charge	Fragment	Product charge	Label	Condition	Subject	Run	Intensity
1	ProteinName	PeptideSequence	PrecursorCharge	FragmentIon	ProductCharge	IsotopeLabelType	Condition	BioReplicate	Run	Intensity
2	ACEA	EILGHEIFFDWELP	3	y3	0	H	1	ReplA	1	66472.3847
3	ACEA	EILGHEIFFDWELP	3	y3	0	L	1	ReplA	1	5764.16228
4	ACEA	EILGHEIFFDWELP	3	y4	0	H	1	ReplA	1	101005.166
5	ACEA	EILGHEIFFDWELP	3	y4	0	L	1	ReplA	1	61.65238
6	ACEA	EILGHEIFFDWELP	3	y5	0	H	1	ReplA	1	90055.4993
7	ACEA	EILGHEIFFDWELP	3	y5	0	L	1	ReplA	1	472.691803
8	ACEA	TDSEAATLISSTID	2	y10	0	H	1	ReplA	1	43506.5425
9	ACEA	TDSEAATLISSTID	2	y10	0	L	1	ReplA	1	217.203553
10	ACEA	TDSEAATLISSTID	2	y7	0	H	1	ReplA	1	68023.0377
11	ACEA	TDSEAATLISSTID	2	y7	0	L	1	ReplA	1	725.284308
12	ACEA	TDSEAATLISSTID	2	y8	0	H	1	ReplA	1	68276.0489
13	ACEA	TDSEAATLISSTID	2	y8	0	L	1	ReplA	1	243.658527

For DDA, ‘Fragment’, ‘ProductCharge’ can be any one value, such as NA

1. Data processing and quality control (QC)



I. Data transformation

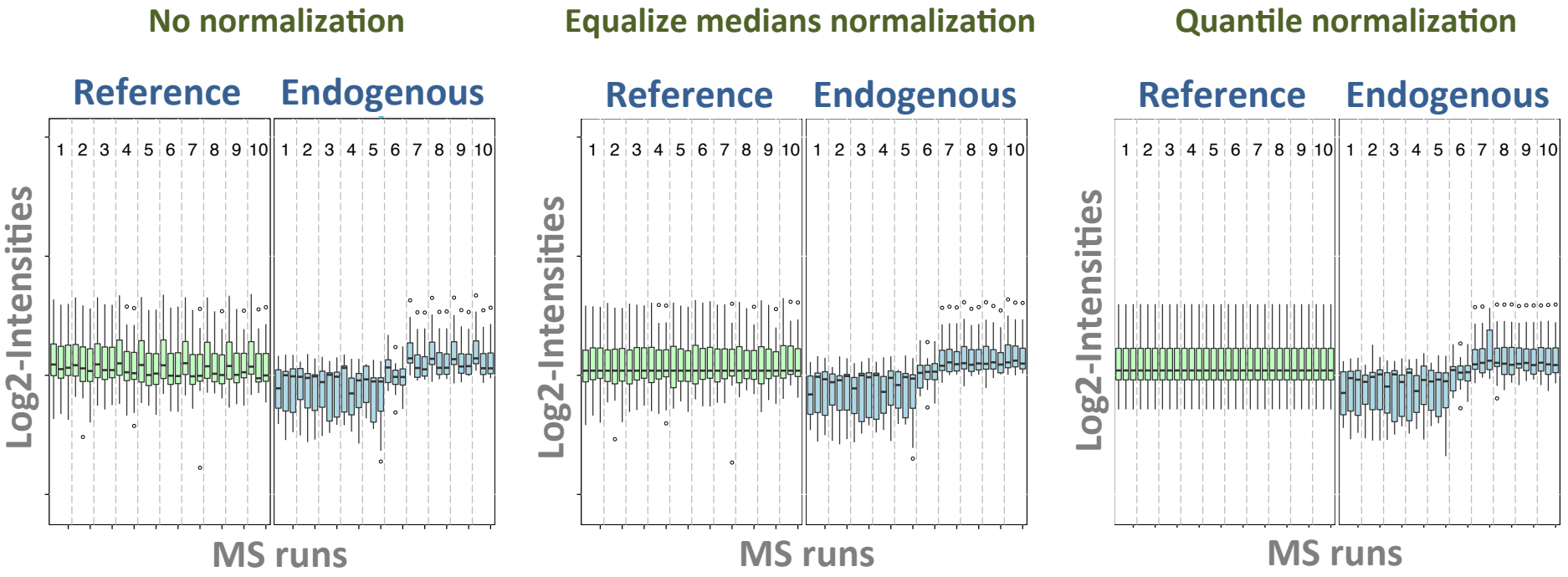
- Import the data for analysis
- Log 2 or 10 transformation of intensities

II. Intensity normalization

- **None** : no normalization is performed.
- **equalizeMedians** : make the same median of
  - reference intensities across runs for label-based experiments
  - Endogenous intensities across runs for label free experiments
- **Quantile** : equalize the distribution of
  - reference intensities across runs for label-based experiments
  - Endogenous intensities across runs for label free experiments
- **Global Standards** : applied to endogenous intensities. Equalize endogenous intensities of global standard protein across runs. Then apply the same between-run shifts to the remaining endogenous proteins.

Quality control plots

- ✓ Distribution of intensities per run
- ✓ Show potential systematic biases between mass spectrometry runs
- ✓ Show how the normalization works for all the proteins combined



Normalization method should be changed based on your design of experiment

!! **equalizeMedians** and **Quantile** normalization assume constant protein abundance across all samples

Only true if you

a) have a labeled reference or protein standards

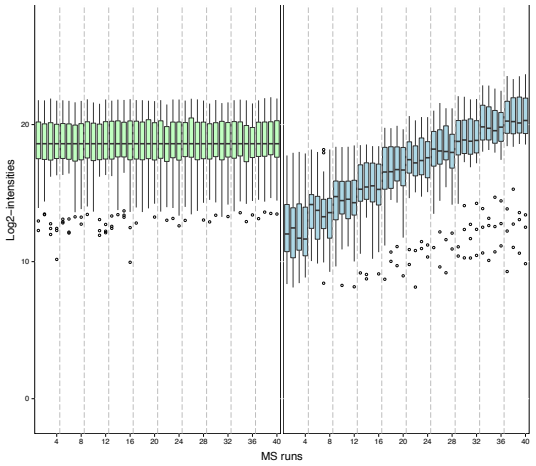
b) expect the majority of your proteins to be constant

Assume label-free SRM :
- **most features are differently abundant**

Equalize medians normalization

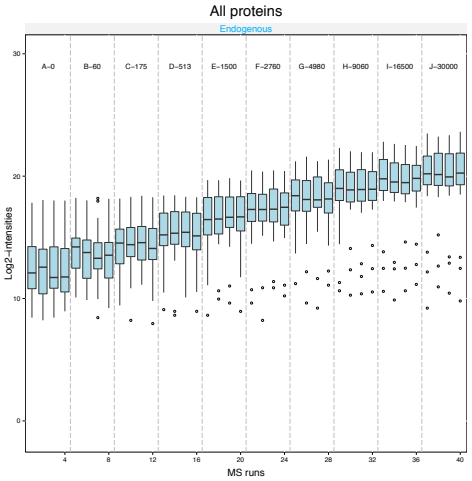
Reference **Endogenous**

All proteins

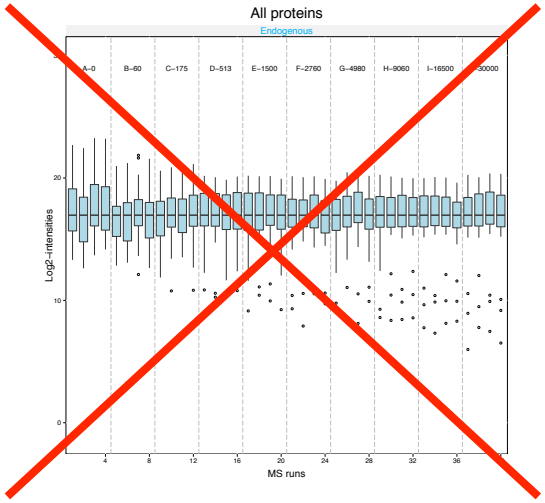


MS runs

No normalization



Equalize medians normalization



Normalization method should be changed based on your design of experiment

!! **equalizeMedians** and **Quantile** normalization assume constant protein abundance across all samples

Only true if you

a) have a labeled reference or protein standards

b) expect the majority of your proteins to be constant

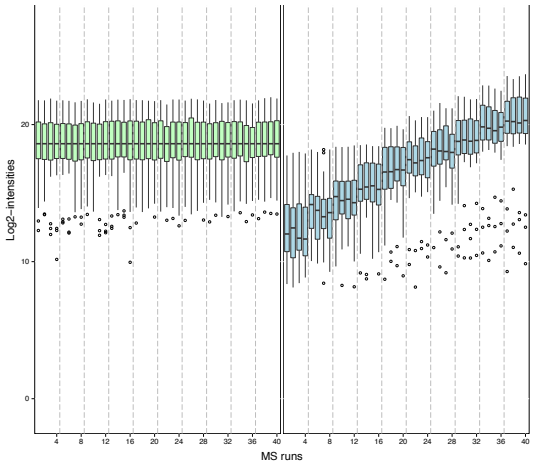
Assume label-free SRM :

- Need to concern normalization method in design stage

Equalize medians normalization

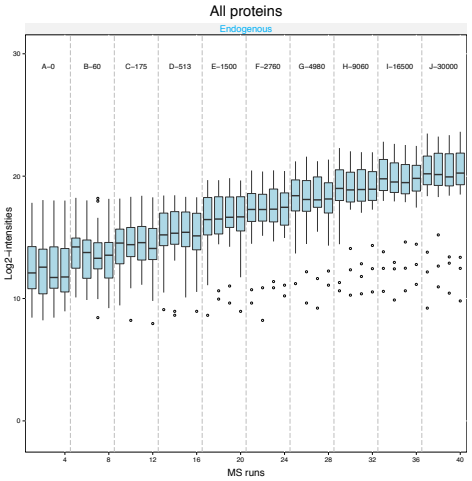
Reference **Endogenous**

All proteins

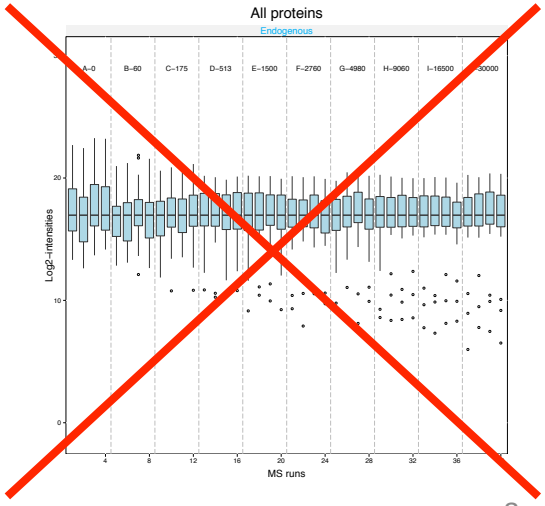


MS runs

No normalization



Equalize medians normalization



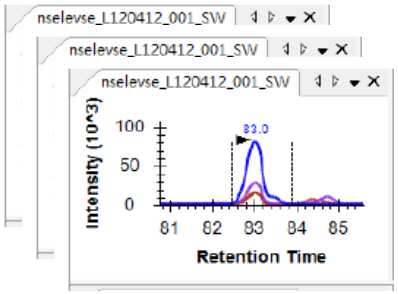
Feature selection

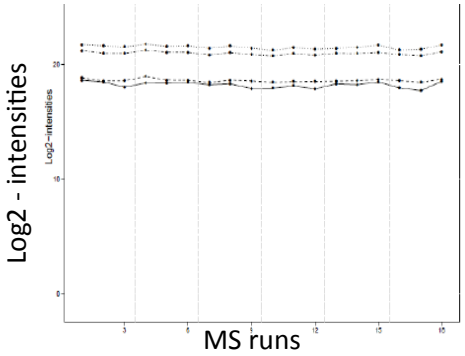
Which features are used for relative quantification and differential expression analysis?

- all features
- Top3 features : highest average of log2(intensity) across runs
- Features without interference

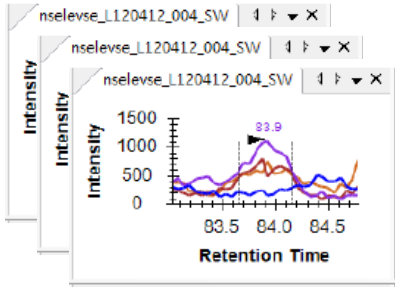
How to detect features with interference? Quantify how parallel features

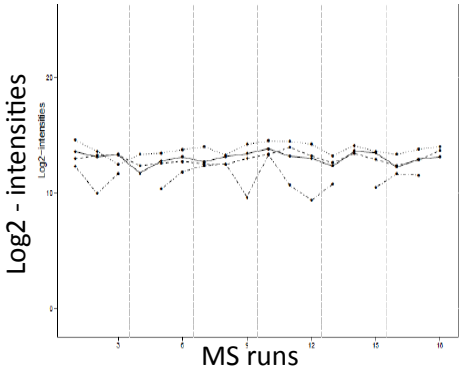
- Clear peaks have almost parallel profiles





- Features with interference has unparalleled profiles



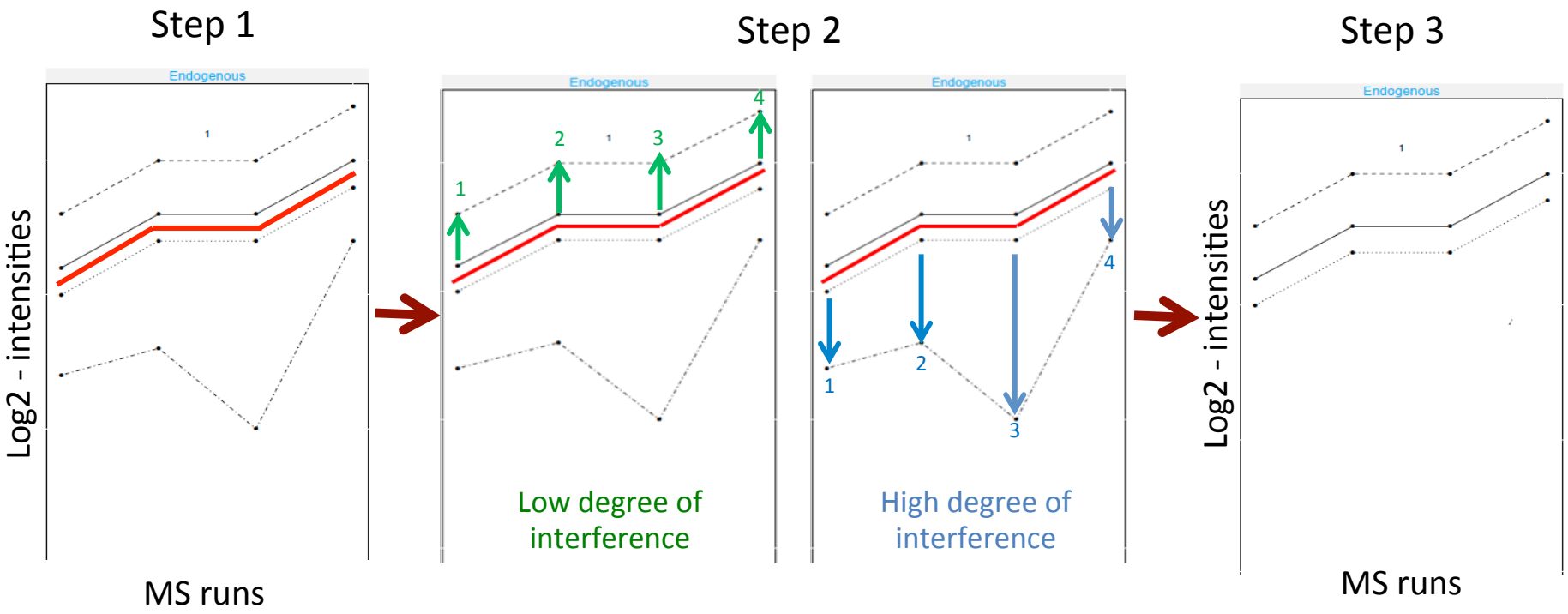


How to detect interference

Step 1: Summarize abundance of each peptide

Step 2: Estimate the degree of interference = variance of residuals

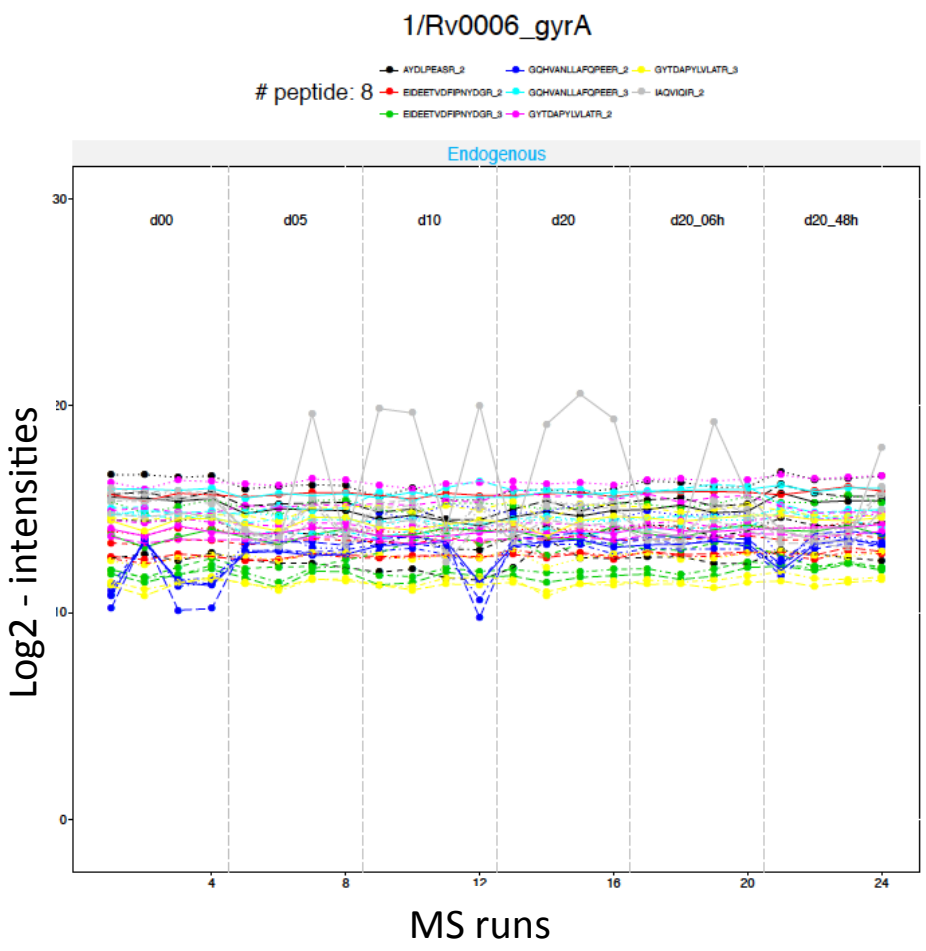
Step 3: Remove features with highest interference until no further improvement



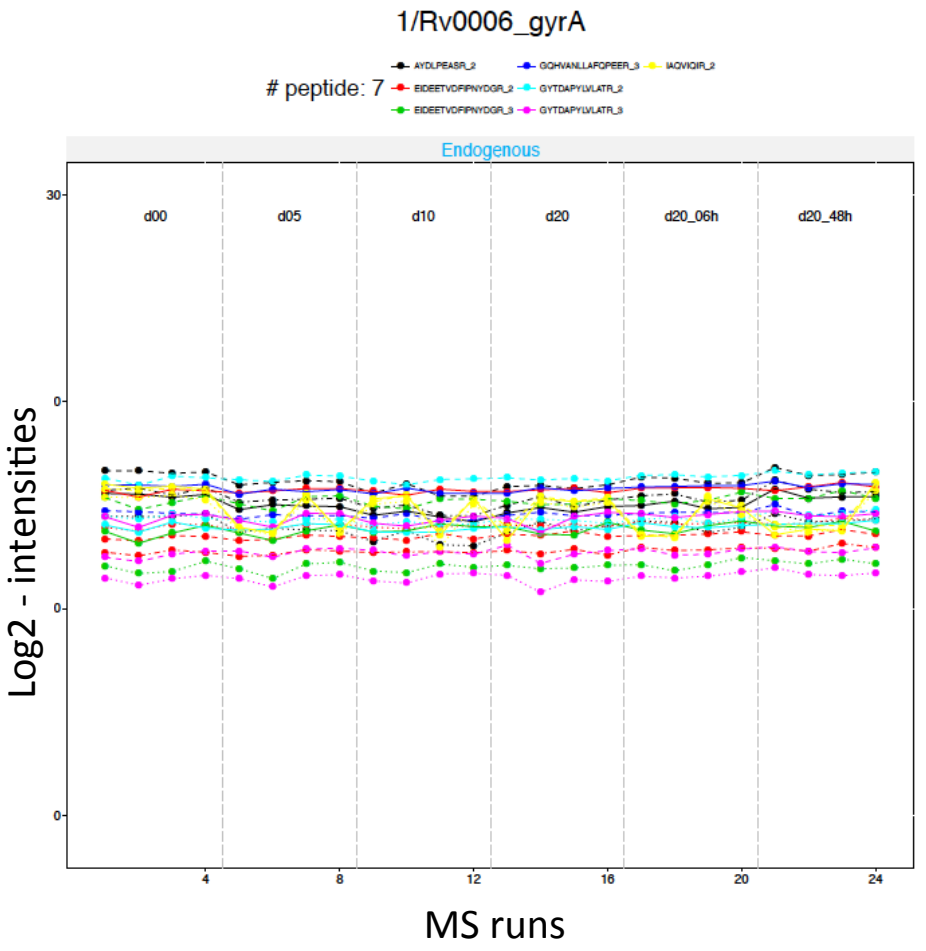
Feature selection removes the features with interference

Example DIA data

All features



After feature selection



Impute censored missing values

Missing values

- Random : technical variability
- Censored : below limit of detection

➔ Impute censored missing values by accelerated failure model

- ✓ Only endogenous intensities
- ✓ Cut off is minimum value for each feature

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	. . .	Run_{JK-2}	Run_{JK-1}	Run_{JK}	. . .	$Run_{(l-1)JK+1}$	$Run_{(l-1)JK+2}$	$Run_{(l-1)JK+3}$	$Run_{(l-1)JK+4}$	$Run_{(l-1)JK+5}$	$Run_{(l-1)JK+6}$	. . .	Run_{lJK-2}	Run_{lJK-1}	Run_{lJK}
$Feature_1$	y	y	y	y	y	y	. . .	y	y	y	. . .	y	y	y	y	y	y	. . .	y	y	y
$Feature_2$	y	y	y	y	y		. . .	y	y	y	. . .	y	y	y	y	y	y	. . .	y	y	y
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
$Feature_L$	y		y			y	. . .		y	y	. . .		y	y	y	y	y	. . .	y		y

↓

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	. . .	Run_{JK-2}	Run_{JK-1}	Run_{JK}	. . .	$Run_{(l-1)JK+1}$	$Run_{(l-1)JK+2}$	$Run_{(l-1)JK+3}$	$Run_{(l-1)JK+4}$	$Run_{(l-1)JK+5}$	$Run_{(l-1)JK+6}$	. . .	Run_{lJK-2}	Run_{lJK-1}	Run_{lJK}
$Feature_1$	y	y	y	y	y	y	. . .	y	y	y	. . .	y	y	y	y	y	y	. . .	y	y	y
$Feature_2$	y	y	y	y	y	y_{imp}	. . .	y	y	y	. . .	y	y	y	y	y	y	. . .	y	y	y
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
$Feature_L$	y	y_{imp}	y	y_{imp}	y_{imp}	y	. . .	y_{imp}	y	y	. . .	y_{imp}	y	y	y	y	y	. . .	y	y_{imp}	y

Run-level subplot summarization

- Estimate relative abundance for each protein per run
- Subplot: only consider feature + run
- Tukey's median polish: Robust parameter estimation method with median across rows and columns

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	. . .	Run_{JK-2}	Run_{JK-1}	Run_{JK}	. . .	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	. . .	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
$Feature_1$	y	y	y	y	y	y	. . .	y	y	y	. . .	y	y	y	y	y	y	. . .	y	y	y
$Feature_2$	y	y	y	y	y	y_{imp}	. . .	y	y	y	. . .	y	y	y	y	y	y	. . .	y	y	y
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
$Feature_L$	y	y_{imp}	y	y_{imp}	y_{imp}	y	. . .	y_{imp}	y	y	. . .	y_{imp}	y	y	y	y	y	. . .	y	y_{imp}	y

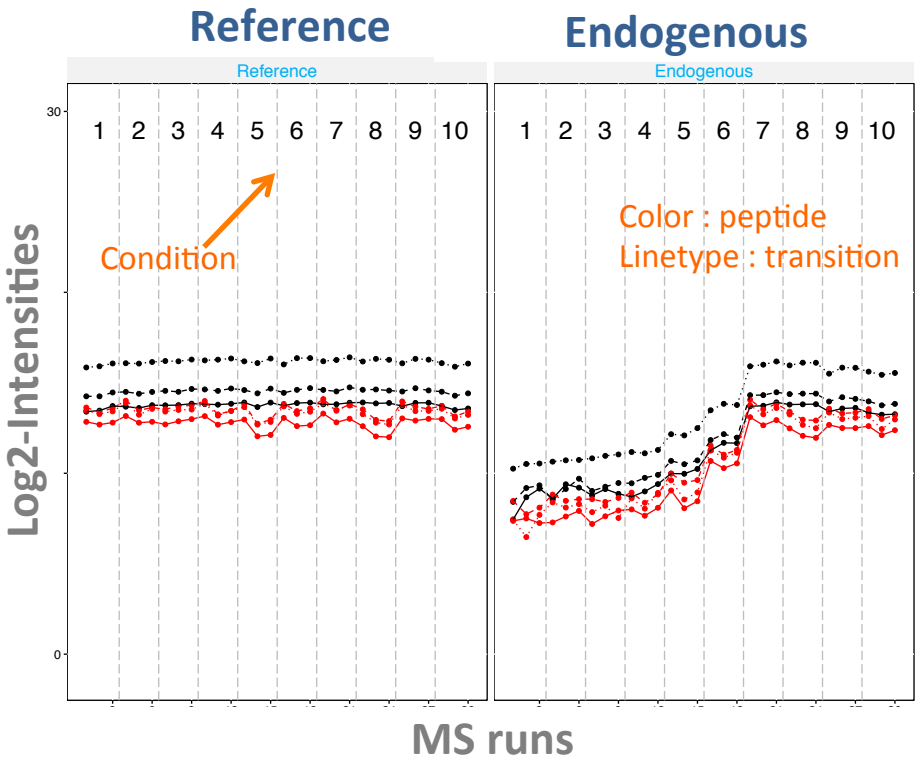
$$y_{ijkl} = \mu + Run_{ijk} + Feature_l + \epsilon_{ijkl} \text{ , where}$$

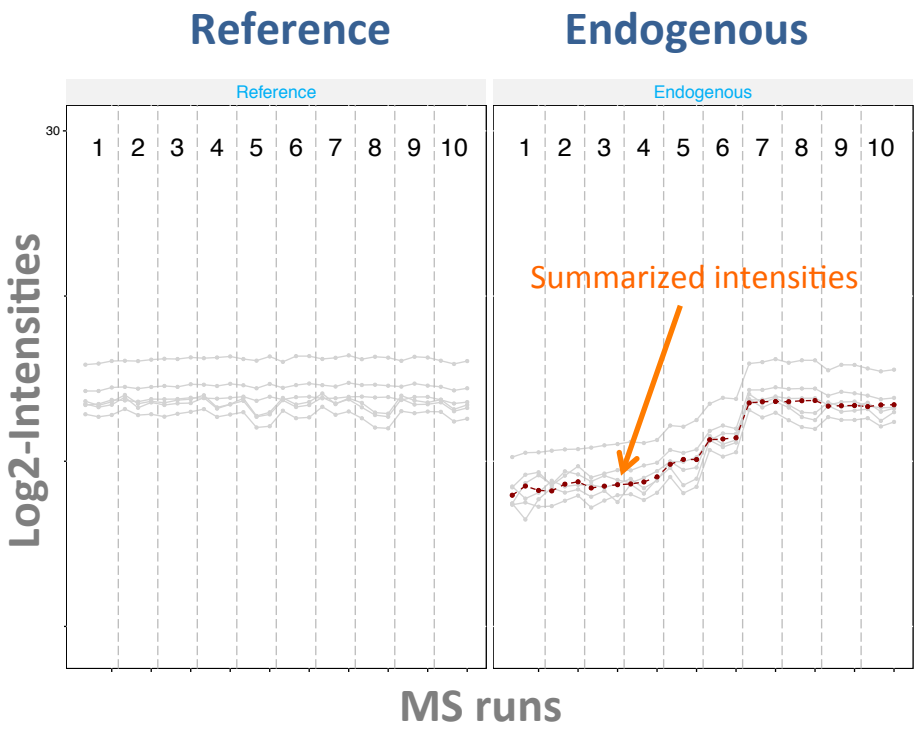
$$median_{ijk}(Run_{ijk}) = 0,\ median_l(Feature_l) = 0,\ and\ median_{ijk}(\epsilon_{ijkl}) = median_l(\epsilon_{ijkl}) = 0$$

	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	. . .	Run_{JK-2}	Run_{JK-1}	Run_{JK}	. . .	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	. . .	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
Summarized	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	. . .	$\hat{y}$	$\hat{y}$	$\hat{y}$	. . .	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	. . .	$\hat{y}$	$\hat{y}$	$\hat{y}$

Profile Plot

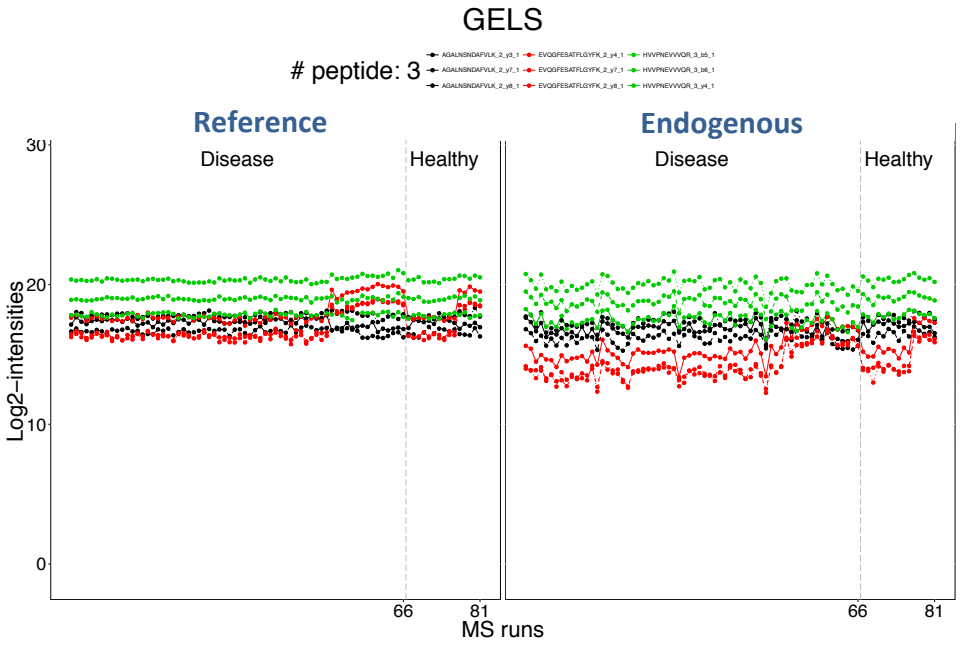
- Visualize individual observations
- Show the potential source of variation, such as Run, Transition, Condition
- Check missing values

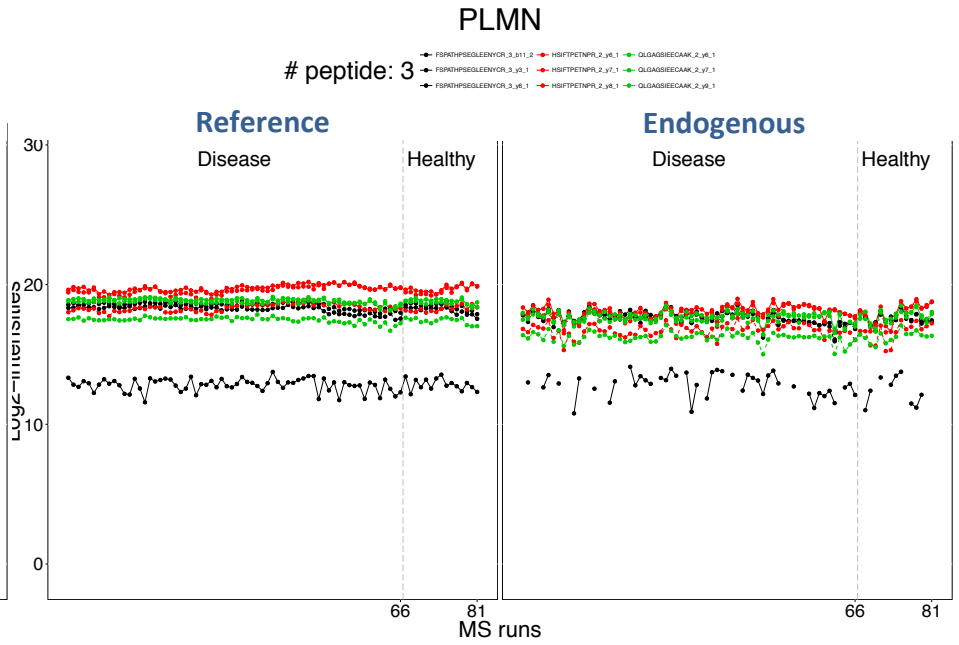




Good quality Profile plot. It shows the source of variation (Run, Condition, Transition)

Profile Plot





Detect the problematical Run or Transition

Show the missing values (Disconnection)

2. Group Comparison : Test for differential abundance



- Hypothesis : Is there a difference in abundance between condition1 and condition2?

  H_0 : log fold change = 0 vs. H_a : log fold change ≠ 0

- Comparisons between conditions are estimated by linear mixed effect models

Whole plot

	$Condition_1$						...	$Subject_J$			...	$Condition_I$						...	$Subject_U$		
	$Subject_1$			$Subject_2$			...				...	$Subject_{(I-1)J+1}$			$Subject_{(I-1)J+2}$			...			
	Run_1	Run_2	Run_3	Run_4	Run_5	Run_6	...	Run_{JK-2}	Run_{JK-1}	Run_{JK}	...	$Run_{(I-1)JK+1}$	$Run_{(I-1)JK+2}$	$Run_{(I-1)JK+3}$	$Run_{(I-1)JK+4}$	$Run_{(I-1)JK+5}$	$Run_{(I-1)JK+6}$	...	Run_{IJK-2}	Run_{IJK-1}	Run_{IJK}
Summarized	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	$\hat{y}$	...	$\hat{y}$	$\hat{y}$	$\hat{y}$

Group comparison

log(peak intensity) = expected reference abundance + condition + biological replicate + random meas. error

$$\hat{y}_{ijk} = \mu + Condition_i + Subject(Condition)_{j(i)} + \psi_{ijk}, \text{where}$$

$$\sum_i Condition_i = 0, \; Subject(Condition)_{j(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{Subject}), \; \psi_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\psi})$$

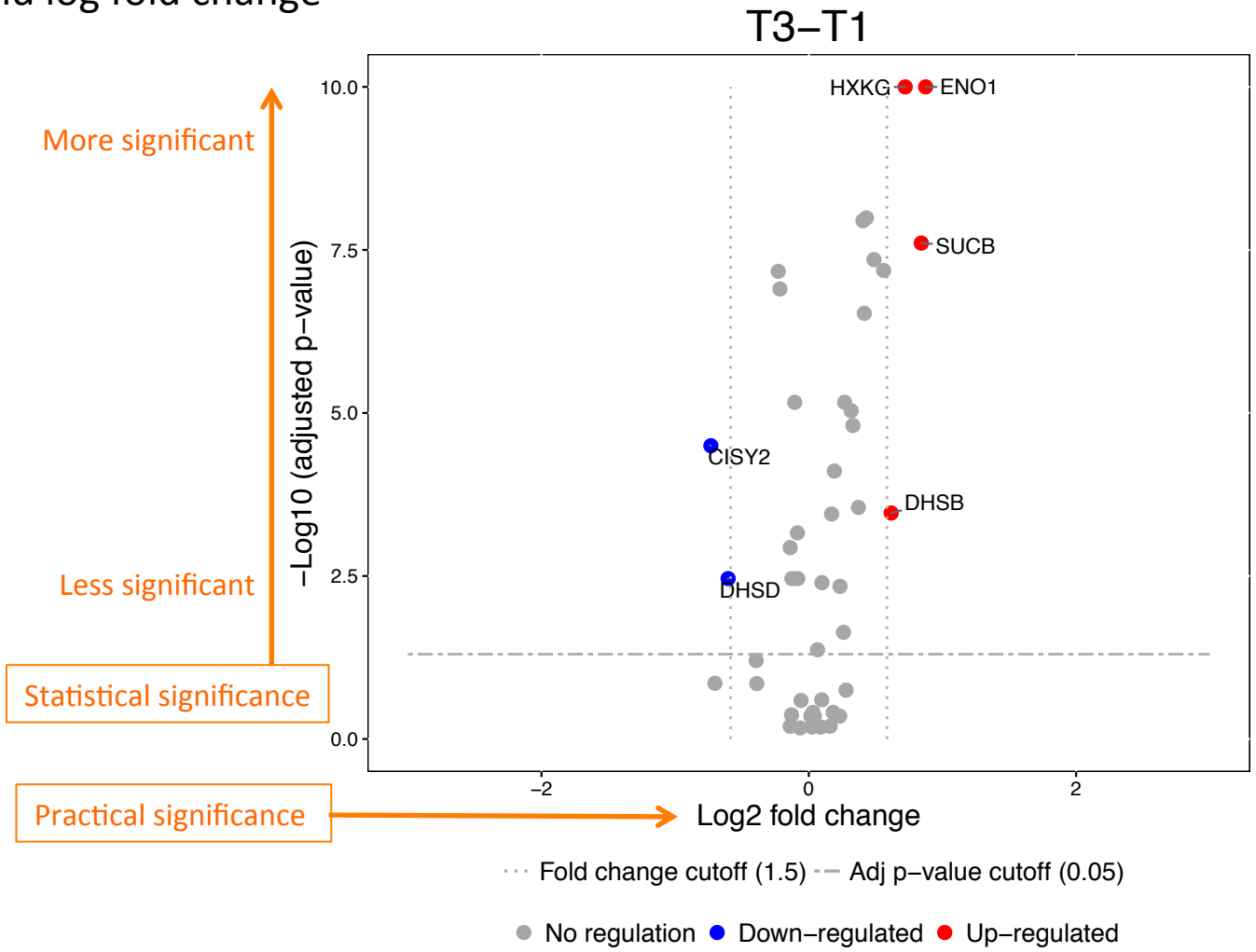
Results table for inference

- **Log2 fold change**: provides an estimate for the difference between the two conditions
- **Adjusted p-value**: probability that your conclusion is false (adjusted = corrected for multiple testing)

Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue
NP_001007697	Disease-Healthy	-3.232825973	0.388549983	-8.3202319	12	2.51E-06	3.01E-05
NP_001008724	Disease-Healthy	-0.356257059	0.402904721	-0.8842216	12	0.39394796	0.54027149
NP_001010968	Disease-Healthy	-0.308483858	0.366600666	-0.8414711	12	0.4165381	0.55538414
NP_001011908	Disease-Healthy	-1.436652196	0.262203616	-5.4791471	12	0.0001409	0.00135261
NP_001012027	Disease-Healthy	-0.093330917	0.388382375	-0.2403068	12	0.81414864	0.90881709
NP_001013967	Disease-Healthy	-1.015265095	0.3575297	-2.8396665	12	0.01490594	0.03736399
NP_001033064	Disease-Healthy	-1.522690232	0.432885764	-3.5175336	12	0.00424265	0.01392513
NP_001101333	Disease-Healthy	-1.162324993	0.331736191	-3.503763	12	0.0043516	0.01392513

Volcano plot

- Per comparison
- All proteins
- Adjusted p-value and log fold change

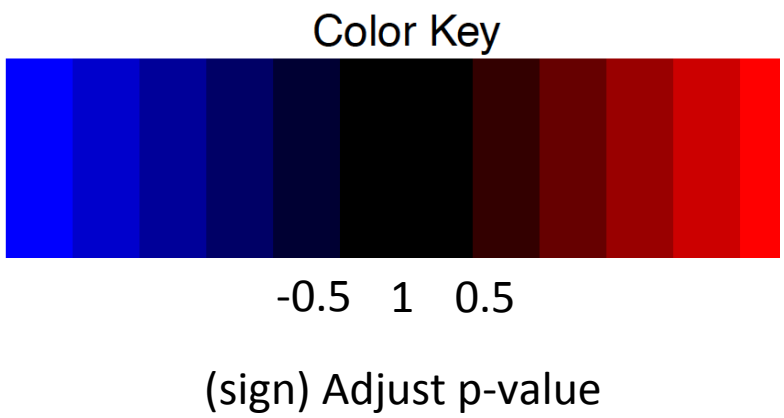


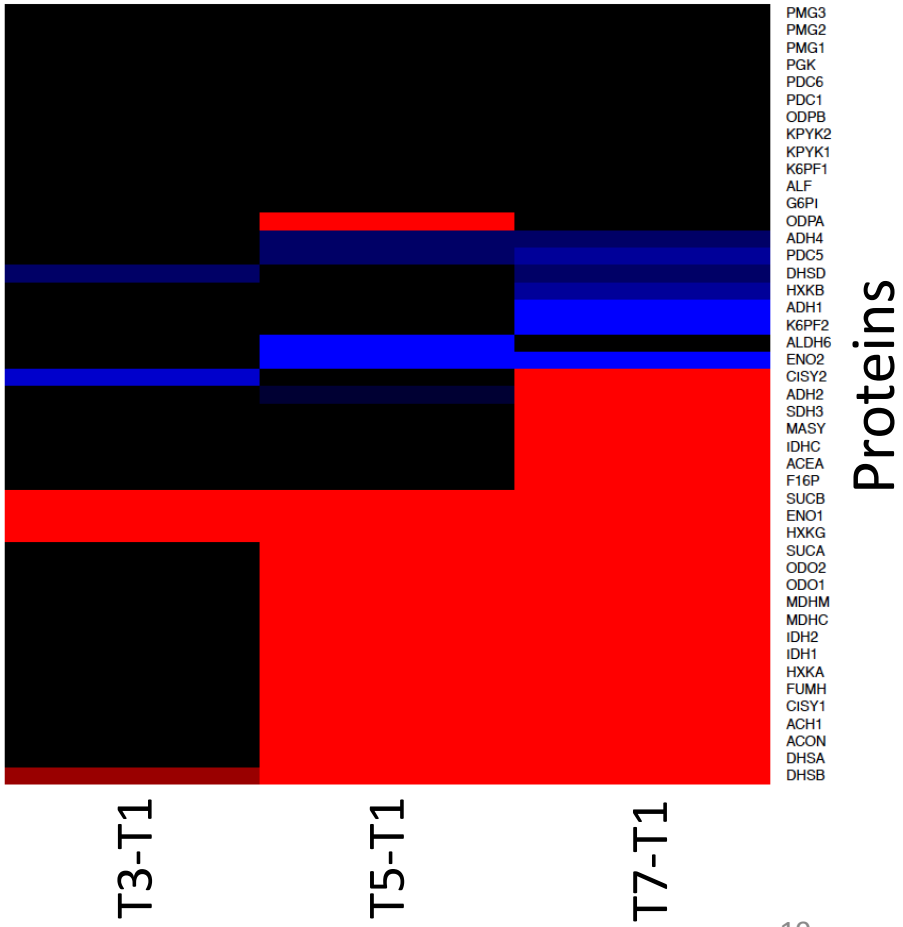
Heatmap

- With all comparisons
- All proteins
- Adjusted p-value and cut-off log fold change

Based on adjusted p-values,

- negative p-value : In cases of a negative fold change below a specified cutoff
- positive p-value : In case of positive fold change above a a specified cutoff

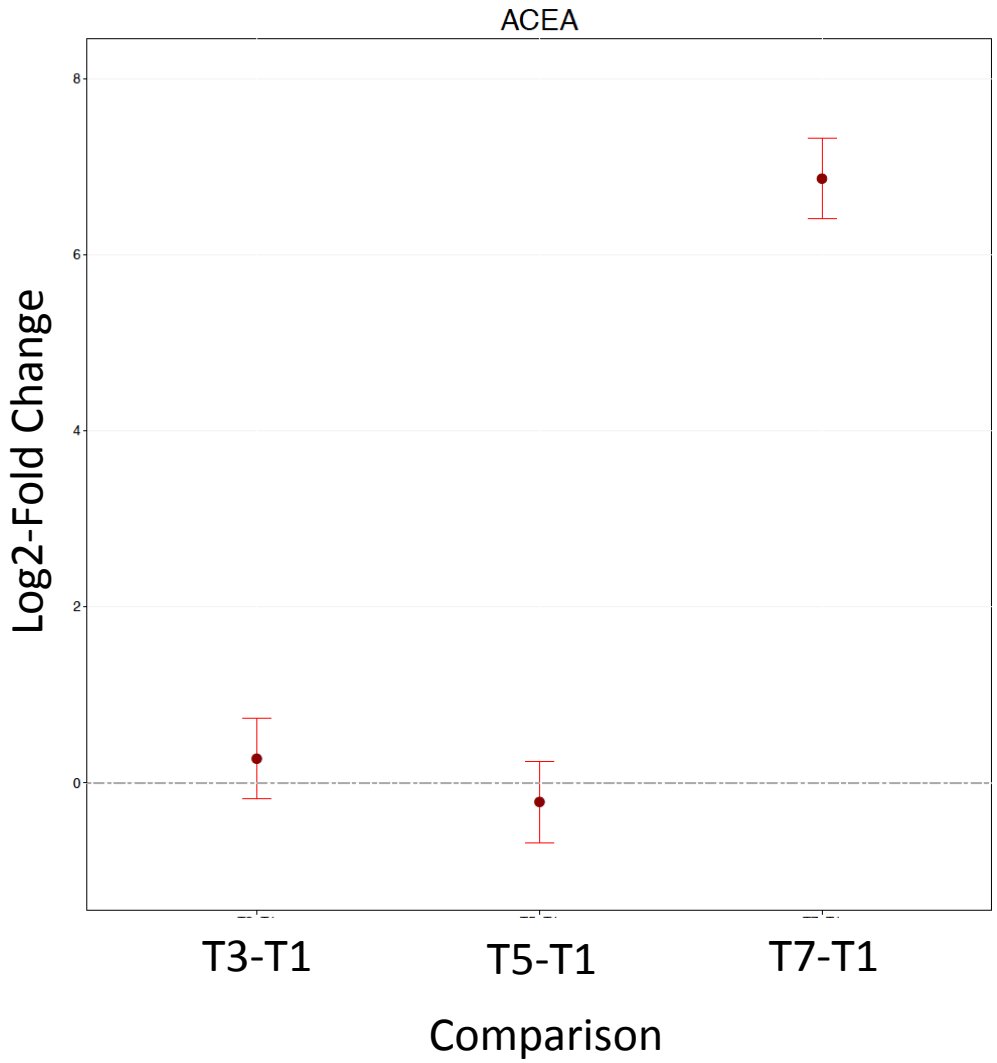




Comparison plot

Represents log fold changes for multiple comparisons per protein

- With all comparisons
- Per protein
- Log fold change and CI



3. Design sample size : Design of future experiment



Use information (variance estimation) from current study to

a) estimate the power of your analysis

b) calculate the required sample size for follow-up experiments

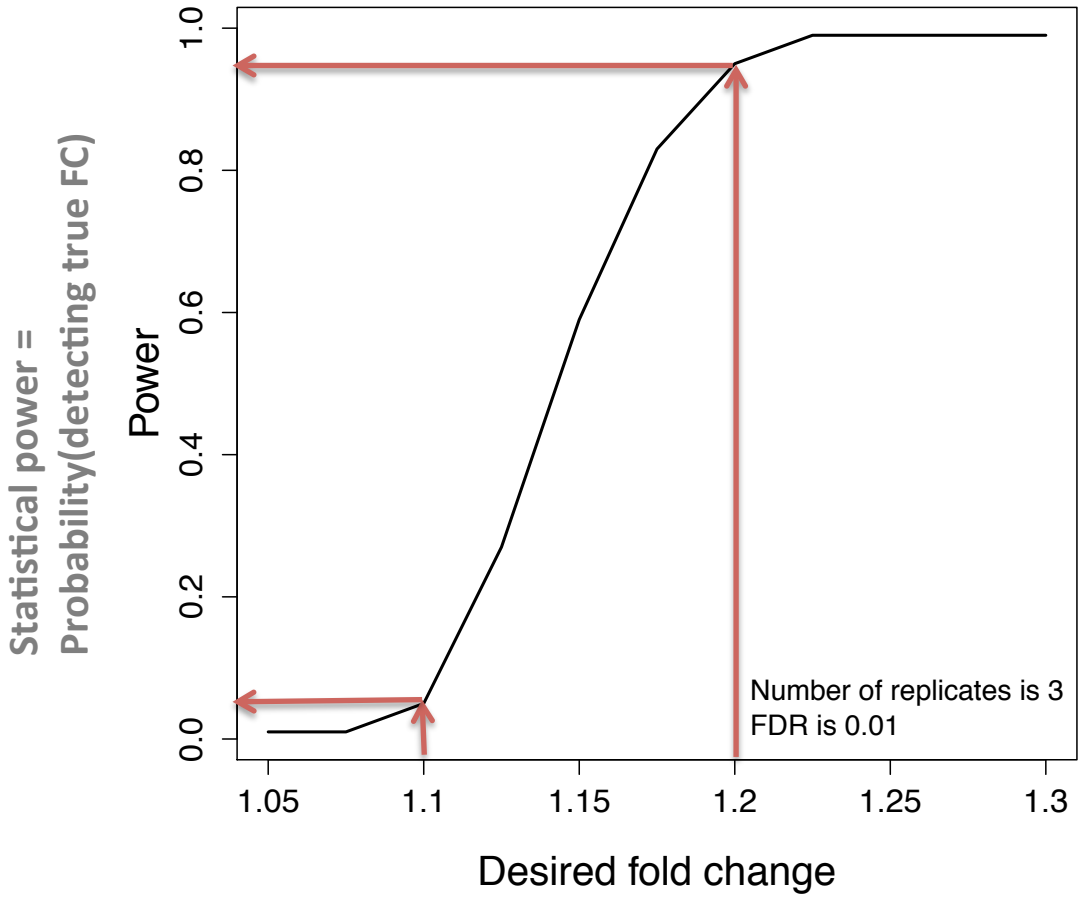
Parameters:

- **Power** (sensitivity): probability of detecting the effect (usually ≥ 0.80)
- **Sample size**: number of biological replicates
- **FDR**: probability of rejecting the null hypothesis even though it is true (usually ≤ 0.05)
- **Variance**: sample variance estimated from current experiment
- **Effect size**: minimum deviation from the null hypothesis that you hope to detect (log fold change)

desiredFC	numSample	numPep	numTran	FDR	power	CV
1.25	4	3	5	0.05	0.9	0.004
1.275	3	3	5	0.05	0.9	0.005
1.3	3	3	5	0.05	0.9	0.005
1.325	2	3	5	0.05	0.9	0.007
1.35	2	3	5	0.05	0.9	0.007
1.375	2	3	5	0.05	0.9	0.006
1.4	2	3	5	0.05	0.9	0.006
1.425	2	3	5	0.05	0.9	0.006

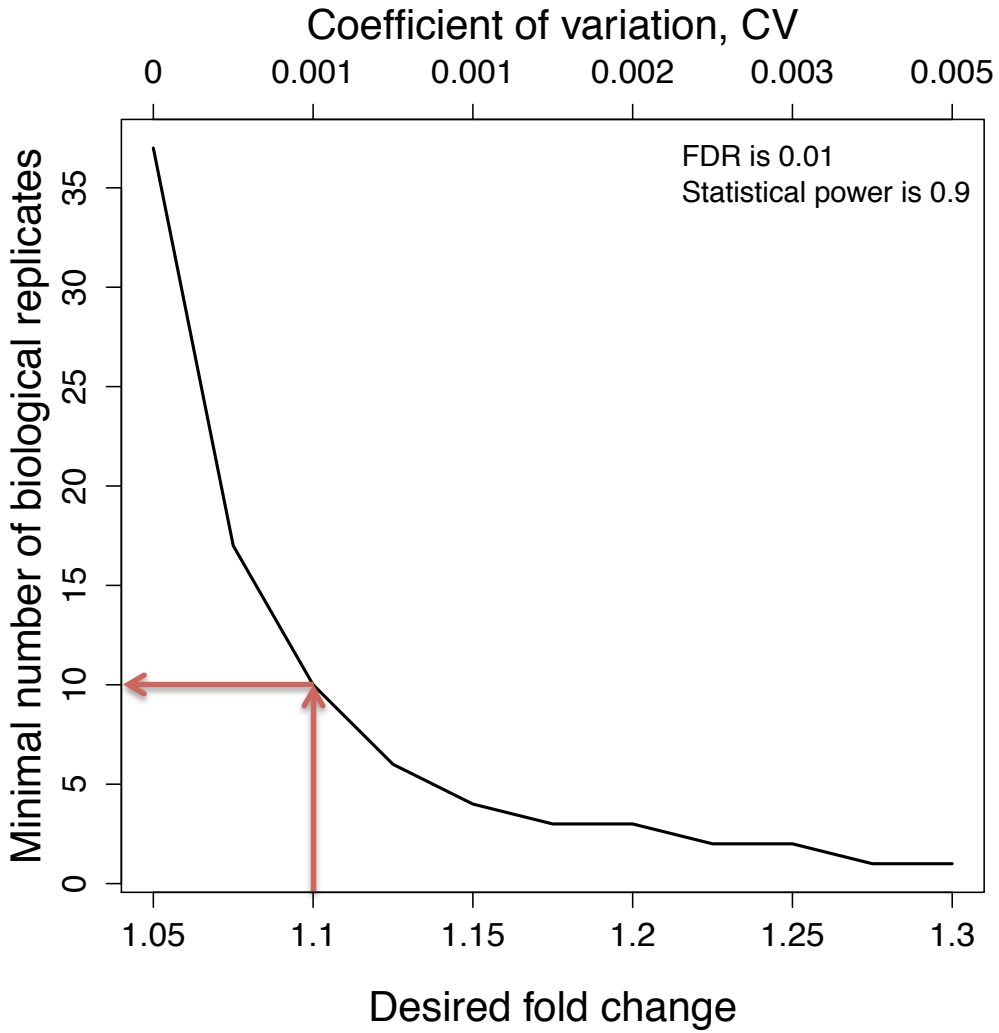
Statistical power estimation

Visualizes the power of an analysis at any desired fold change with a fixed number of biological replicates and FDR



Sample size calculation

Visualizes the minimal number of biological replicates required to observe a desired fold change at a fixed statistical power and FDR



4. Progress report : msstats.log

- Includes
  - R version, loaded software libraries, Options selected by the user, Data structure MSstats recognizes, Completion of intermediate analysis steps, Warning messages
- Help troubleshoot potential problems

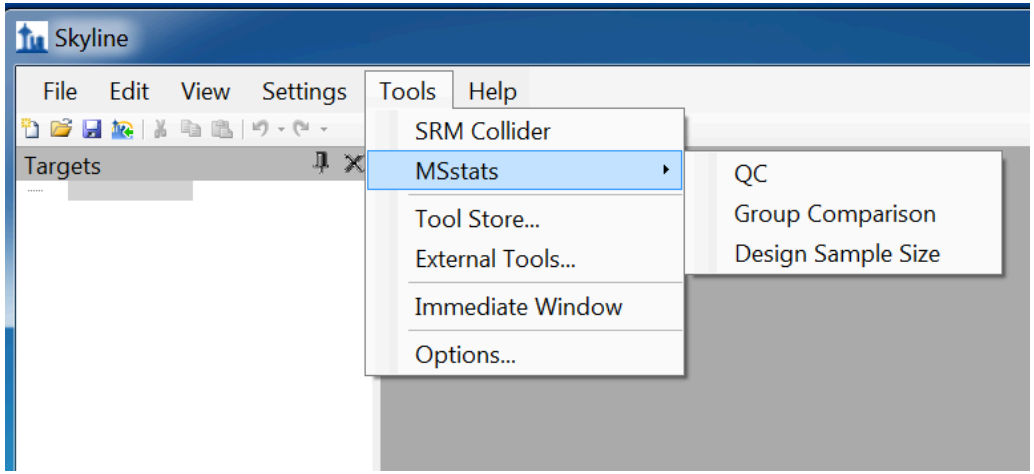
```
R.version.3.2.3..2015.12.10.
Platform: x86_64-apple-darwin13.4.0 (64-bit)
...
other attached packages:
[1] MSstats_3.3.4  reshape2_1.4.1 Rcpp_0.12.2    ggplot2_2.0.0
...
MSstats - dataProcess function

The required input : provided - okay
summaryMethod : TMP
cutoffCensored : minFeature
censoredInt : NA
New input format : made new columns for analysis - okay
...
MSstats - groupComparison function

labeled = TRUE
scopeOfBioReplication = expanded
Time course design of experiment - okay
Finished a comparison for protein  ACEA ( 1  of  45 )
Finished a comparison for protein  ACH1 ( 2  of  45 )
...
```

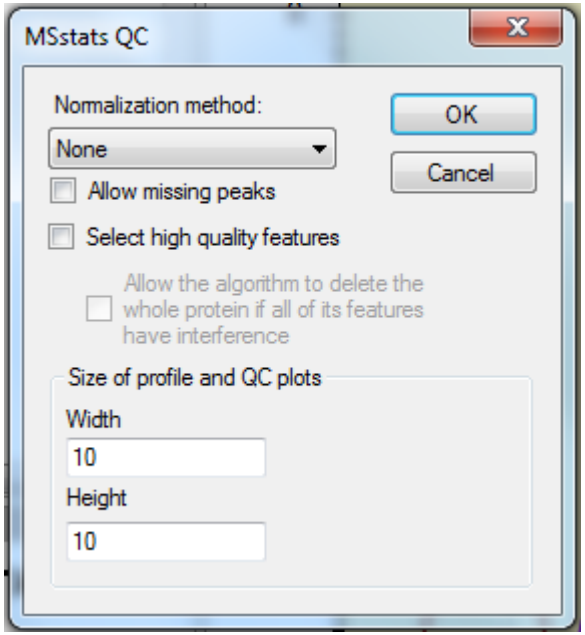
MSstats as an external tool in Skyline



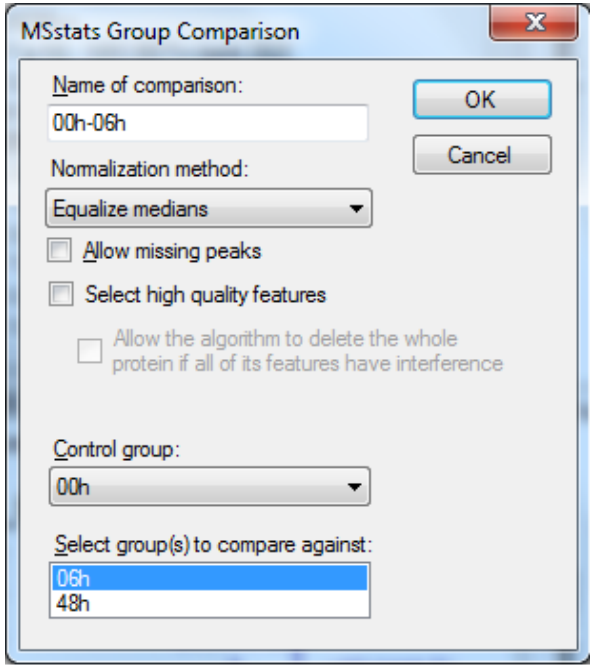


- For the beginner of R or other statistical tools, we can do statistical analysis with default options through Skyline easily.
- In R, more options and functions are available.

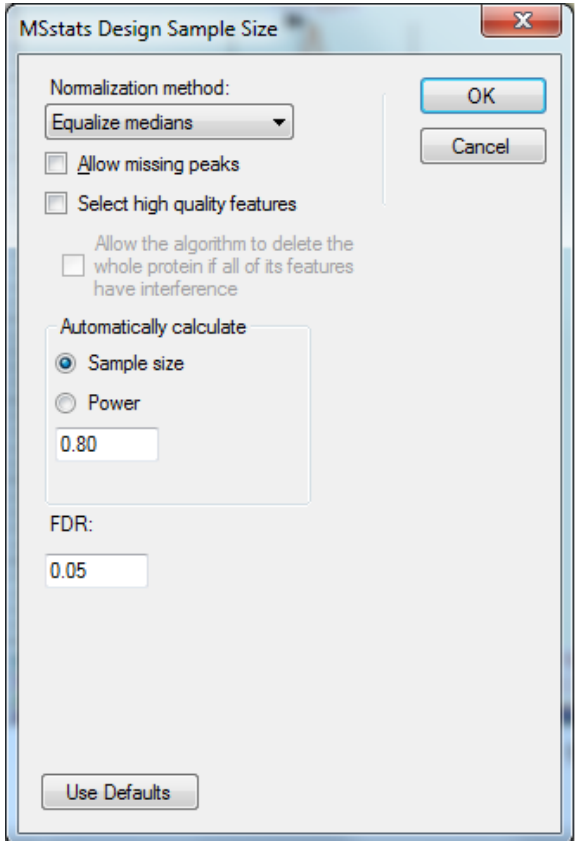
MSstats functions in Skyline



- Select normalization method
- High quality feature selection



- Select which groups to compare



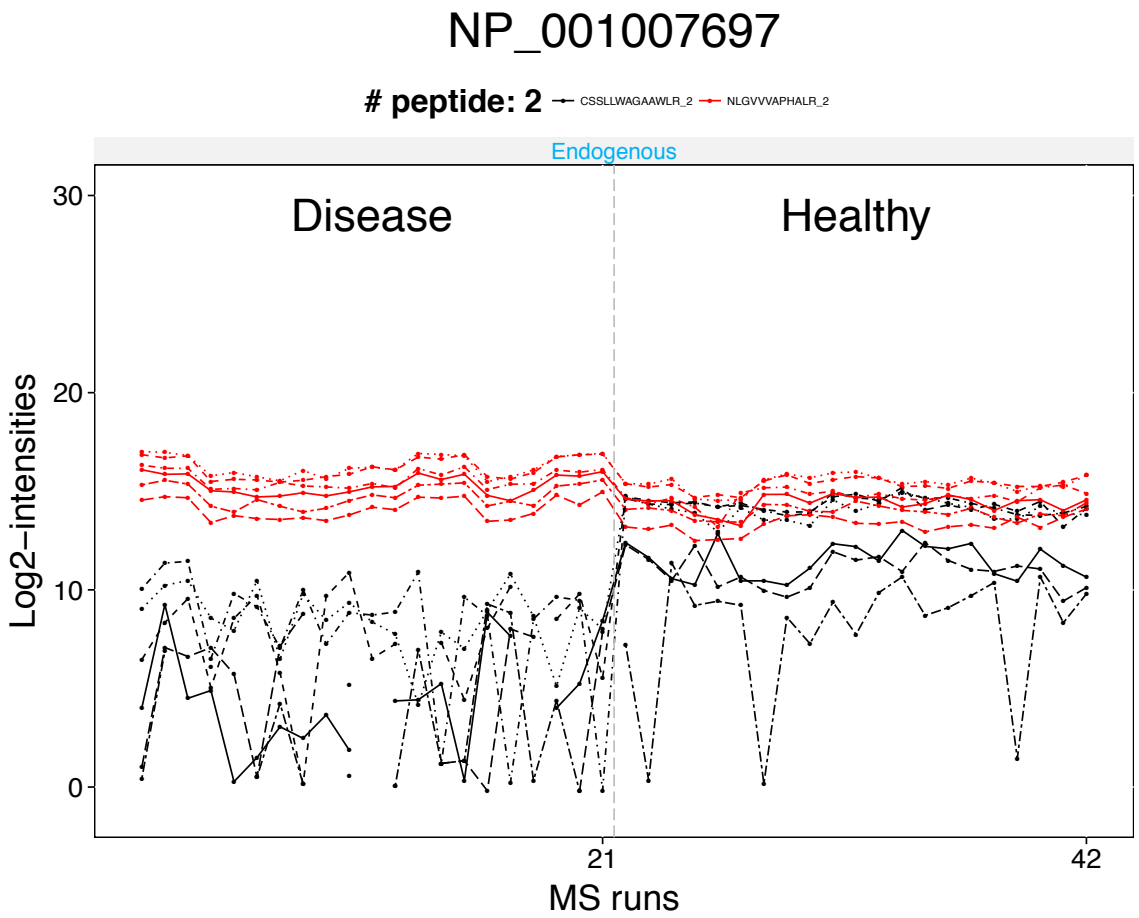
- Select sample size or power calculation
- Specify the desired FDR

Data : Rat-plasma for Risk of heart disease

- Label-free SRM experiment
- High salt (7) vs. Low salt (7)
- 3 Technical replicates
- Total 42 injections (Runs)
- 48 proteins
- Comparison : High Salt – Low Salt (Disease-Healthy)

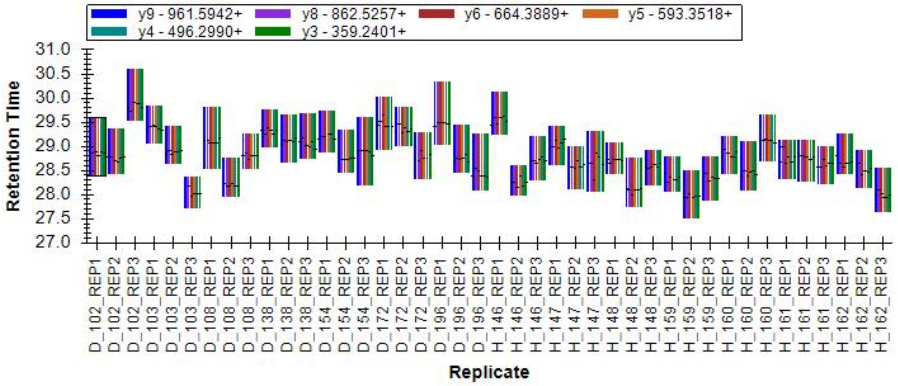
Each Protein	High salt (Disease)									Low salt (Healthy)								
	Sub1			…			Sub7			Sub8			…			Sub14		
	T1	T2	T3				T1	T2	T3	T1	T2	T3				T1	T2	T3
Pep*Tran1	X	X	X		…		X	X	X	X	X	X		…		X	X	X
Pep*Tran2	X	X	X		…		X	X	X	X	X	X		…		X	X	X
Pep*Tran3	~~X~~	X	X		…		X	X	X	X	X	X		…		X	X	X

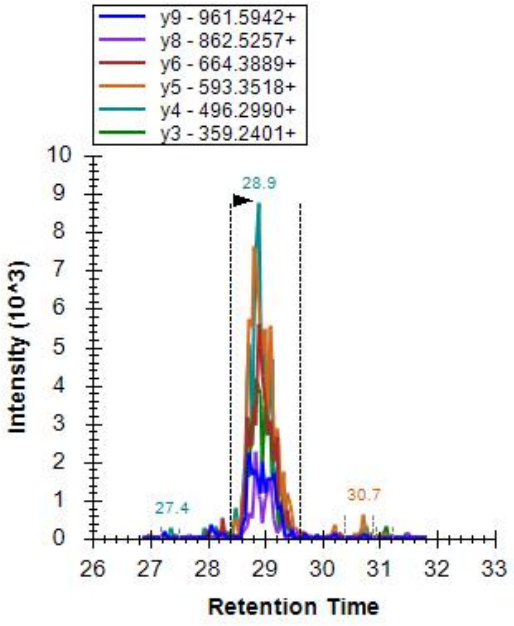
Examples of inconsistent (poor quality?) peptides

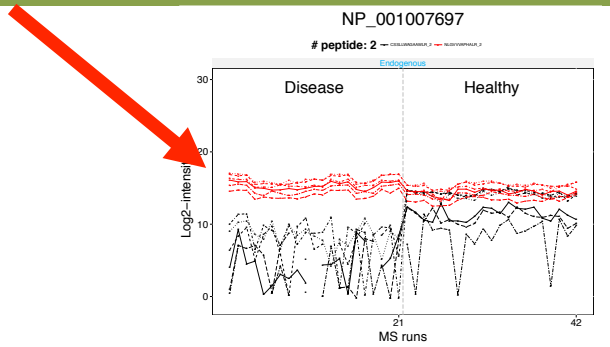


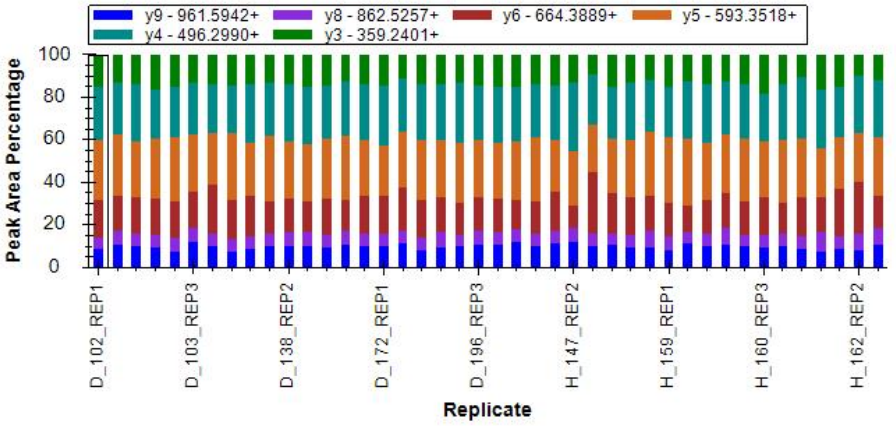
Profile plot show the problematic peptides or transitions. We need to check what happen in this peptide.

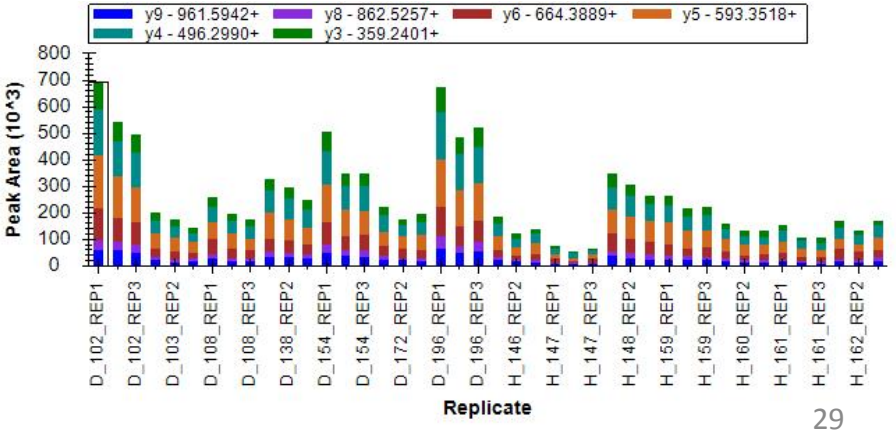
NLGVVVAPHALR



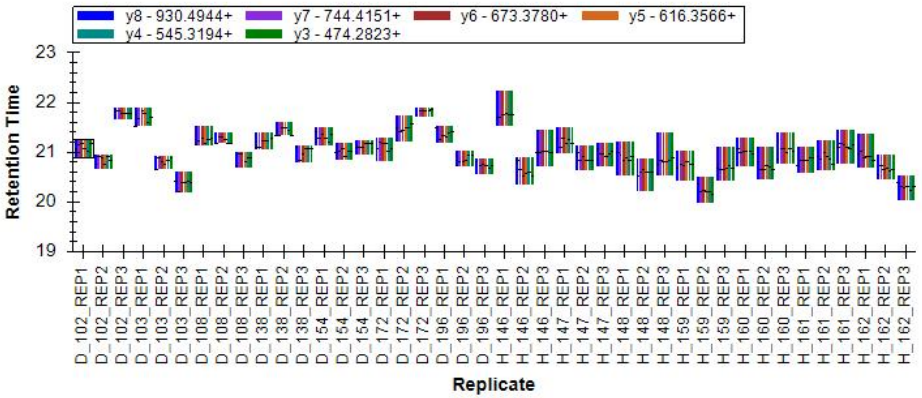


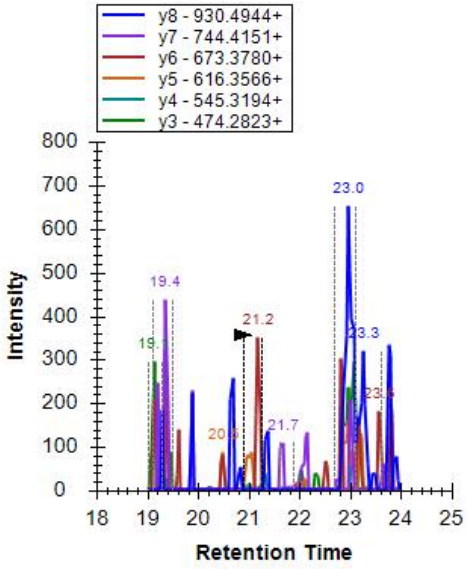


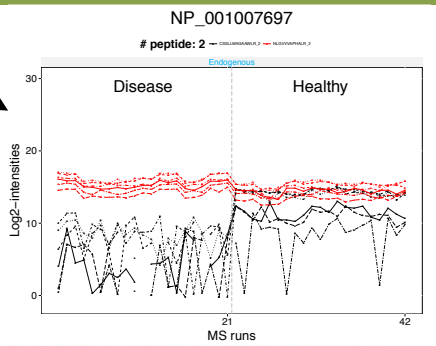


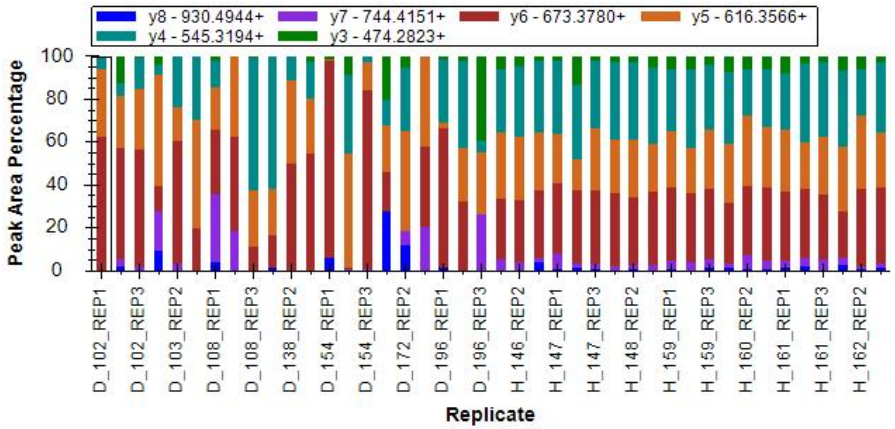


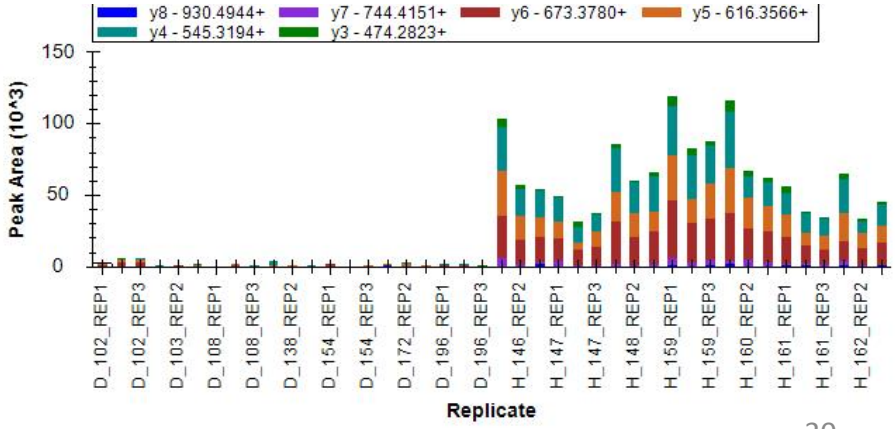
CSSLLWAGAAWLR



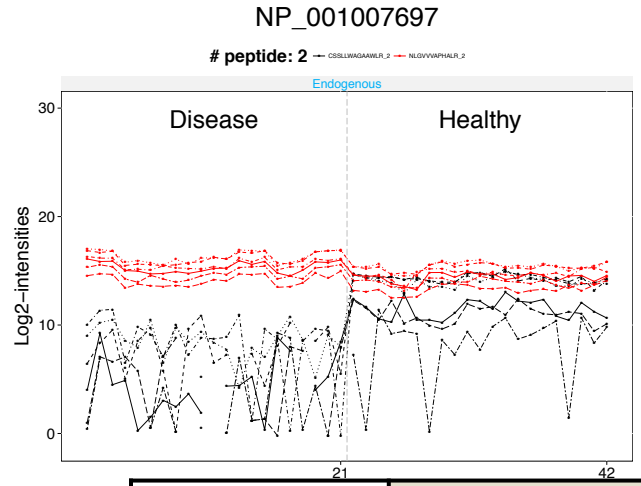


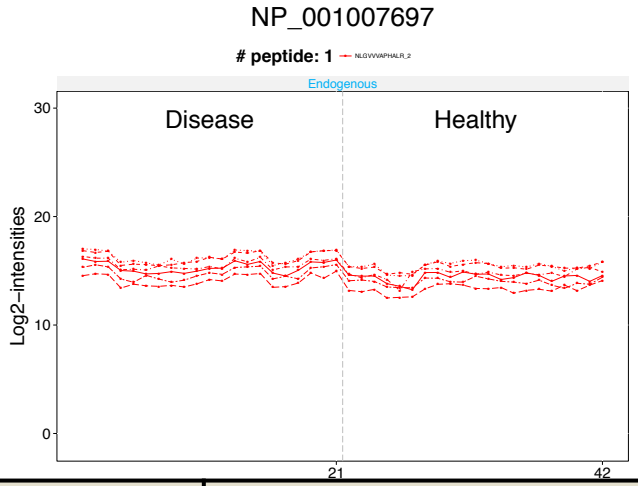


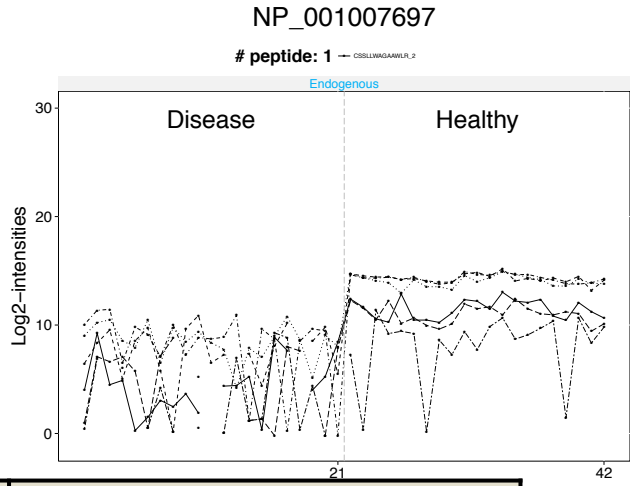




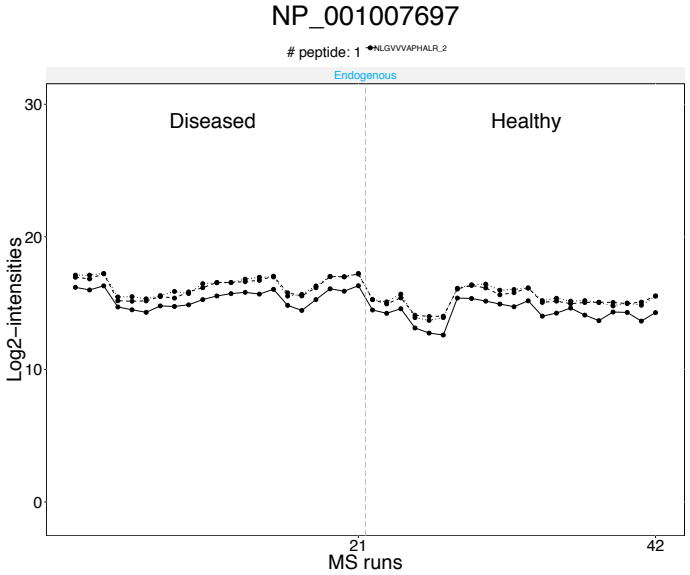
Log2 FC and variation are different between before and after removing peptides





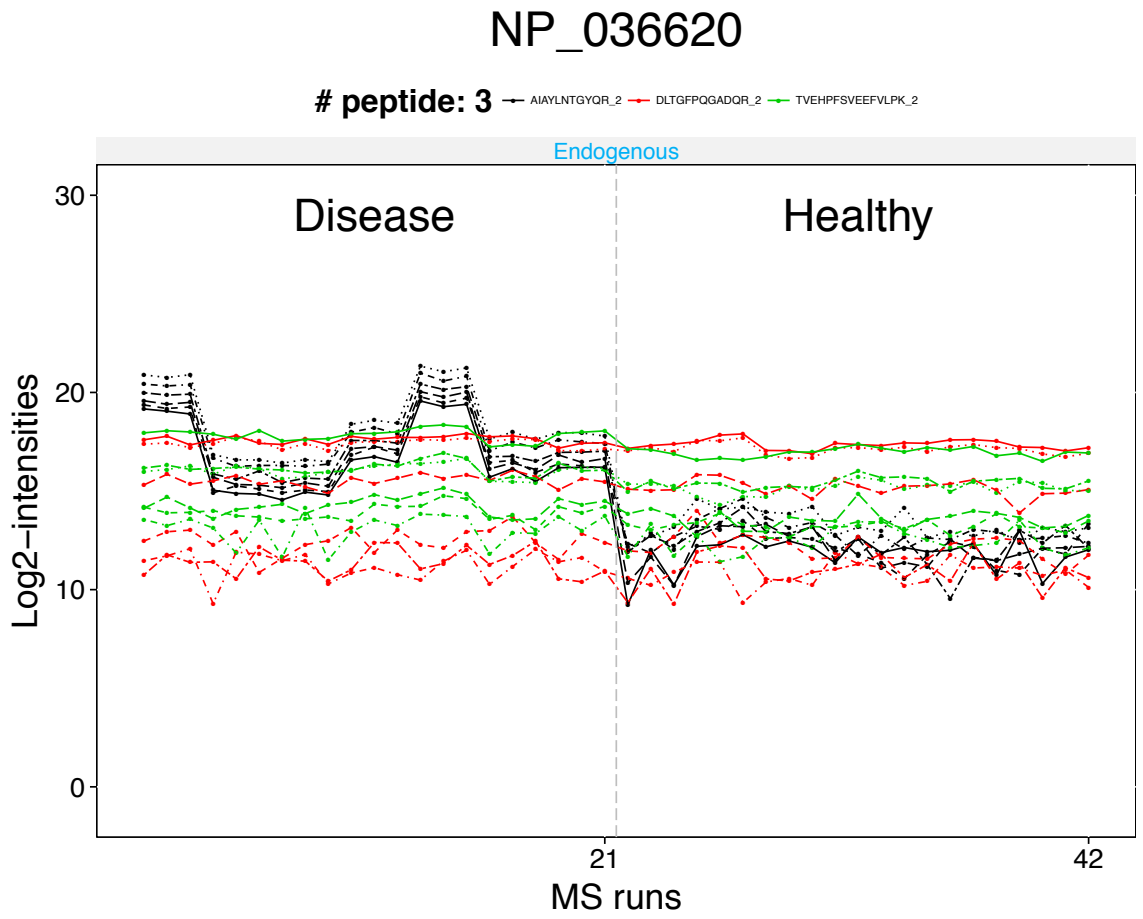


MS runs	All features			Only NLGV (red lines)			Only CSSL (black lines)		
	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value
Result	-2.6701	0.2214	<0.0001	0.8750	0.2399	0.0066	-6.2187	0.4152	<0.0001



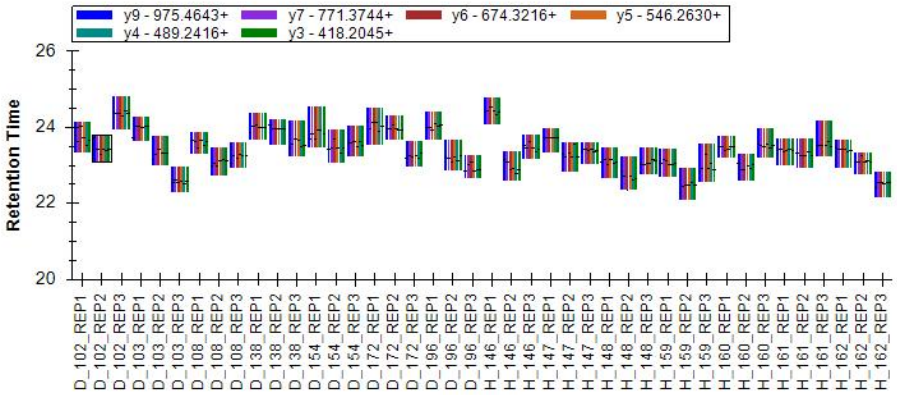
	Featuares without interference		
	log2FC	SE	Adj p-value
Result	1.094	0.3880	0.0087

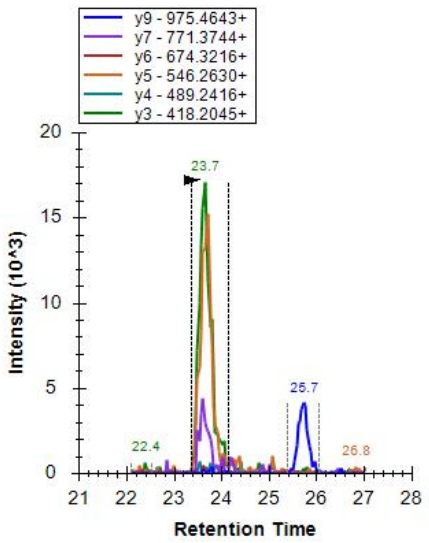
Examples of inconsistent peptides

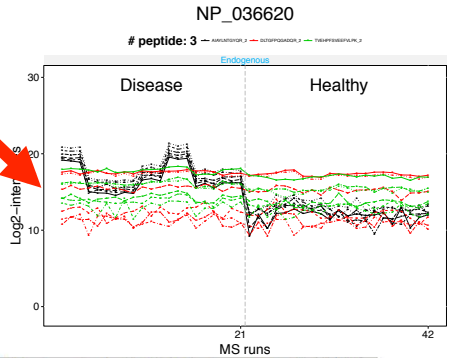


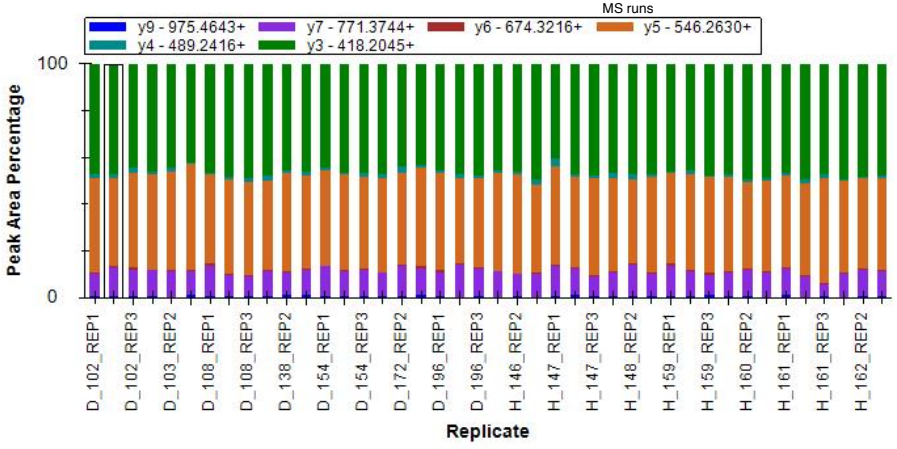
Profile plot show inconsistent pattern per peptides. We need to check that is there any measurement problem.

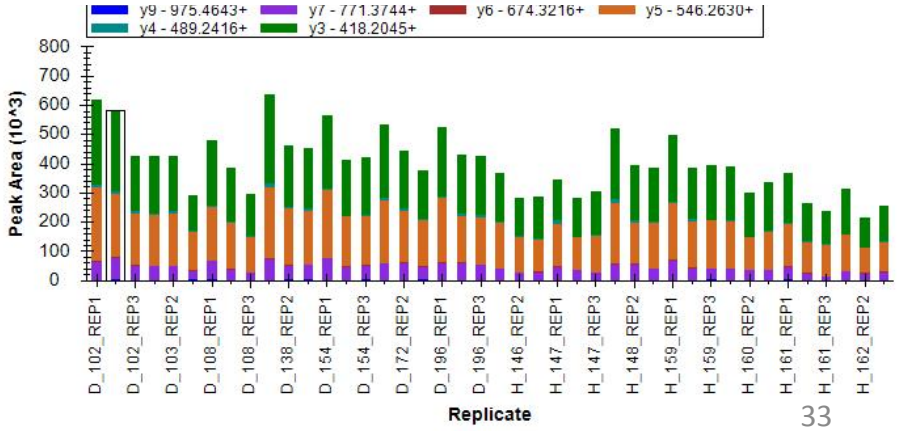
DLTGFPQGADQR



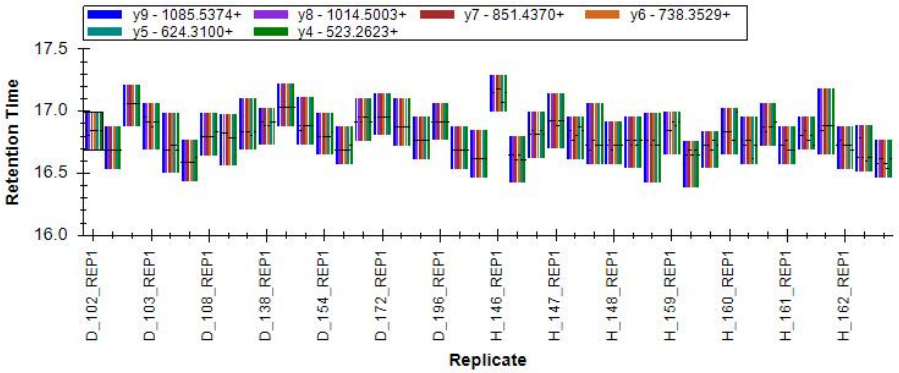


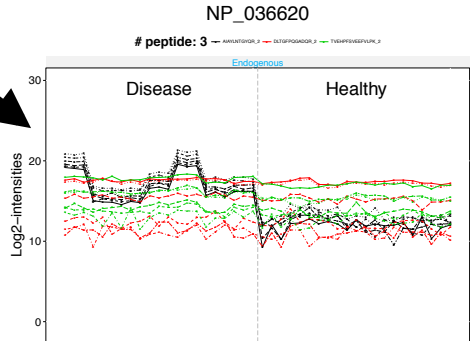


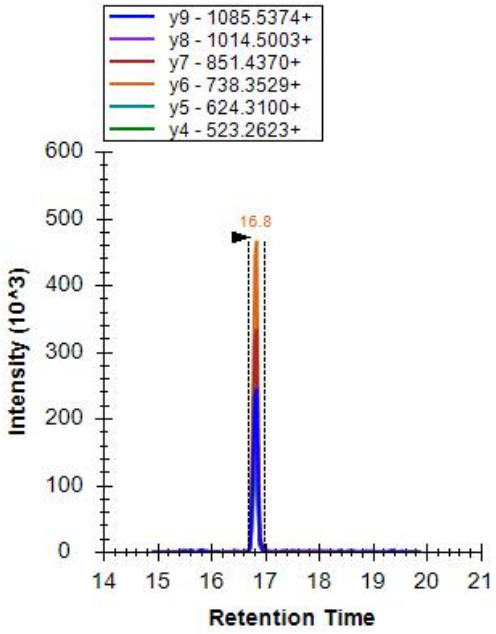


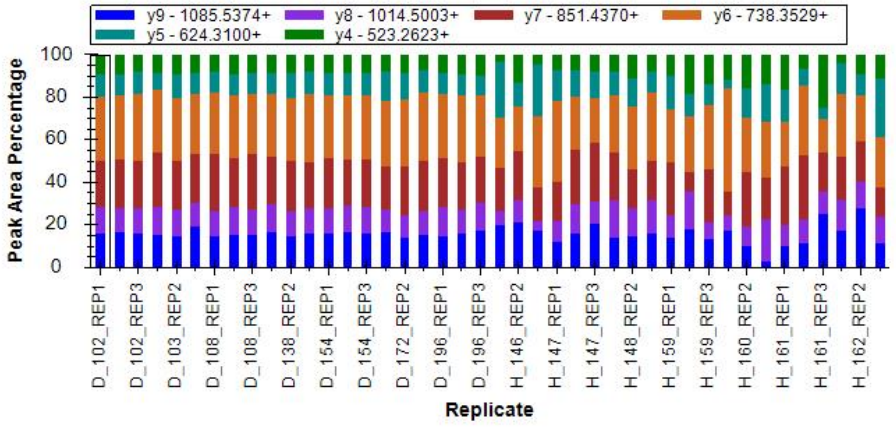


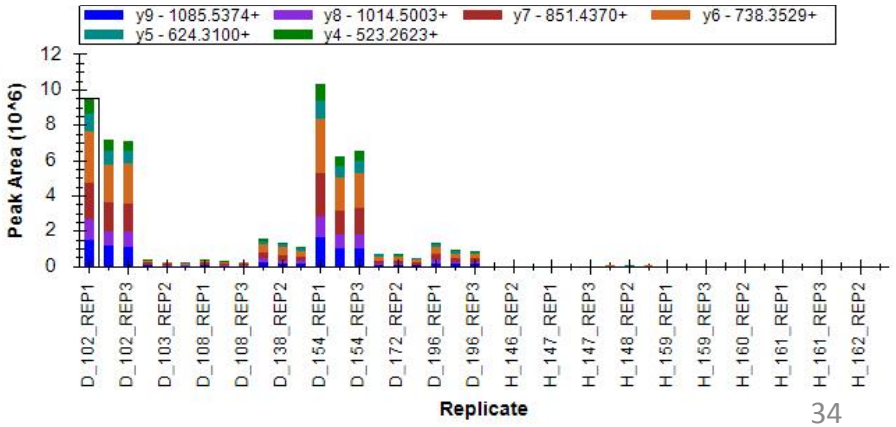
AIAYLNTGYQR



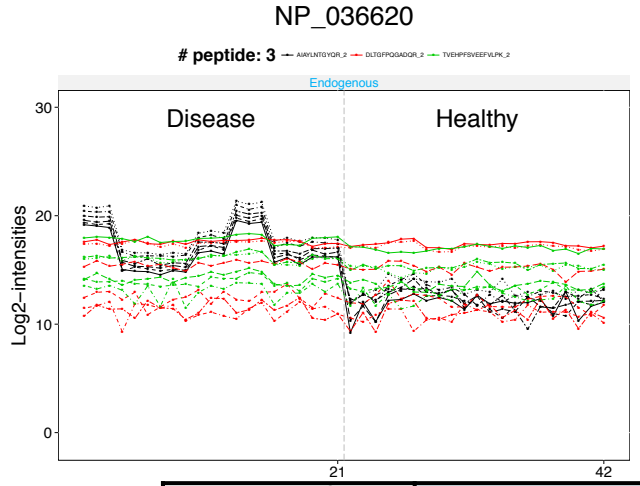


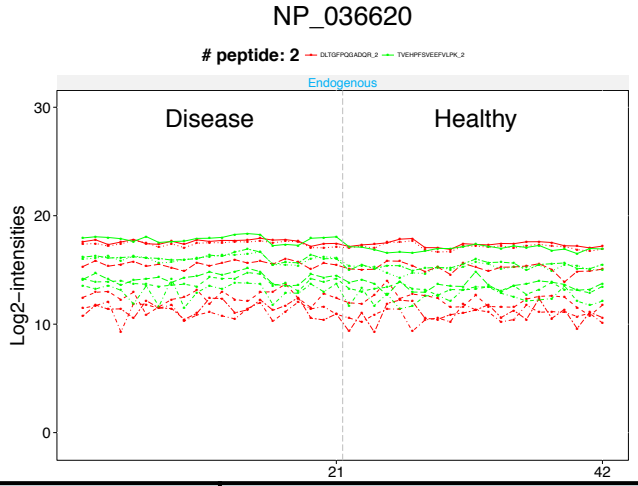


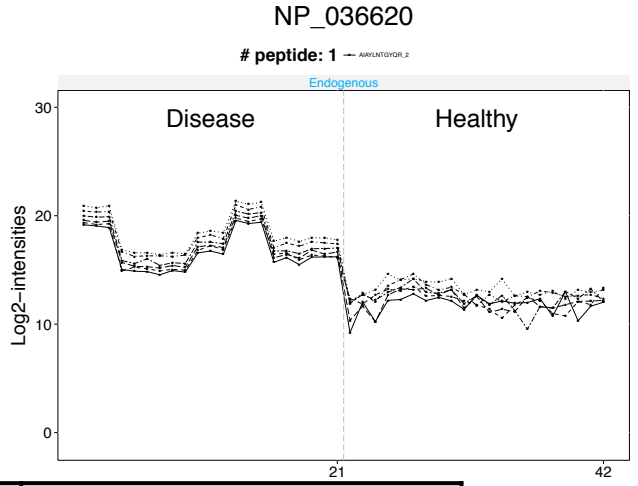




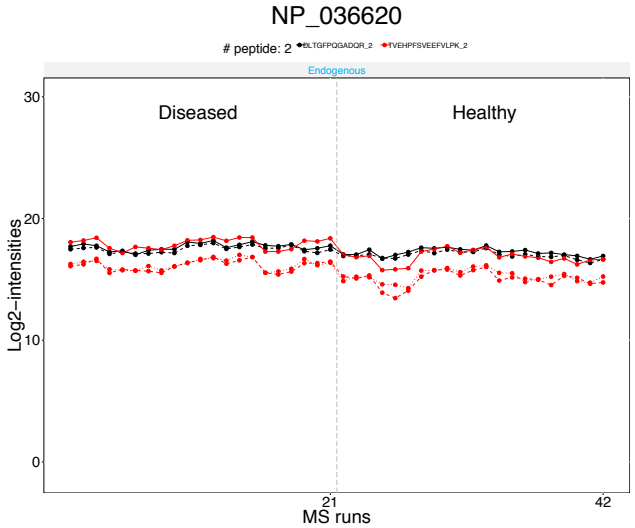
Log2 FC and variation are quite different depending on peptides.





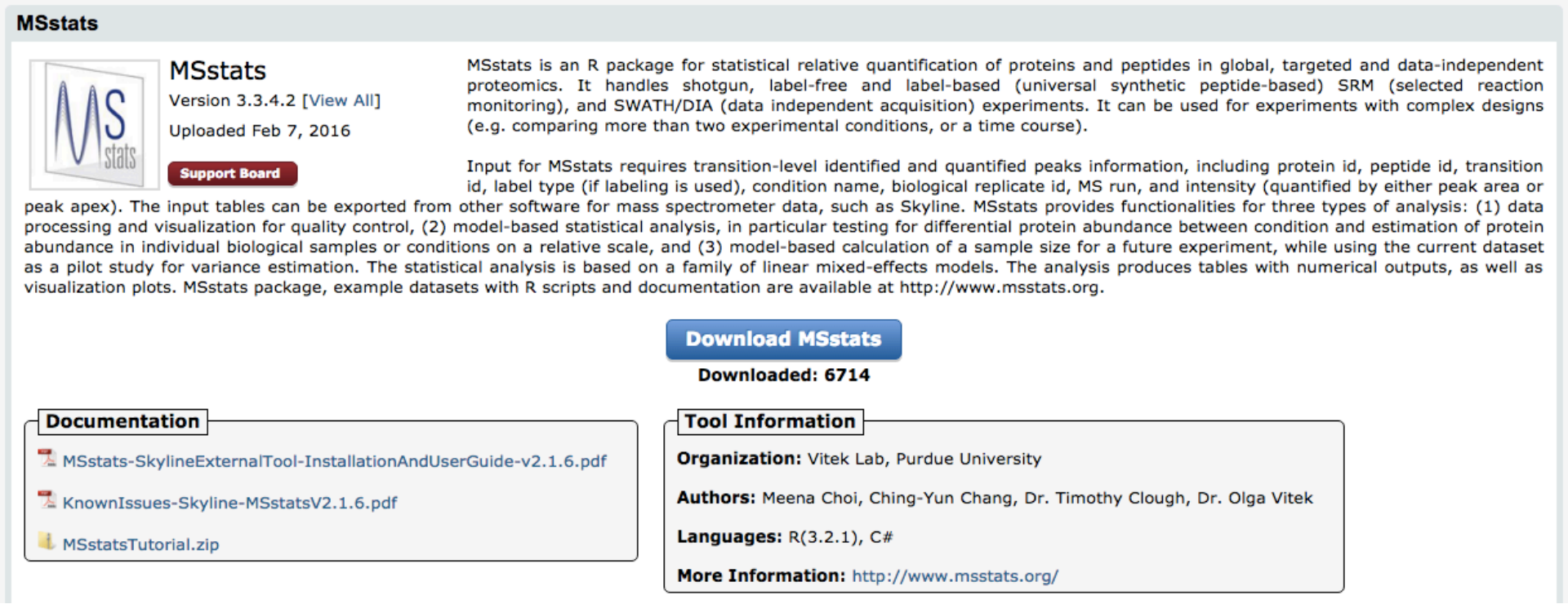


	All features			Only DLTG and TVEH			Only AIAY		
	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value
Results	2.0642	0.2966	<0.0001	0.6167	0.1137	0.0005	5.0812	0.7390	<0.0001



	Features without interference		
	log2FC	SE	Adj p-value
Results	0.8535	0.2334	0.001

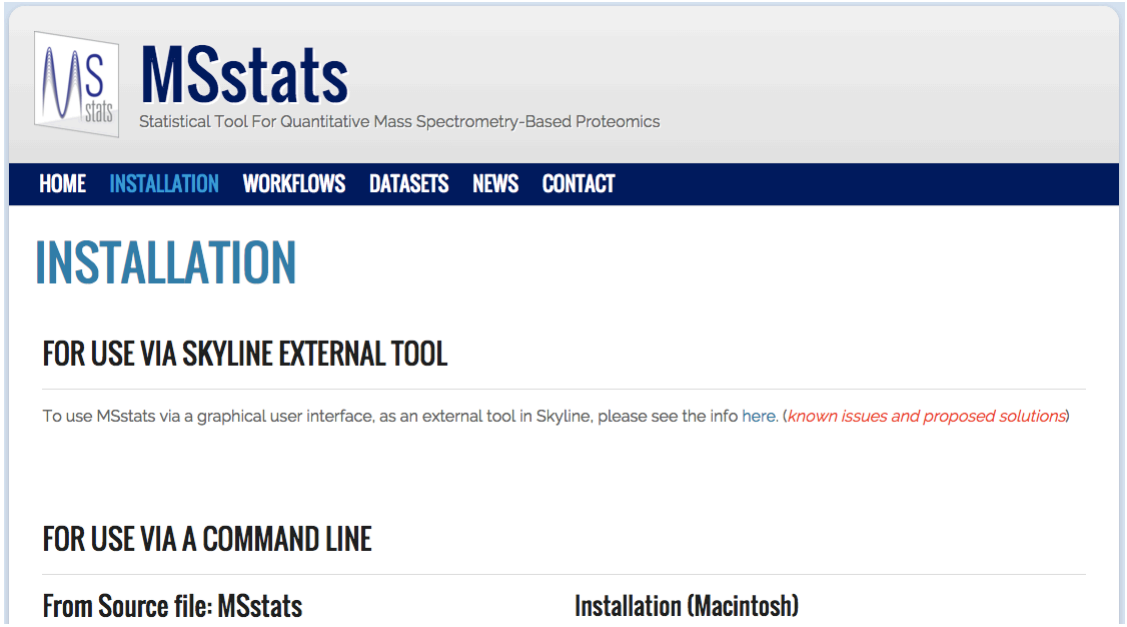
External tool in Skyline



- From MSstats external tool webpage or 'Tool store'
- Automatic installations for all related software and packages
- One-click analysis

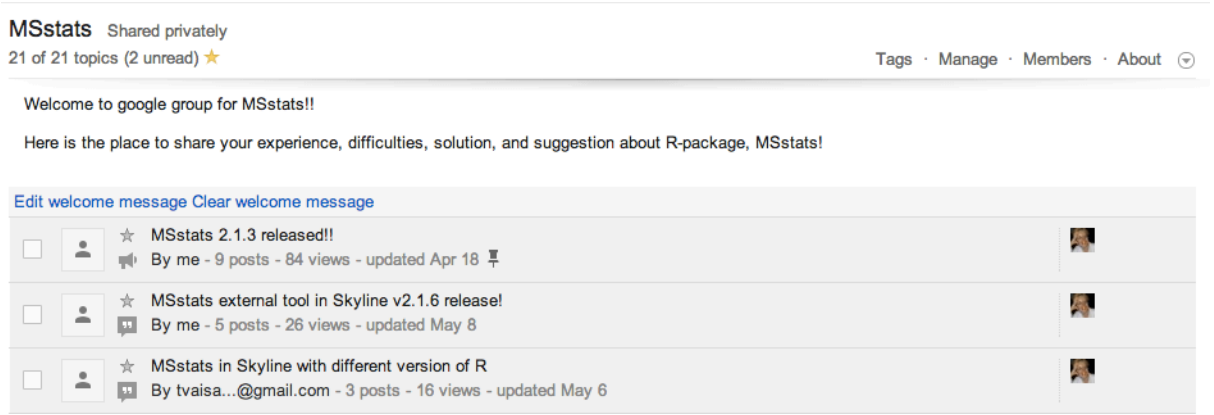
- Tutorial is available (https://skyline.gs.washington.edu/labkey/skyts/home/software/Skyline/tools/details.view?name=Msstats)

msstats.org and MSstats google group



- News about Msstats
- Tutorials for different workflows (under 'WORKFLOWS')
- Example datasets with R-scripts
- Related publications

- Announce new release or news in the mailing list
- Question and answer
- Discussion and suggestion



Acknowledgements

Purdue University

- Prof. Olga Vitek
- Mike Cheng
- Tsung-Heng Tsai

ETH Zürich

- Prof. Ruedi Aebersold
- Prof. Bernd Wollscheid
- Maria Pavlou
- Isabell Bludau

University of Washington

- Prof. Mike MacCoss and Lab
- Brendan MacLean
- Yuval Boss

Centre for Genomic Regulation

- Eduard Sabido
- Cristina Chiva

UCSF

- Erik Verschueren

Related works

Poster 038 : A system Suitability Monitoring Method of LC MS/MS Proteomic Experiment, Eralp Dogu

Poster 041 : Relative Protein Quantification in Mass Spectrometry-based Proteomics – A Split Plot Approach, Meena Choi

Poster 046 : Nonlinear Regression Avoids Overly Optimistic Assay Characterization, Cyril Galitzine

Oral presentation : Monday 10:40am-10:55am

gMODs : An Open Source Data Analysis Tool for Quantifying the Differential Site Occupancy of Therapeutic Protein Modification : Tsung-Heng Tsai

Parallel evening workshop : Monday 6:30 pm – 8 pm

Computation and statistics for quantitative proteomics, Grand C