Statistical analysis with MSstats

US HUPO short course 2015:

Design and analysis of quantitative proteomic experiment.

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

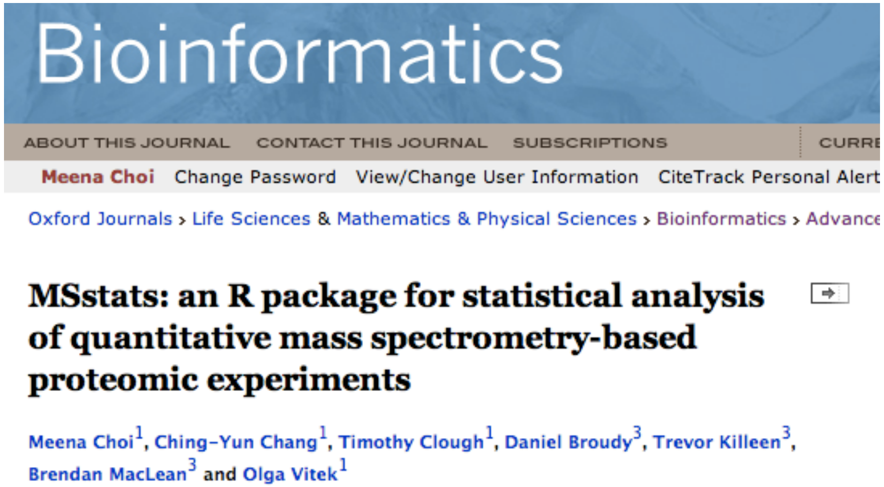
Outline

1. MSstats : statistical tool for quantitative MS proteomics
   - Workflow of MSstats
   - MSstats as an external tool
     - Integration of Skyline improves analysis workflow
     - User interface

2. Study of poor quality of peaks

3. How to access MSstats

MSstats : statistical tool for quantitative MS proteomics



Bioinformatics

ABOUT THIS JOURNAL
CONTACT THIS JOURNAL
SUBSCRIPTIONS
CURRE
Meena Choi
Change Password
View/Change User Information
CiteTrack Personal Alert
Oxford Journals › Life Sciences & Mathematics & Physical Sciences › Bioinformatics › Advance

MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments

Meena Choi[1], Ching-Yun Chang[1], Timothy Clough[1], Daniel Broudy[3], Trevor Killeen[3], Brendan MacLean[3] and Olga Vitek[1]

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

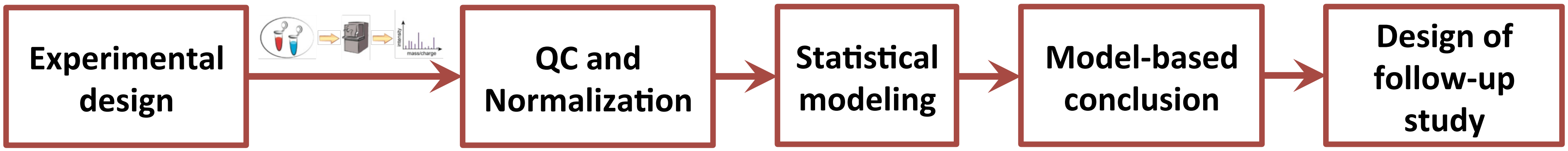
What we can do in MSstats

1. Test proteins for differential abundance
2. Quantify proteins in biological samples
3. Design of experiment

Type of experimental design

- Label-free workflows or workflows that use stable isotope labeled reference proteins and peptides
- SRM, DDA or shotgun, DIA or SWATH
- Comparisons of experimental conditions or times, or paired design

MSstats workflow : Experimental design



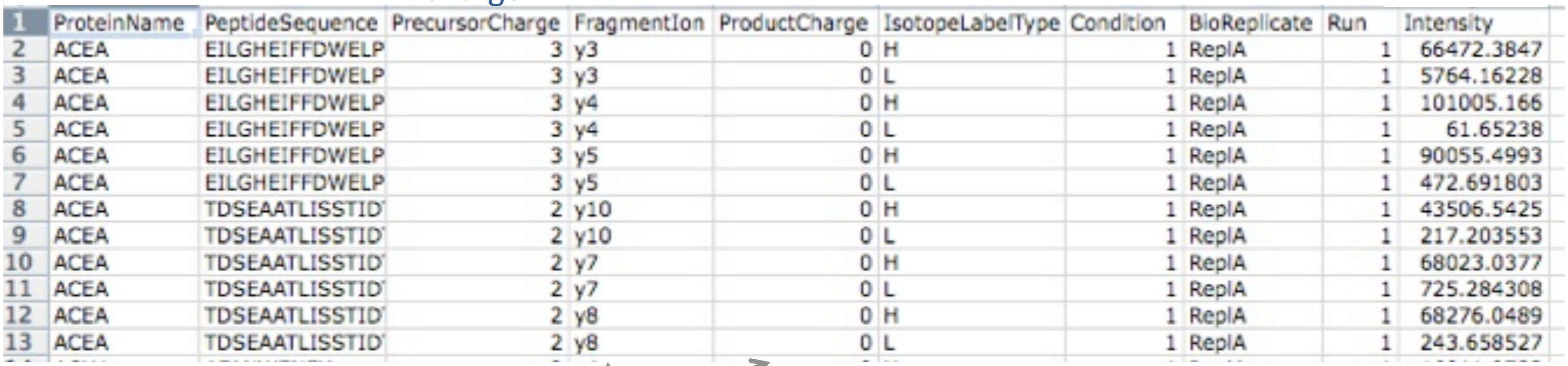
QC

Group Comparison

Design Sample Size

Input format

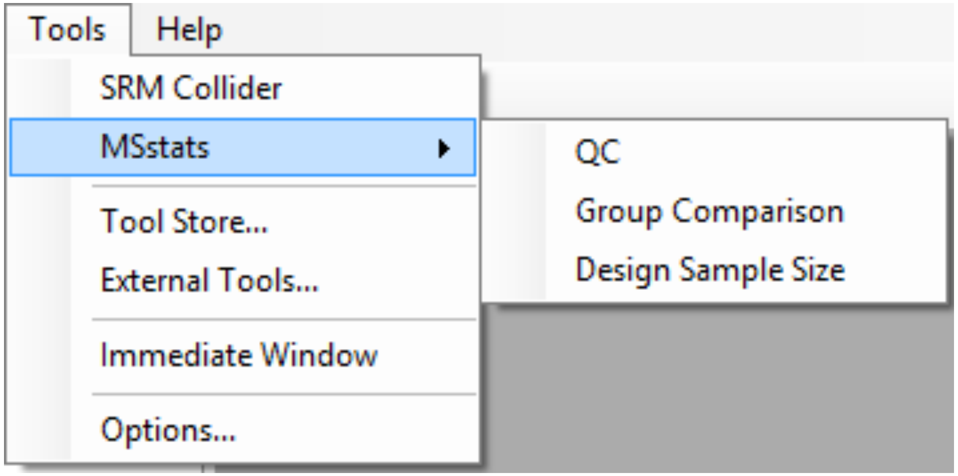
Protein Peptide Precursor charge Fragment Product charge Label Condition Subject Run Intensity



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

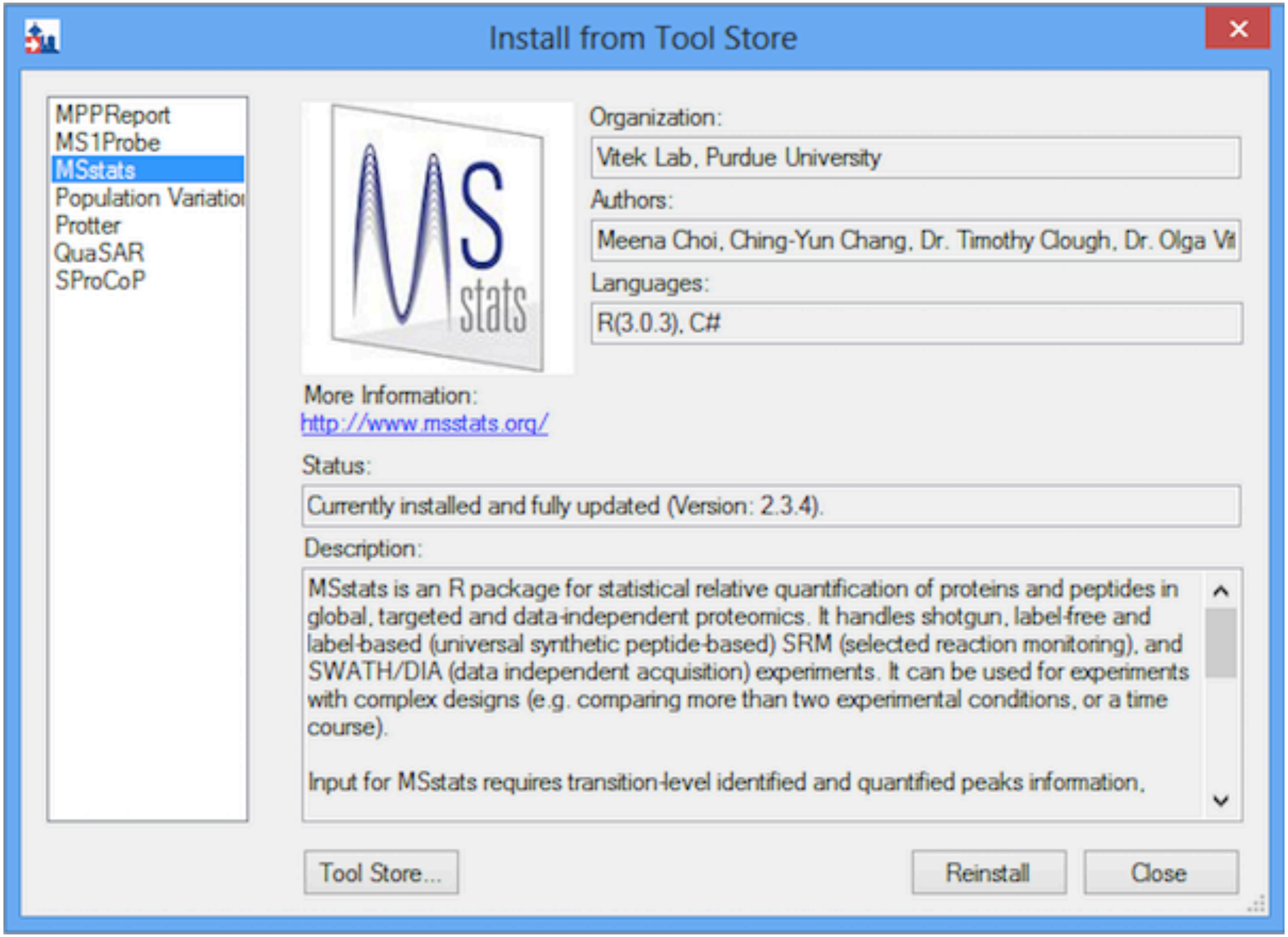
For DDA, ‘Fragment’, ‘ProductCharge’ can be any one value, such as NA

MSstats as an external tool



- Use as an external tool
- Automatically run the functions for
  - **Data processing** : Preprocessing the data, Drawing the profile plots, Quality control plots, Condition plots
  - **Group Comparison** : Comparing between groups, Drawing the plots with results
  - **Design Sample Size** : Calculating the sample size
- For the beginner of R or other statistical tools, we can do statistical analysis with default options through Skyline easily.

Set up MSstats as external tool



Organization: Vitek Lab, Purdue University

Authors: Meena Choi, Ching-Yun Chang, Dr. Timothy Clough, Dr. Olga Vi

Languages: R(3.0.3), C#

More Information: http://www.msstats.org/

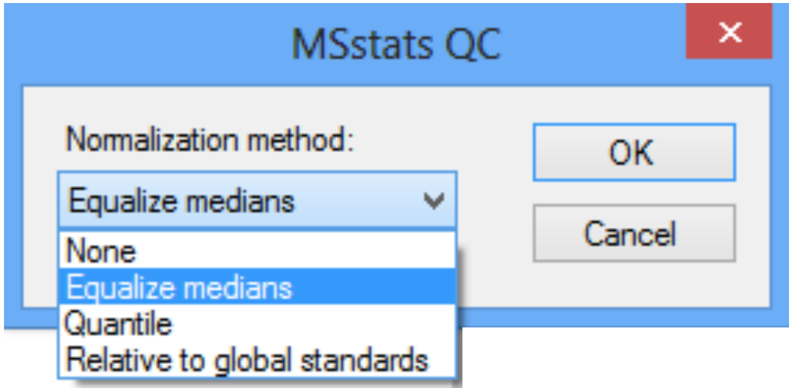
Status: Currently installed and fully updated (Version: 2.3.4).

Description:

MSstats is an R package for statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. It handles shotgun, label-free and label-based (universal synthetic peptide-based) SRM (selected reaction monitoring), and SWATH/DIA (data independent acquisition) experiments. It can be used for experiments with complex designs (e.g. comparing more than two experimental conditions, or a time course).

Input for MSstats requires transition-level identified and quantified peaks information,

1. QC : Data processing and normalization



Normalization method: Equalize medians

Data processing : Input with the report from Skyline

- get required report for analysis
- Log 2 or 10 transformation

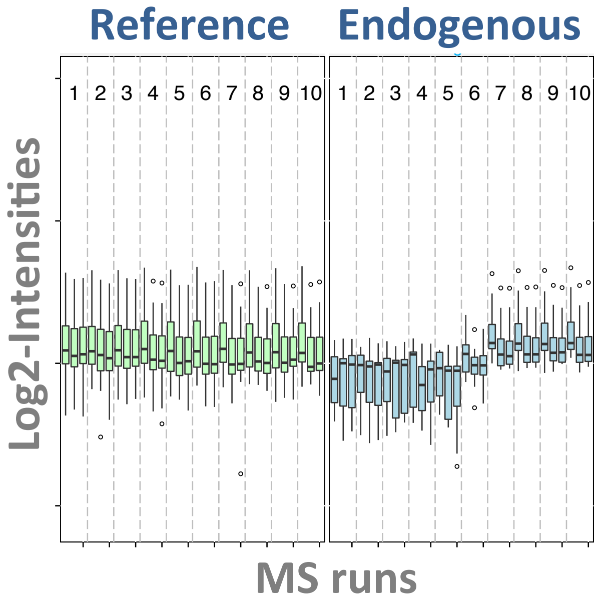
Normalization

- None : no normalization is performed.
- Constant : make the same median of reference intensities across runs.
- Quantile : equalize the distribution of reference intensities across runs.
- 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.

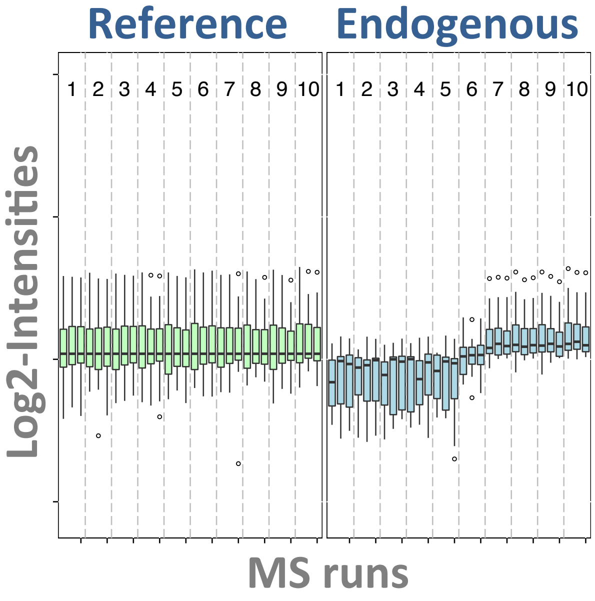
Quality control plots

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

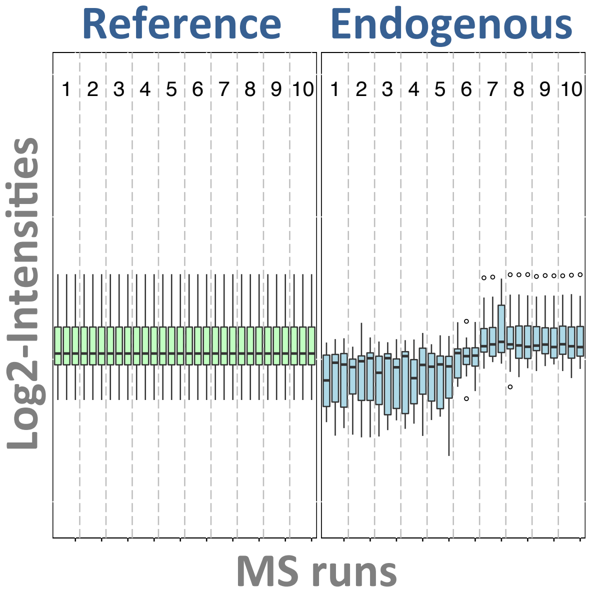
No normalization



Constant (equalize medians) normalization



Quantile normalization



Normalization method should be changed based on your design of experiment

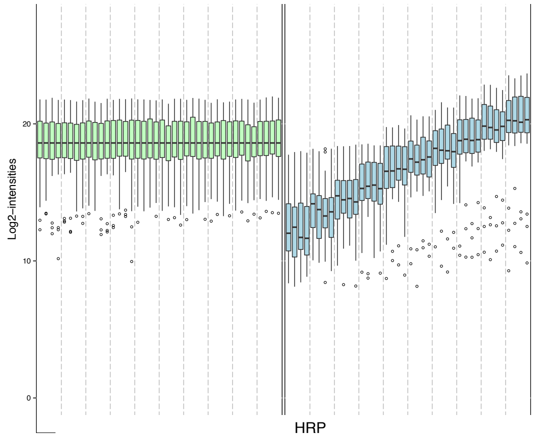
Constant (equalize medians) normalization

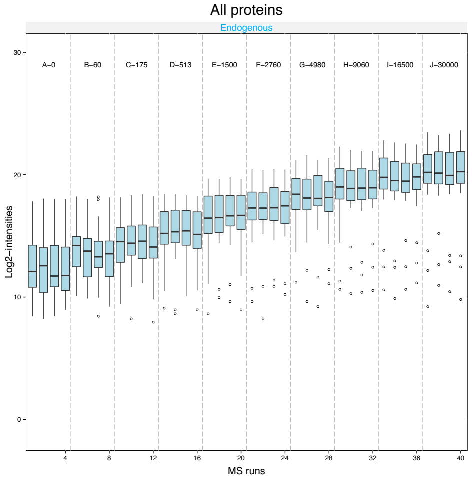
Assume label-free SRM :

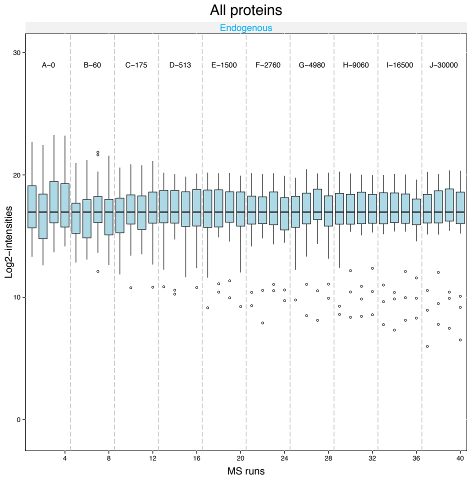
- most features are differently abundant

Reference Endogenous

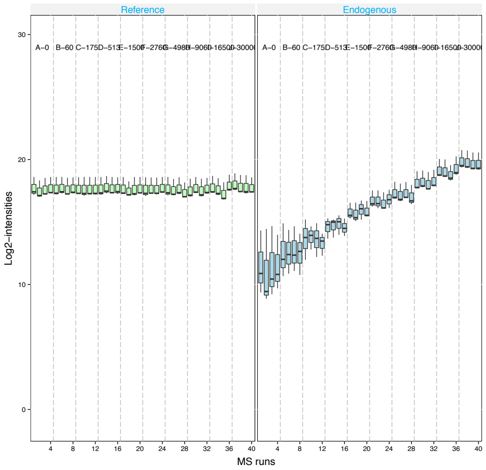
All proteins

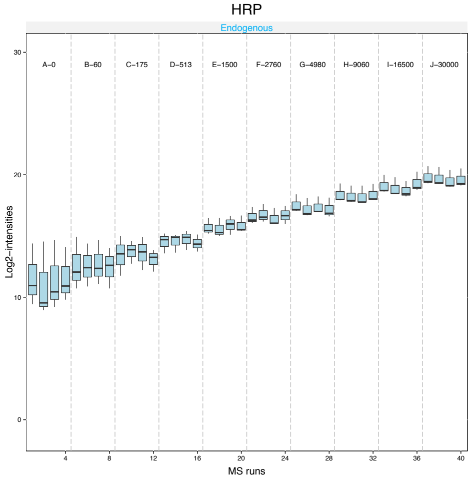


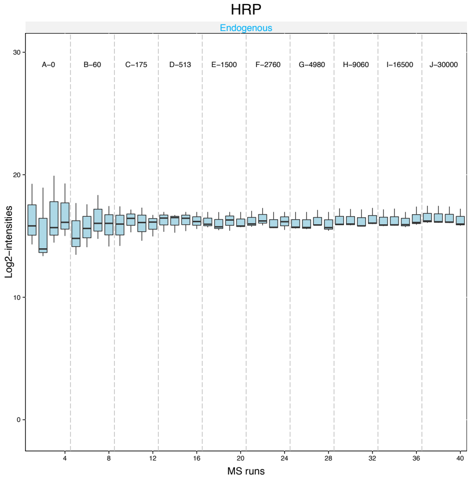




HRP







MS runs

Normalization method should be changed based on your design of experiment

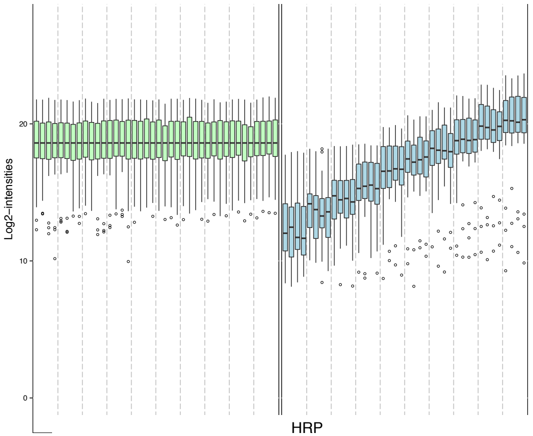
Assume label-free SRM :

- Need to concern normalization method in design stage

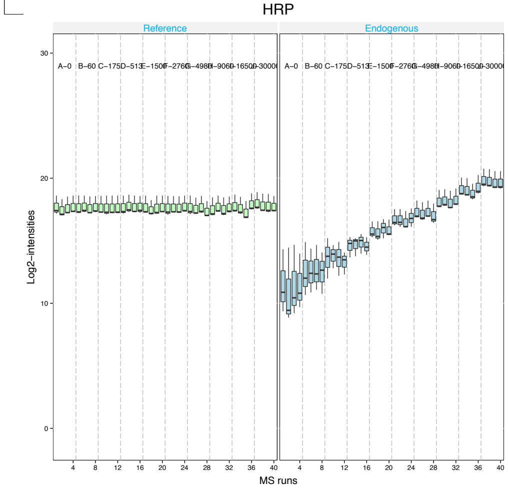
Constant (equalize medians) normalization

Reference **Endogenous**

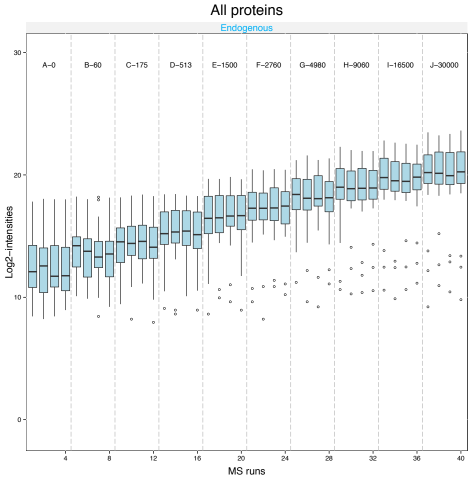
All proteins

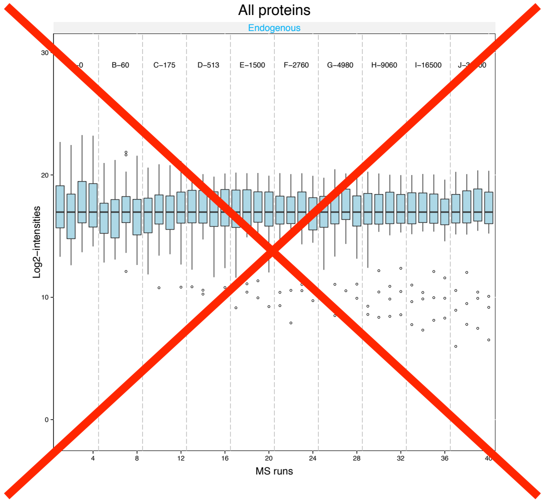


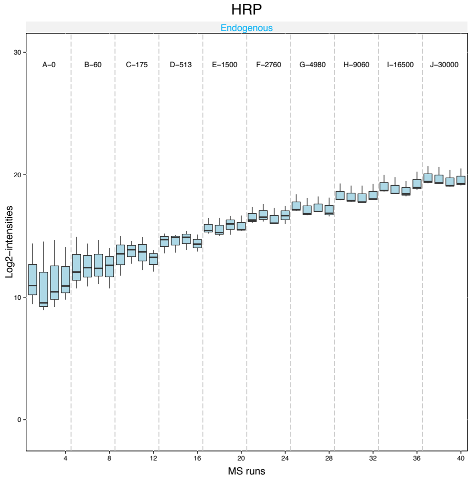
HRP

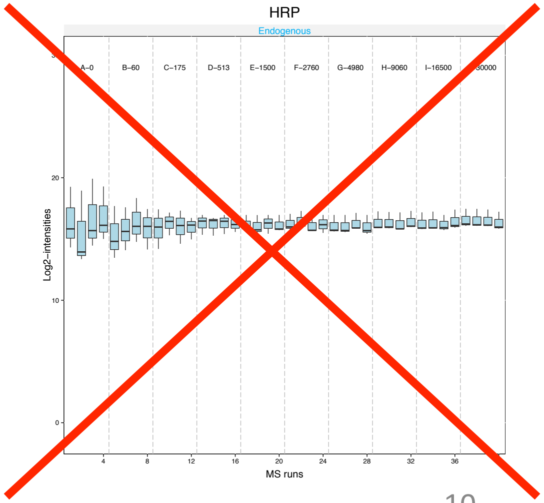


MS runs



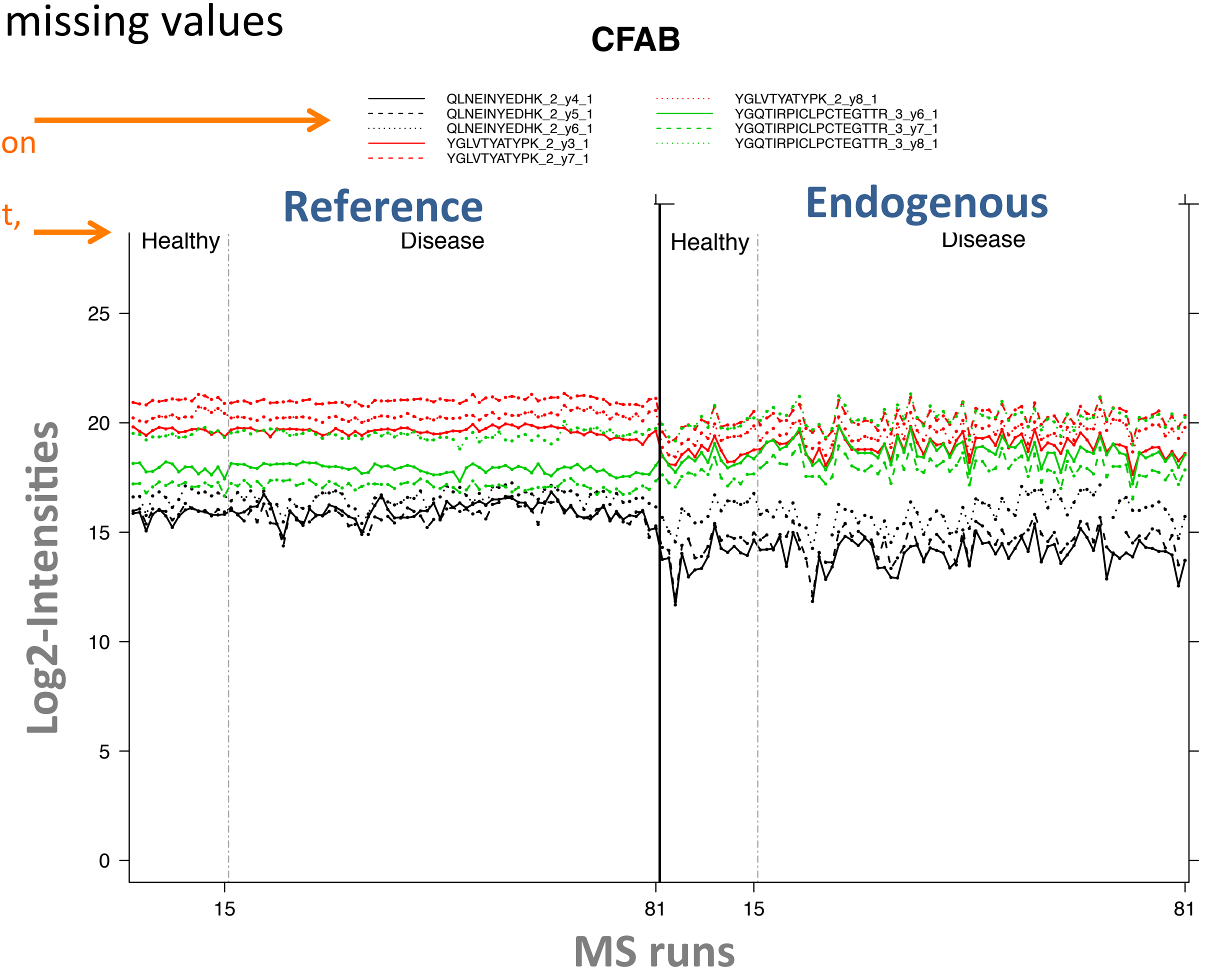






Profile Plot

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

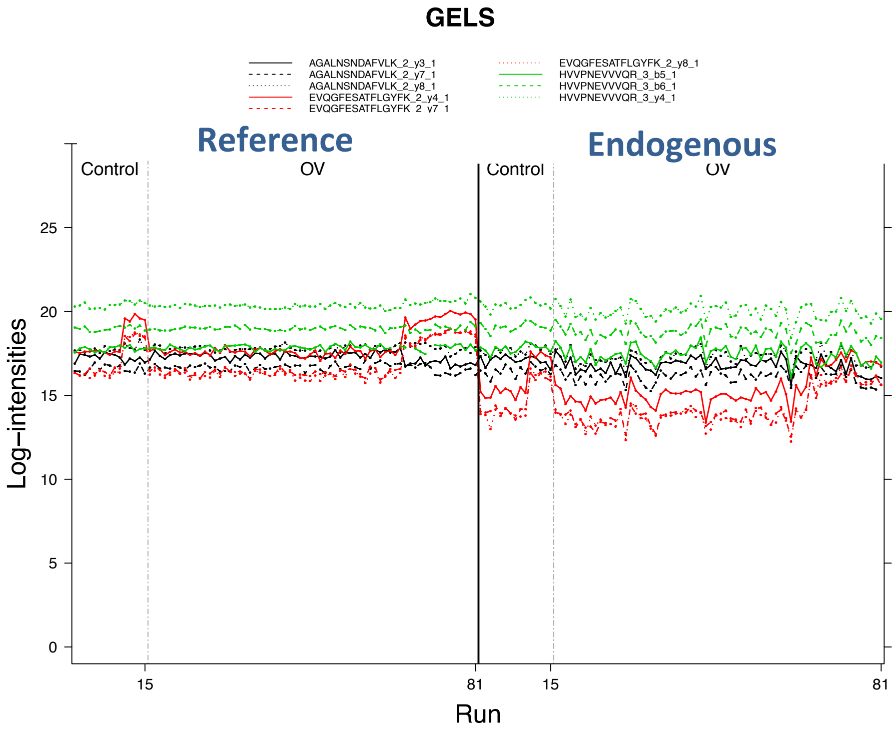


Color : peptide
Linetype : transition

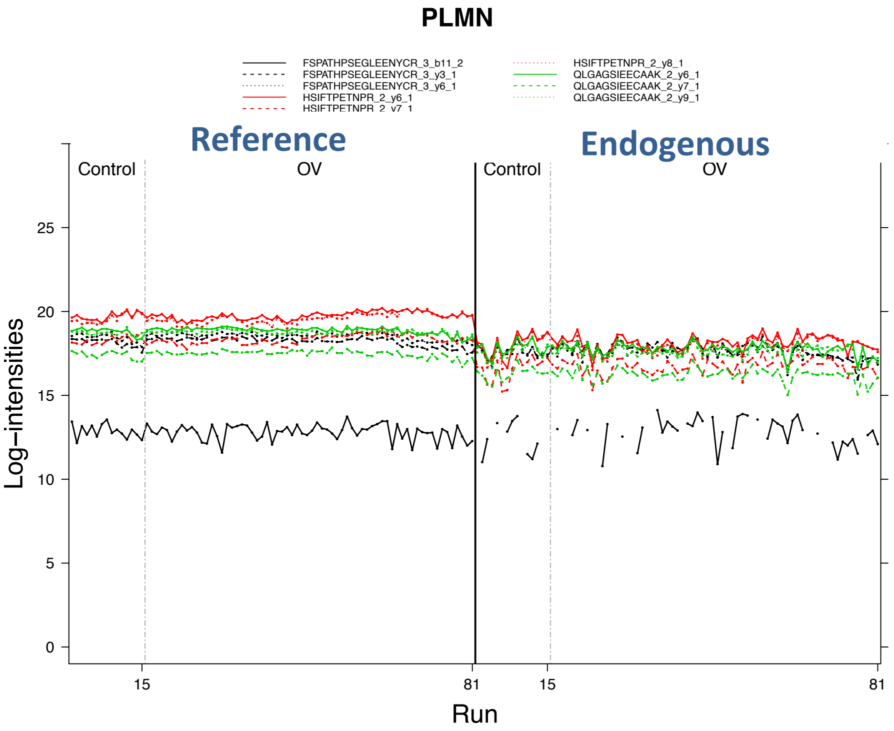
Reference or not,
Condition

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

Profile Plot



Detect the problematical Run or Transition

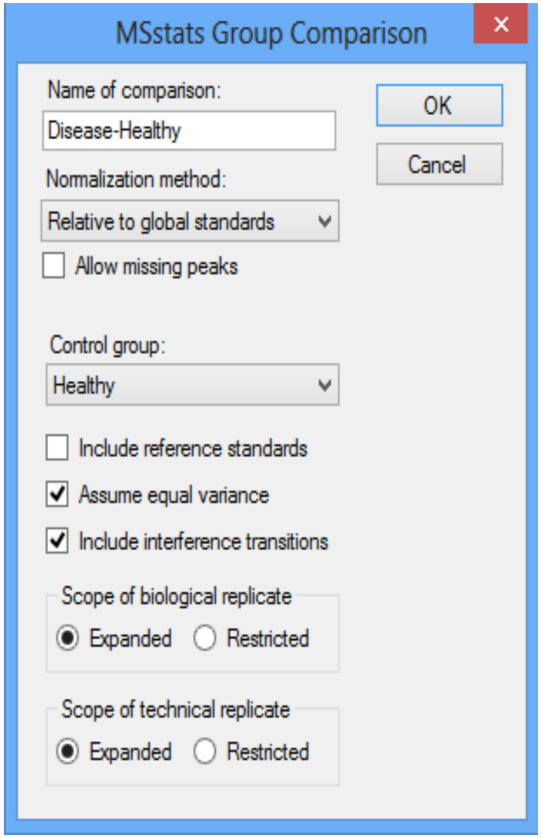


Show the missing values (Disconnection)

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



MSstats Group Comparison

Name of comparison: Disease-Healthy

Normalization method: Relative to global standards

□ Allow missing peaks

Control group: Healthy

□ Include reference standards

☑ Assume equal variance

☑ Include interference transitions

Scope of biological replicate: ● Expanded ○ Restricted

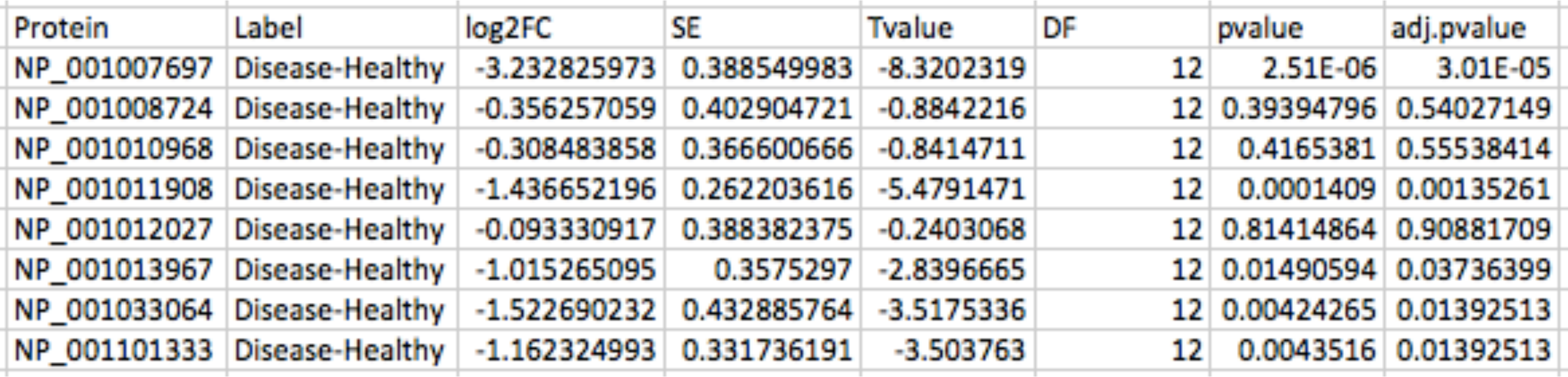
Scope of technical replicate: ● Expanded ○ Restricted

- Automatically detect the properties of the experimental design
  - Case-control study, Time-course study, Paired design

- Can choose the model
  - Presence of stable isotope labeled reference peptides
  - Assumption that all the features have equal noise variation between runs
  - Interference
    - contain interference transitions, need additional model interaction
  - with the desired scope of conclusion
    - Scope of biological replication : restricted / expanded
    - Scope of technical MS run replication : restricted/ expanded

Model-based conclusion

- Quantify the uncertainty
- Adjust p-values to control FDR
- Result will be saved in *TestingResult.csv*

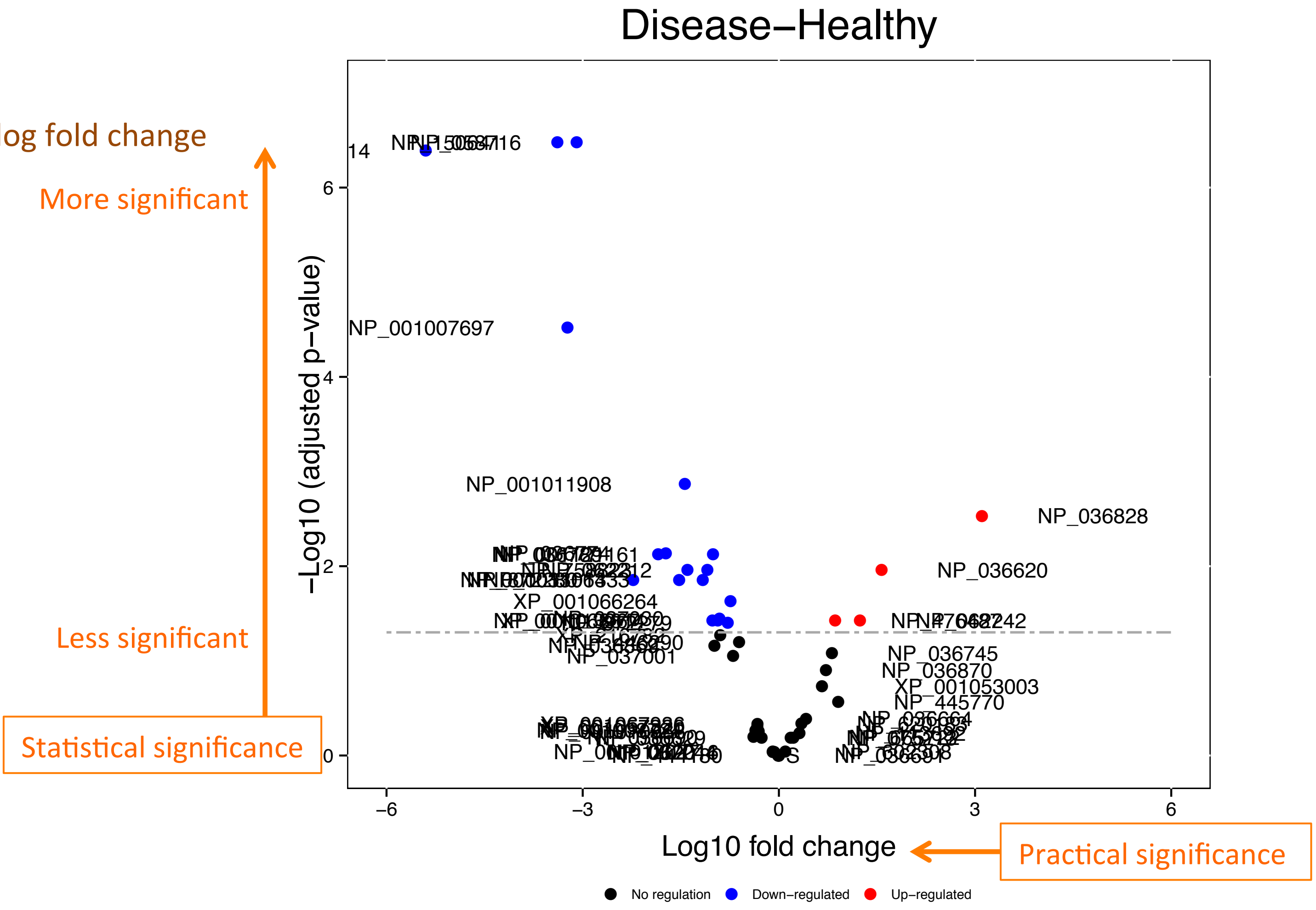


Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue
NP_001007697	Disease-Healthy	-3.232825973	0.388549983	-8.3202319	12	2.51E-06	3.01E-05
NP_001008724	Disease-Healthy	-0.356257059	0.402904721	-0.8842216	12	0.39394796	0.54027149
NP_001010968	Disease-Healthy	-0.308483858	0.366600666	-0.8414711	12	0.4165381	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

Volcano plot :

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



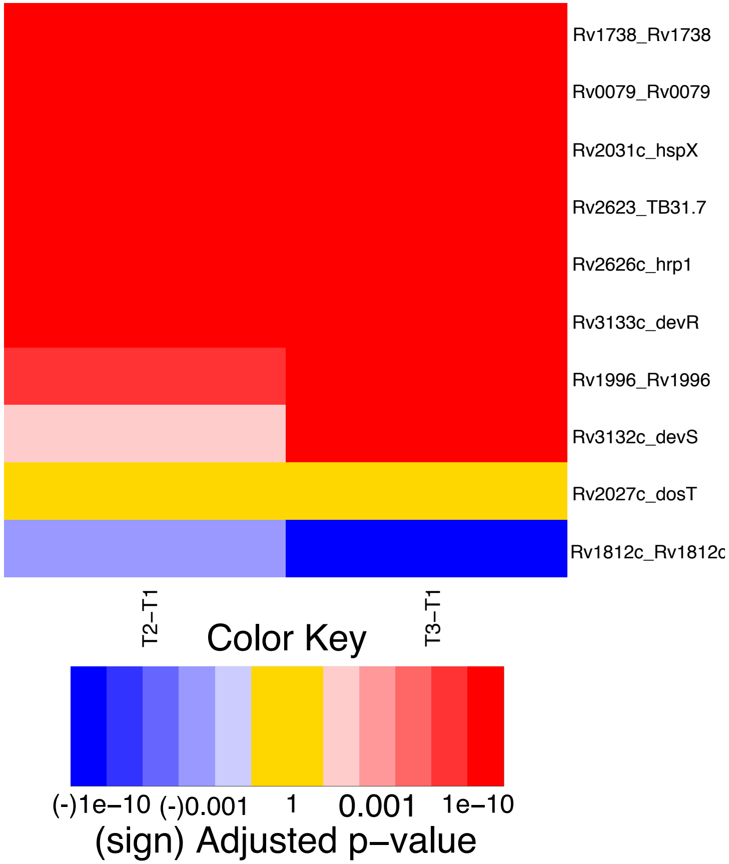
More significant

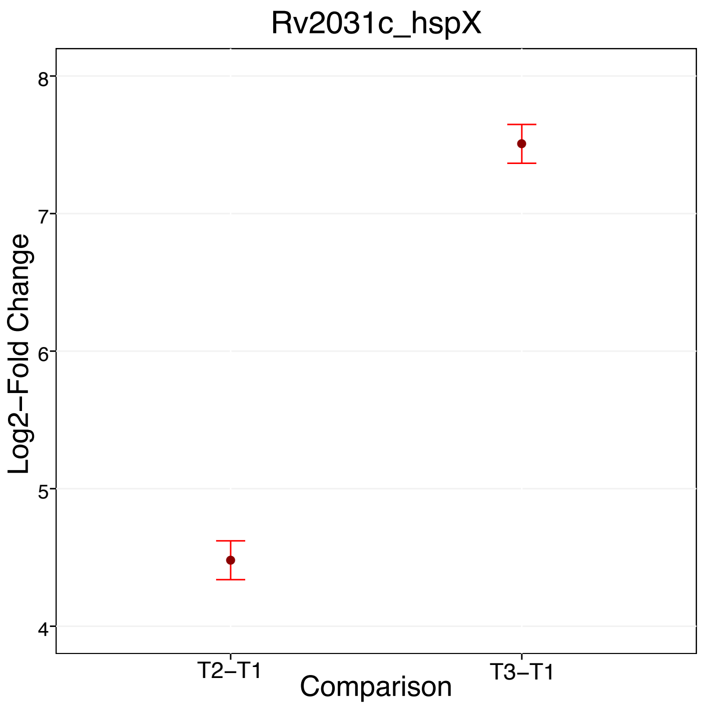
Less significant

Statistical significance

Practical significance

Visualization for multiple comparisons





Heatmap:

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

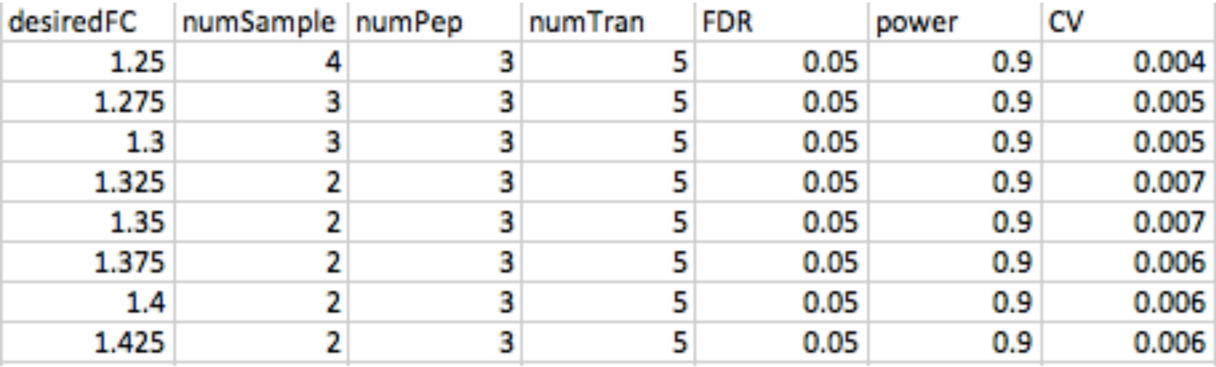
Comparison plot:

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

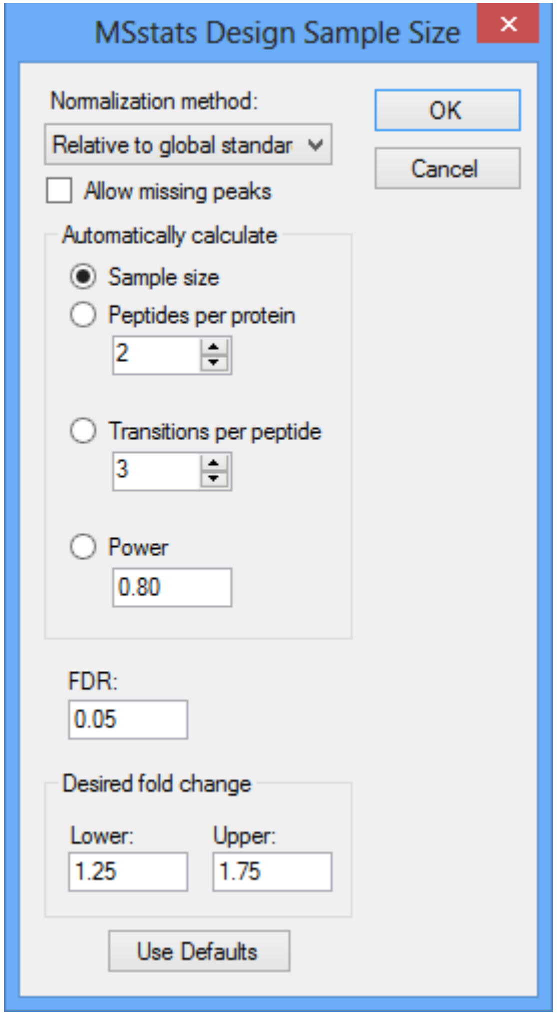
3. Design sample size : Design of future experiment

- Use the current dataset for variance estimation
- Also calculate
  - The number of peptide per protein
  - The number of transition per peptide
  - Power : the probability of detecting a true fold changes

- Result will be saved in *SampleSizeCalculation.csv*



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



MSstats Design Sample Size

Normalization method: Relative to global standar

□ Allow missing peaks

Automatically calculate

● Sample size

○ Peptides per protein: 2

○ Transitions per peptide: 3

○ Power: 0.80

FDR: 0.05

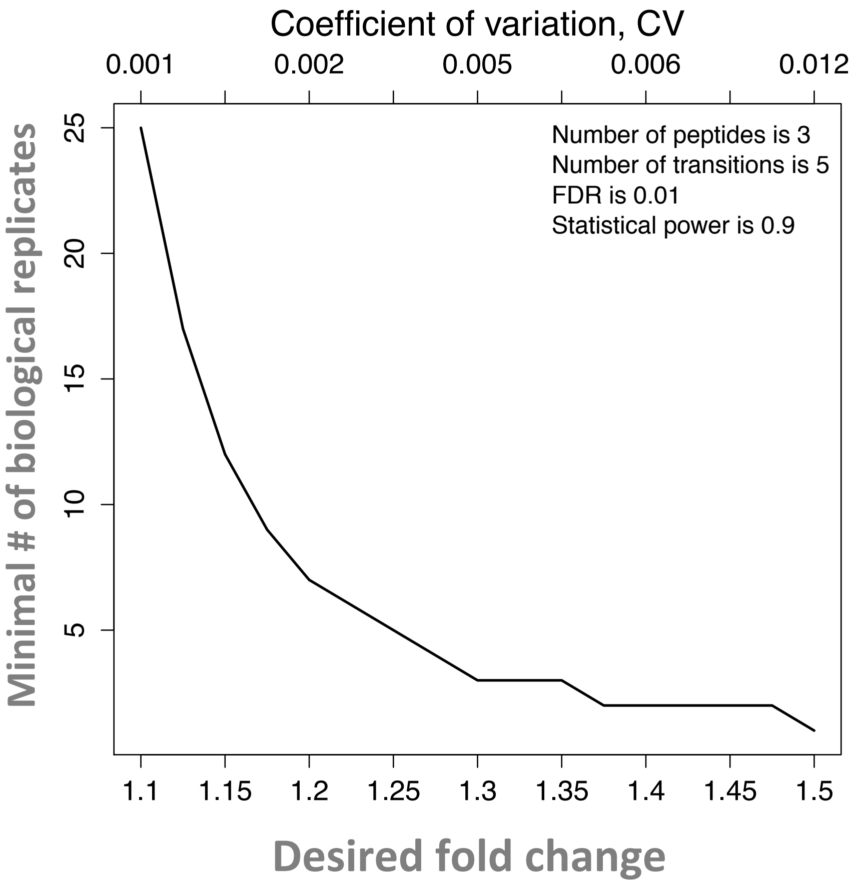
Desired fold change

Lower: 1.25

Upper: 1.75

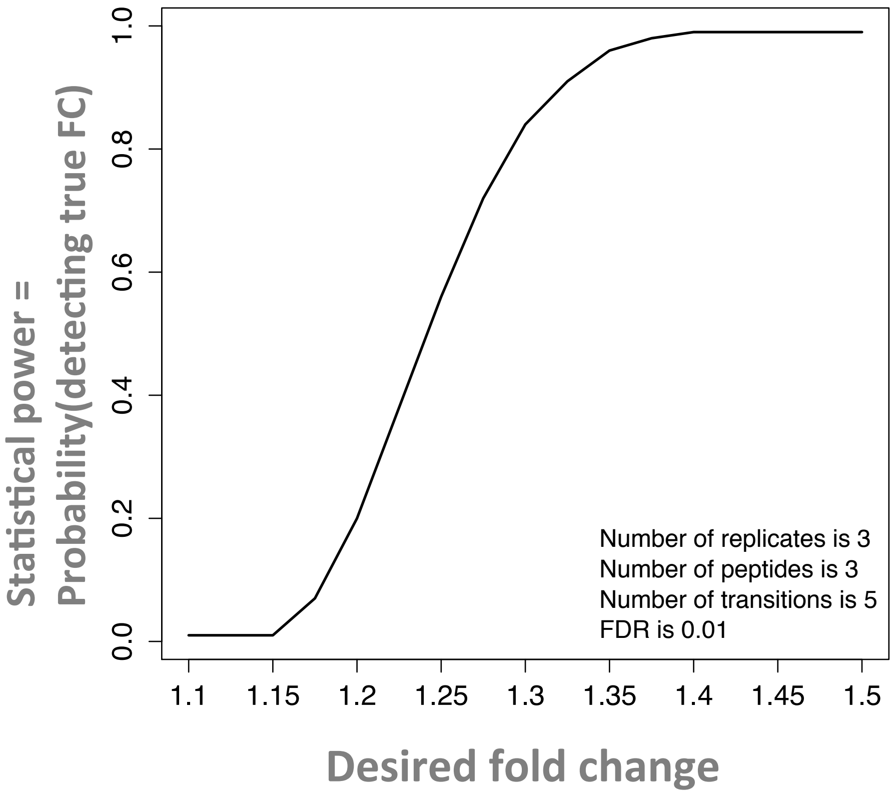
Sample size calculation and power

Sample size calculation



Number of peptides is 3
Number of transitions is 5
FDR is 0.01
Statistical power is 0.9

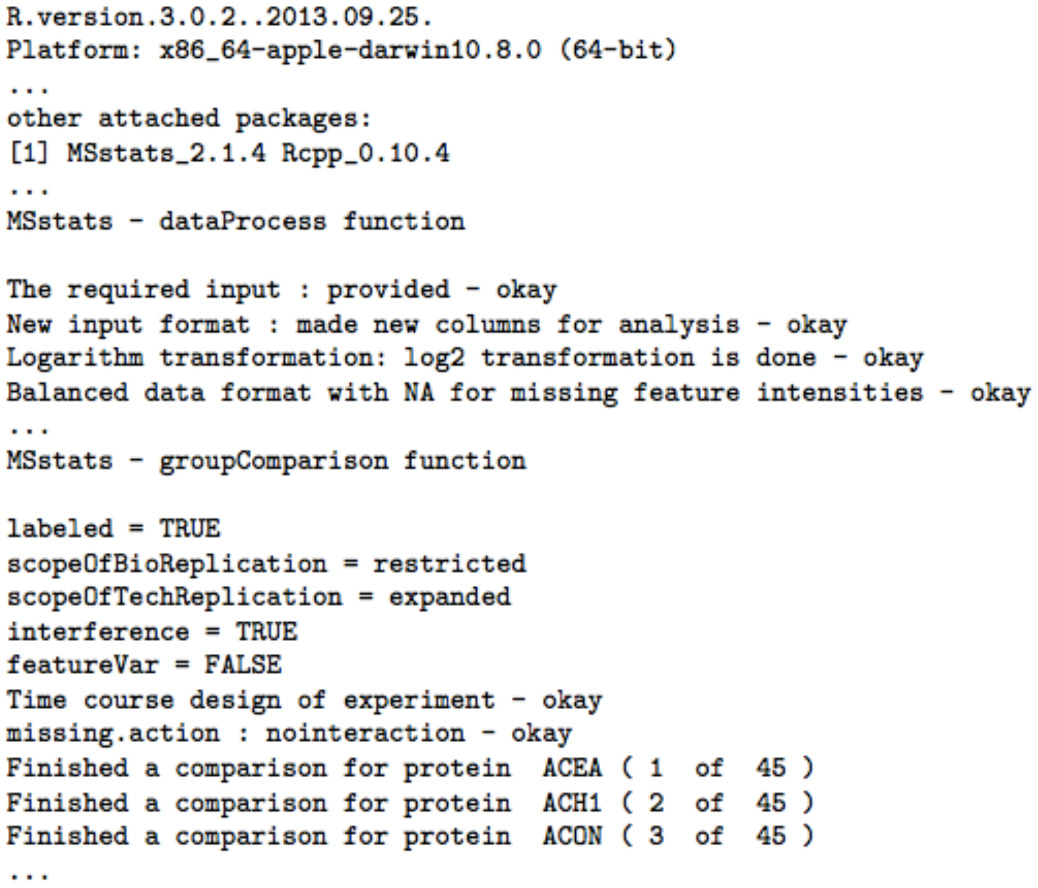
Power calculation



Number of replicates is 3
Number of peptides is 3
Number of transitions is 5
FDR is 0.01

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.3.0.2..2013.09.25.
Platform: x86_64-apple-darwin10.8.0 (64-bit)
...
other attached packages:
[1] MSstats_2.1.4 Rcpp_0.10.4
...
MSstats - dataProcess function

The required input : provided - okay
New input format : made new columns for analysis - okay
Logarithm transformation: log2 transformation is done - okay
Balanced data format with NA for missing feature intensities - okay
...
MSstats - groupComparison function

labeled = TRUE
scopeOfBioReplication = restricted
scopeOfTechReplication = expanded
interference = TRUE
featureVar = FALSE
Time course design of experiment - okay
missing.action : nointeraction - okay
Finished a comparison for protein  ACEA ( 1  of  45 )
Finished a comparison for protein  ACH1 ( 2  of  45 )
Finished a comparison for protein  ACON ( 3  of  45 )
...
```

Outline

1. MSstats : statistical tool for quantitative MS proteomics
   - Workflow of MSstats
   - MSstats as an external tool
     - Integration of Skyline improves analysis workflow
     - User interface

2. Study of poor quality of peaks

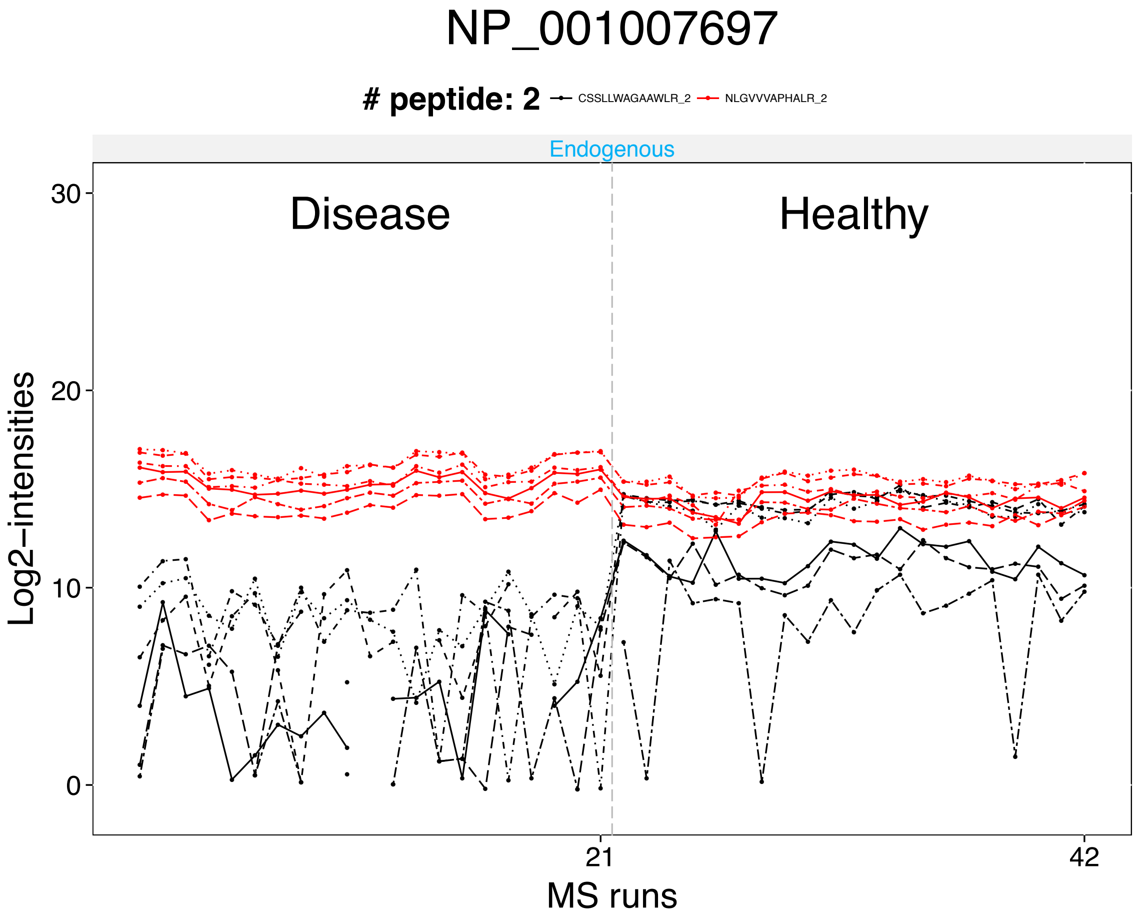
3. How to access MSstats

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

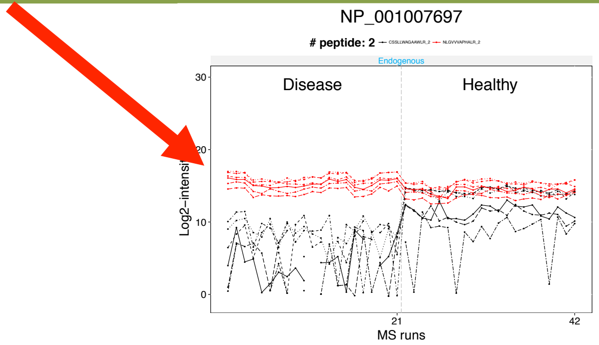
Examples of inconsistent (poor quality?) peptides

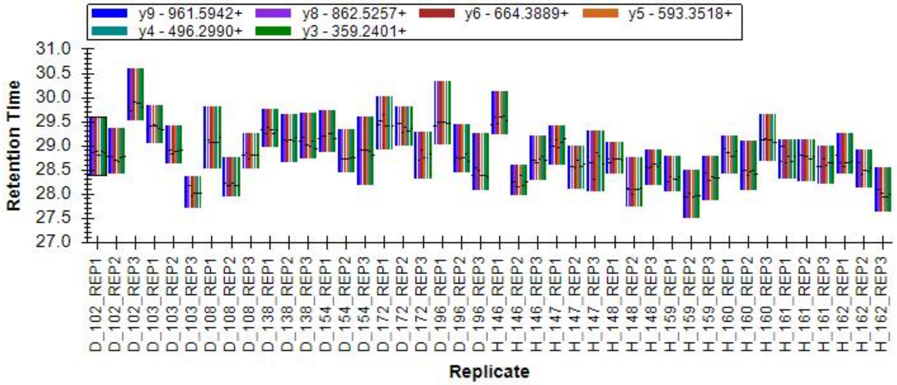


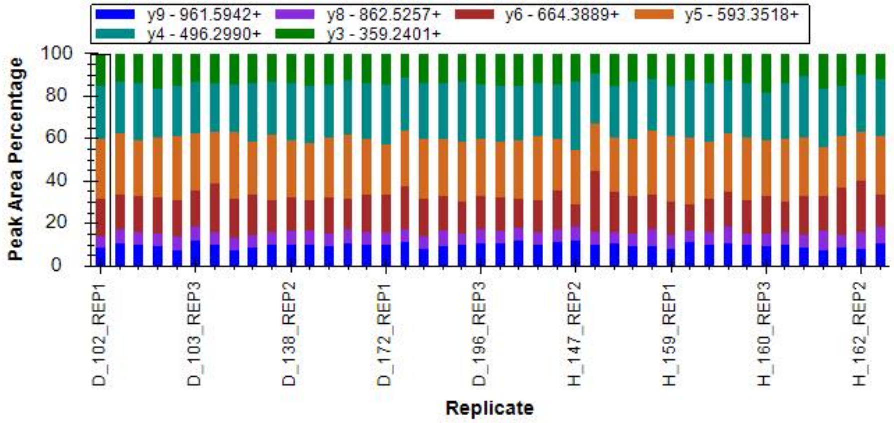
Profile plot show the problematic peptides or transitions. We need to check what happen in this peptide.

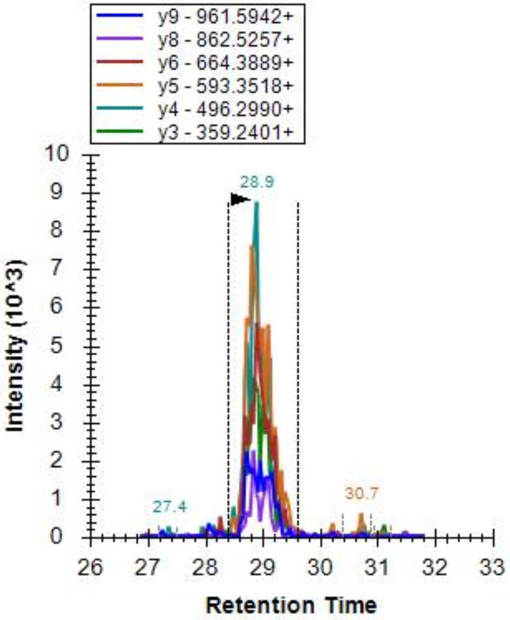
Rat Plasma : label-free SRM

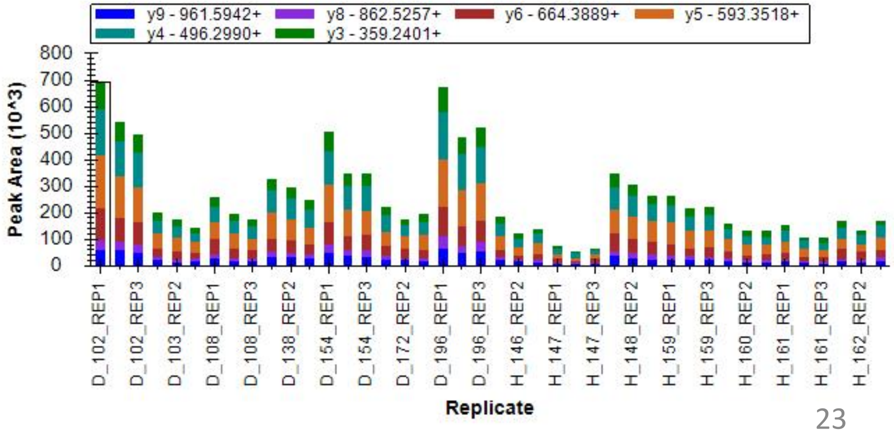
NLGVVVAPHALR





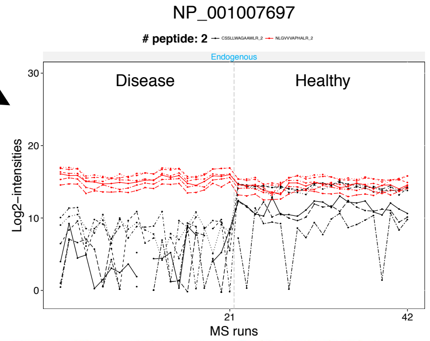


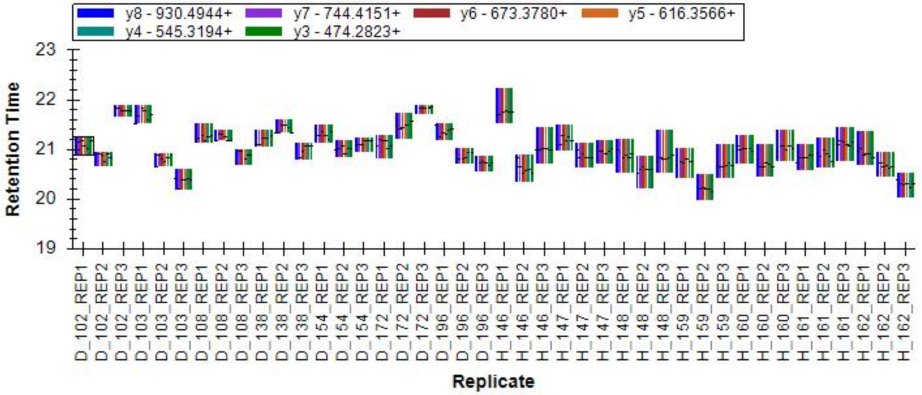


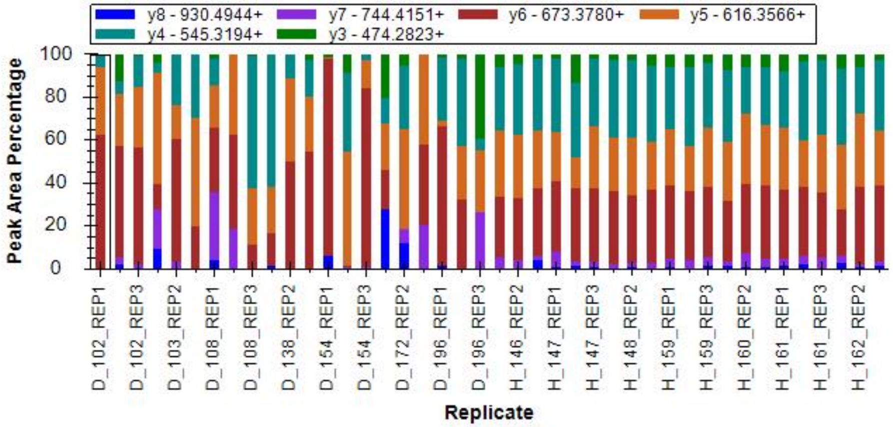


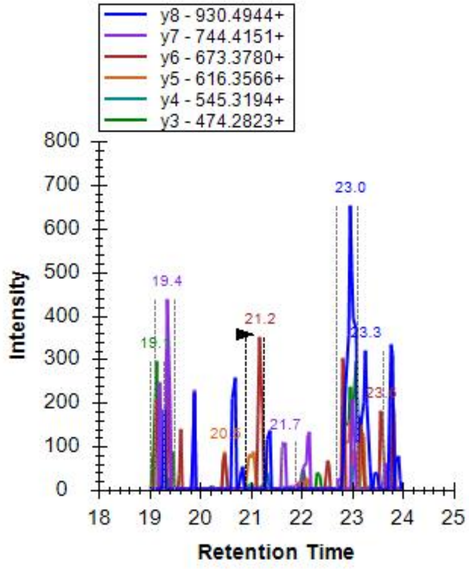
Rat Plasma : label-free SRM

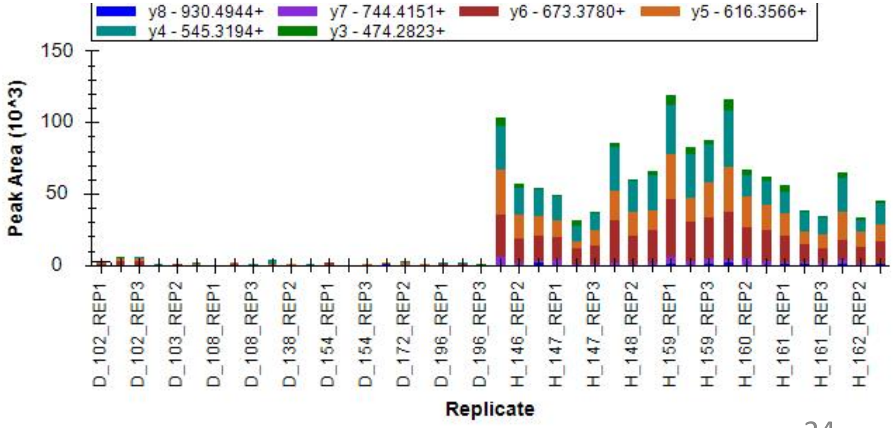
CSSLLWAGAAWLR



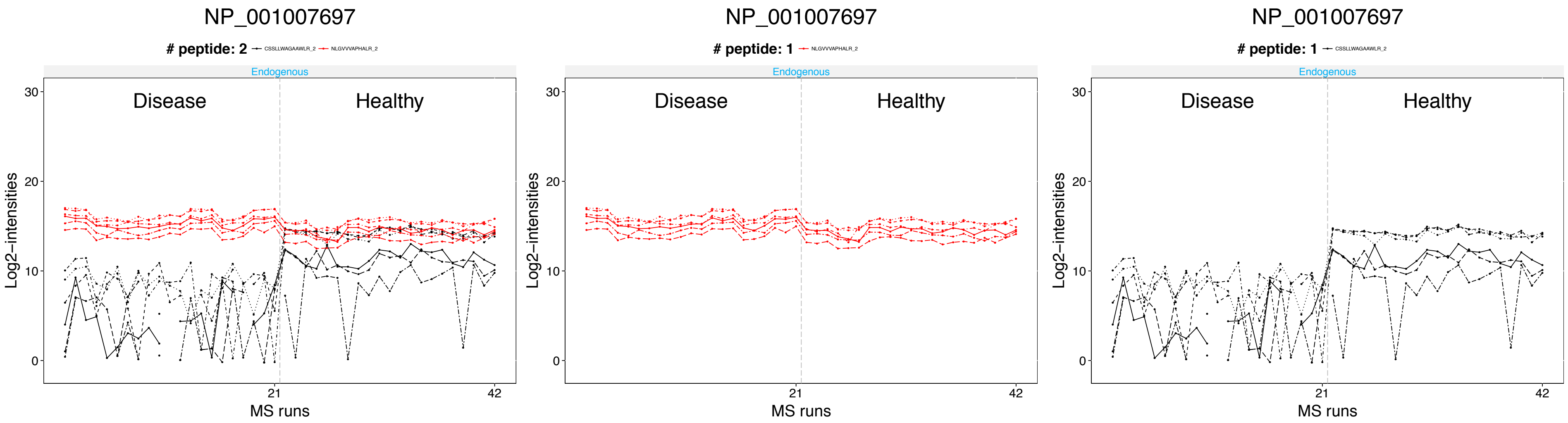






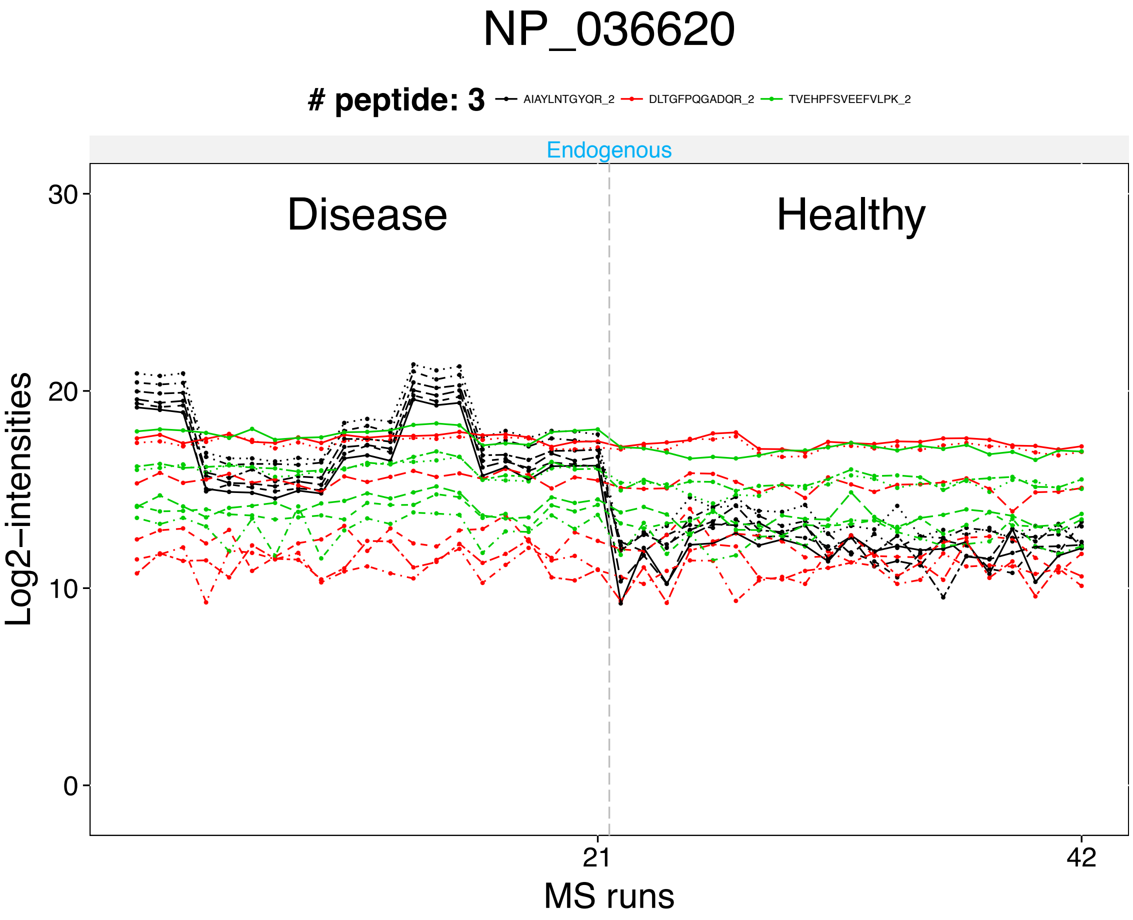


Log2 FC and variation are different between before and after removing peptides



	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
Fixed Subject	-2.6721	0.1439	<0.0001	0.8750	0.0260	<0.0001	-6.2272	0.2868	<0.0001
Random Subject	-2.6701	0.2214	<0.0001	0.8750	0.2399	0.0066	-6.2187	0.4152	<0.0001

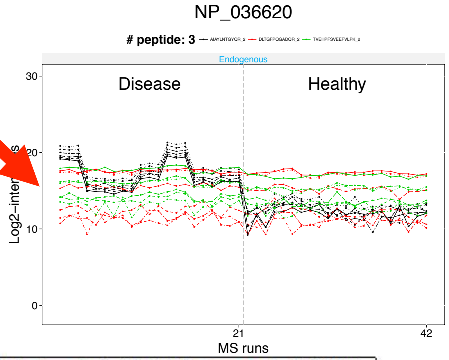
Examples of inconsistent peptides

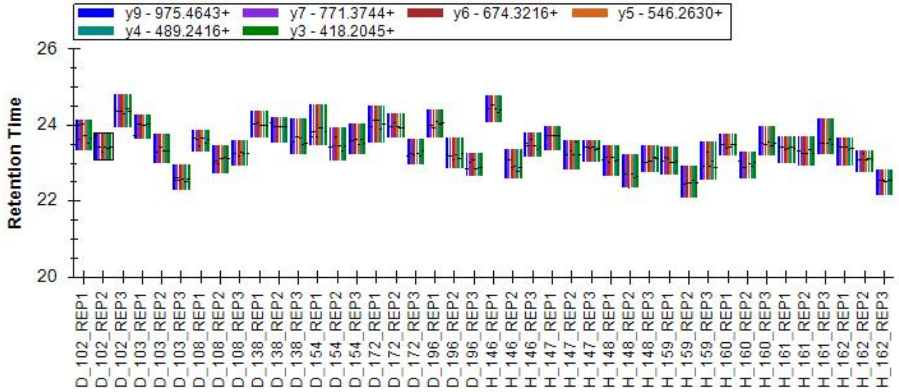


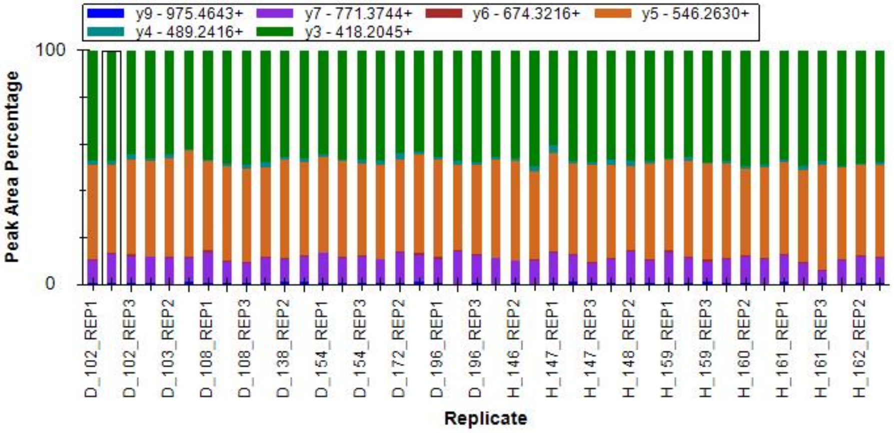
Profile plot show inconsistent pattern per peptides. We need to check that is there any measurement problem.

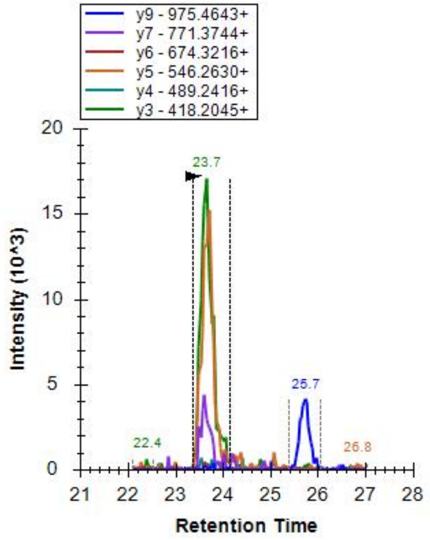
Rat Plasma : label-free SRM

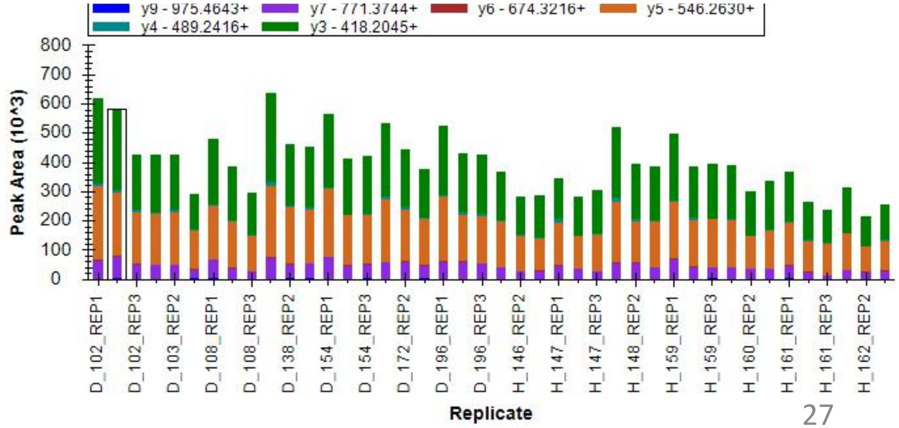
DLTGFPQGADQR





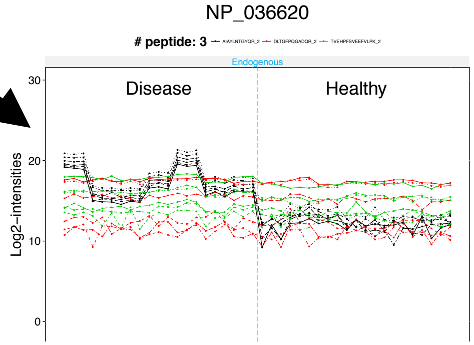


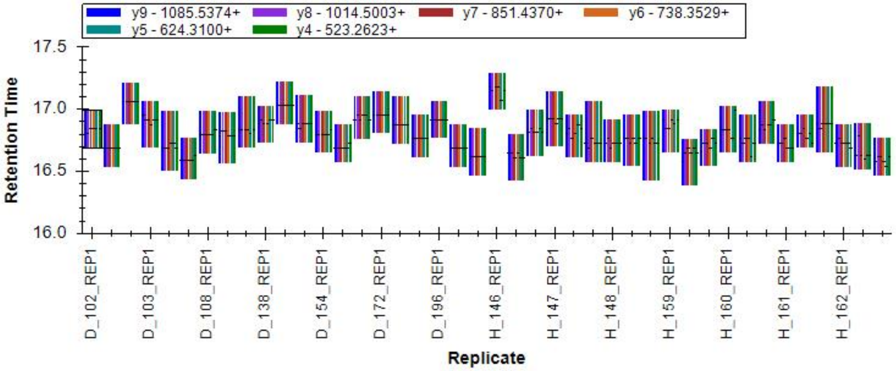


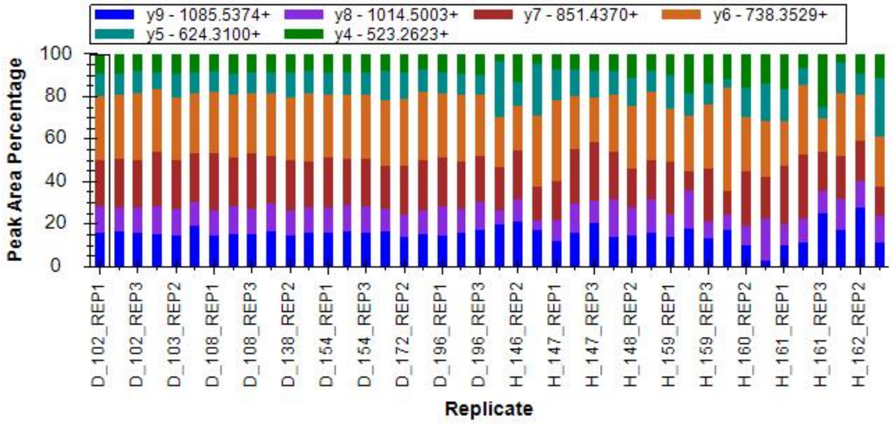


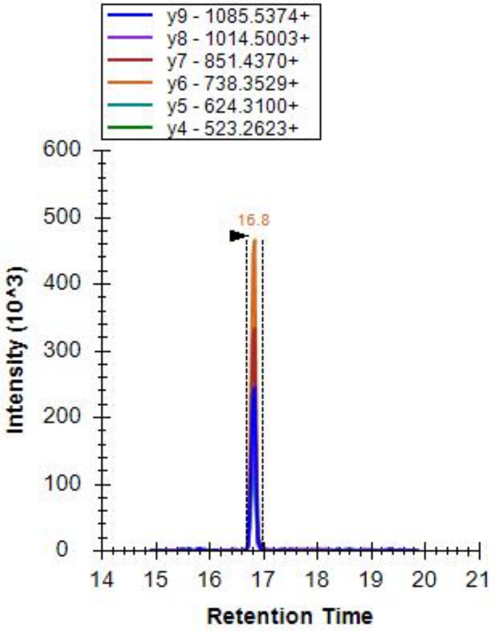
Rat Plasma : label-free SRM

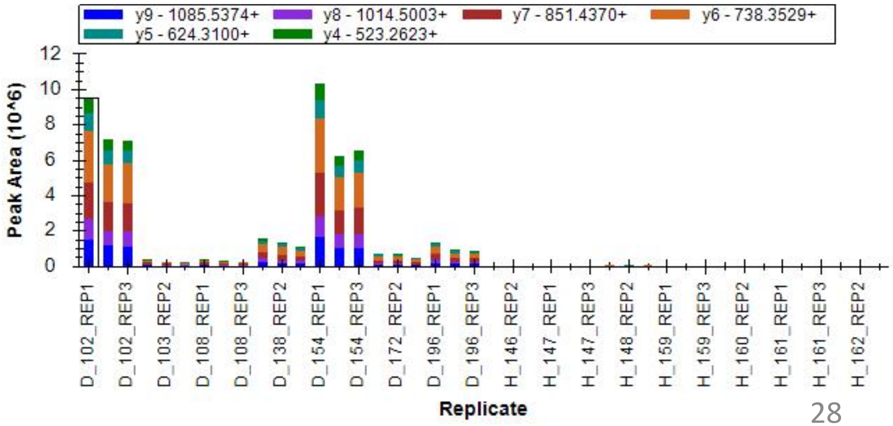
AIAYLNTGYQR





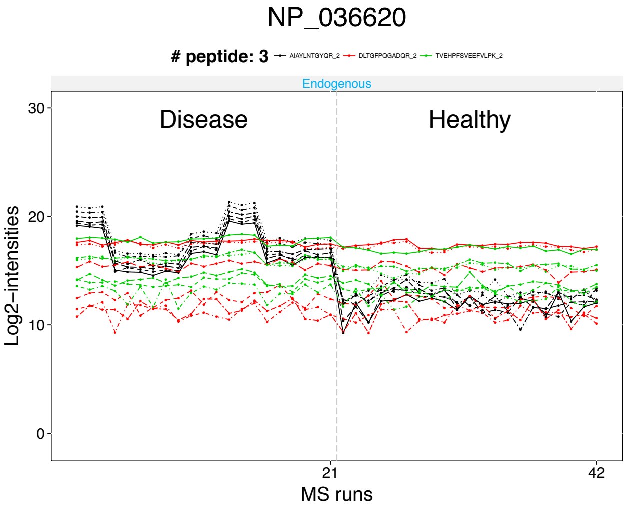


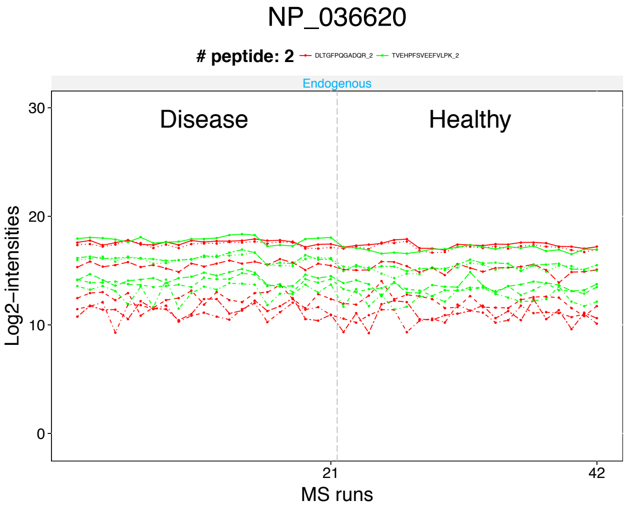


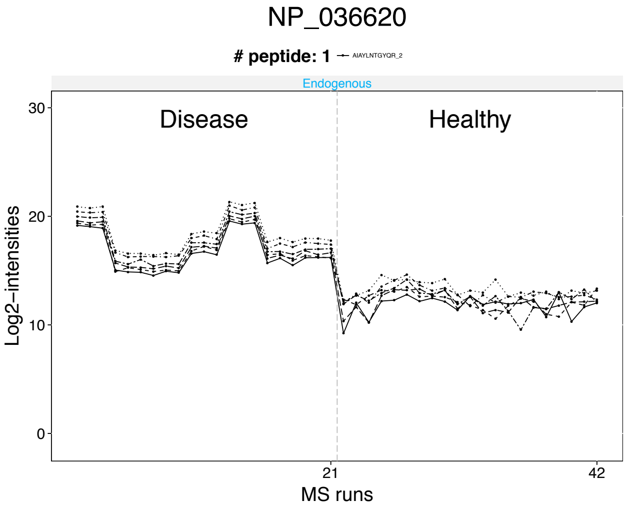


Rat Plasma : label-free SRM

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
Fixed Subject	2.0642	0.0951	<0.0001	0.6167	0.0414	<0.0001	5.0812	0.0591	<0.0001
Random Subject	2.0642	0.2966	<0.0001	0.6167	0.1137	0.0005	5.0812	0.7390	<0.0001

Summary of poor quality peptides

- Profile plot show inconsistent pattern per peptides. We need to check that is there any measurement problem.
- Less certainty that you look at the correct peptide,
  - Due to different reasons such as any phosphorylation and modification in peptide level.
  - suggestion : re-measure in label-based way.
- Need to investigate further a subset of peptides that we find interesting for some reason.

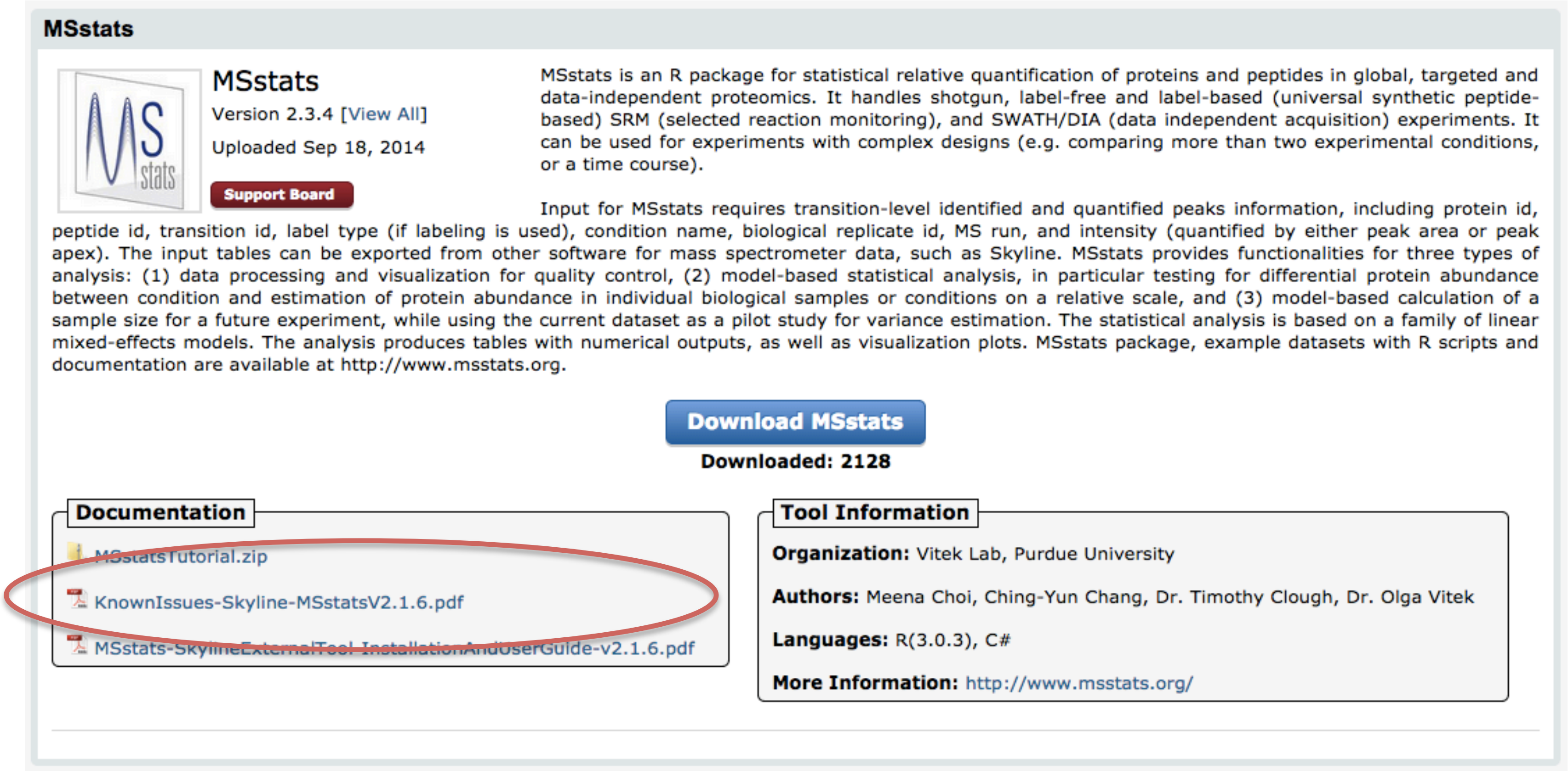
Outline

1. MSstats : statistical tool for quantitative MS proteomics
    - Workflow of MSstats
    - MSstats as an external tool
        - Integration of Skyline improves analysis workflow
        - User interface

2. Study of poor quality of peaks

3. How to access MSstats

External tool in Skyline



MSstats

MSstats

Version 2.3.4 [View All]

Uploaded Sep 18, 2014

Support Board

MSstats is an R package for statistical relative quantification of proteins and peptides in global, targeted and data-independent proteomics. It handles shotgun, label-free and label-based (universal synthetic peptide-based) SRM (selected reaction monitoring), and SWATH/DIA (data independent acquisition) experiments. It can be used for experiments with complex designs (e.g. comparing more than two experimental conditions, or a time course).

Input for MSstats requires transition-level identified and quantified peaks information, including protein id, peptide id, transition id, label type (if labeling is used), condition name, biological replicate id, MS run, and intensity (quantified by either peak area or peak apex). The input tables can be exported from other software for mass spectrometer data, such as Skyline. MSstats provides functionalities for three types of analysis: (1) data processing and visualization for quality control, (2) model-based statistical analysis, in particular testing for differential protein abundance between condition and estimation of protein abundance in individual biological samples or conditions on a relative scale, and (3) model-based calculation of a sample size for a future experiment, while using the current dataset as a pilot study for variance estimation. The statistical analysis is based on a family of linear mixed-effects models. The analysis produces tables with numerical outputs, as well as visualization plots. MSstats package, example datasets with R scripts and documentation are available at http://www.msstats.org.

Download MSstats

Downloaded: 2128

Documentation

- MSstatsTutorial.zip
- KnownIssues-Skyline-MSstatsV2.1.6.pdf
- MSstats-SkylineExternalTool-InstallationAndUserGuide-v2.1.6.pdf

Tool Information

Organization: Vitek Lab, Purdue University

Authors: Meena Choi, Ching-Yun Chang, Dr. Timothy Clough, Dr. Olga Vitek

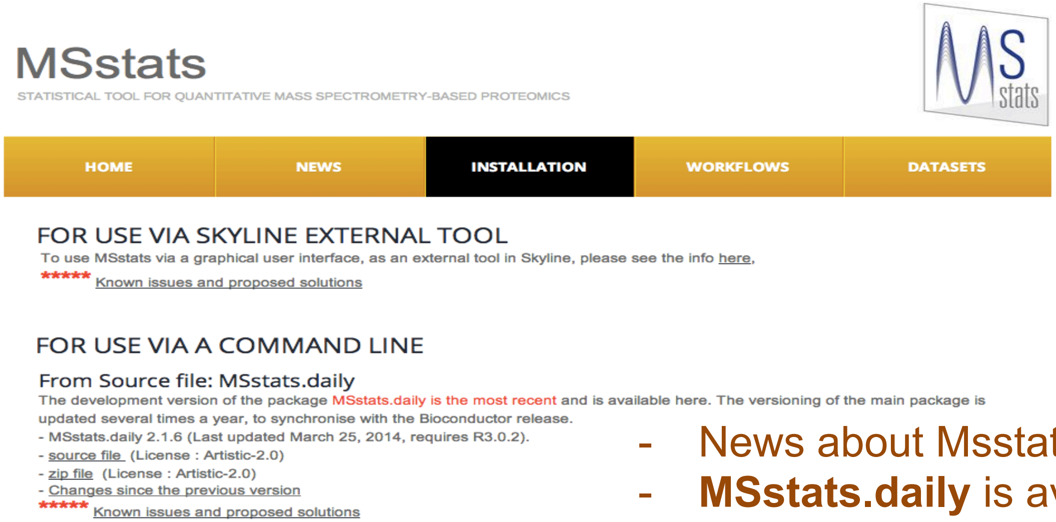
Languages: R(3.0.3), C#

More Information: http://www.msstats.org/

- 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



MSstats

STATISTICAL TOOL FOR QUANTITATIVE MASS SPECTROMETRY-BASED PROTEOMICS

FOR USE VIA SKYLINE EXTERNAL TOOL

To use MSstats via a graphical user interface, as an external tool in Skyline, please see the info here,

****** Known issues and proposed solutions

FOR USE VIA A COMMAND LINE

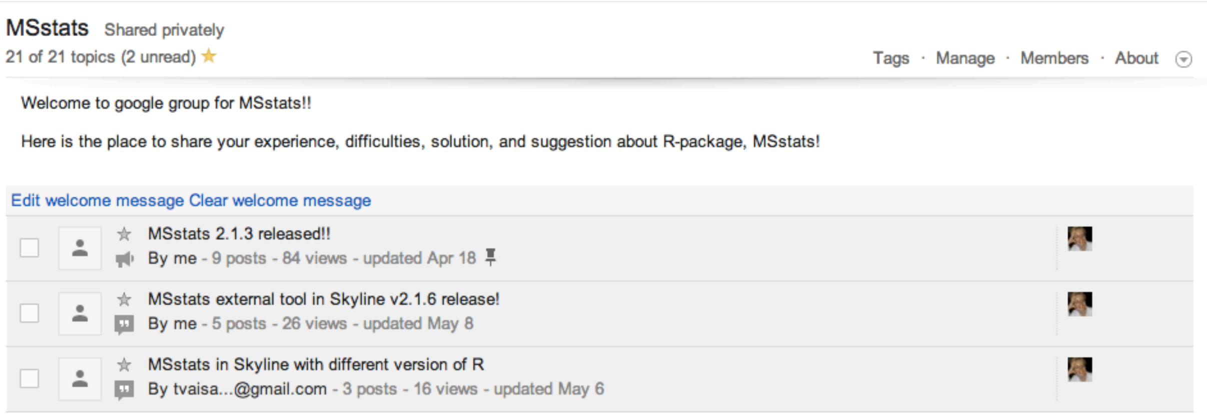
From Source file: MSstats.daily

The development version of the package MSstats.daily is the most recent and is available here. The versioning of the main package is updated several times a year, to synchronise with the Bioconductor release.

- MSstats.daily 2.1.6 (Last updated March 25, 2014, requires R3.0.2).
- source file (License : Artistic-2.0)
- zip file (License : Artistic-2.0)
- Changes since the previous version

****** Known issues and proposed solutions

- News about Msstats
- **MSstats.daily** is available : development version available
- Tutorials for different workflows (under ‘WORKFLOWS’)
- Example datasets with R-scripts
- Related publications



MSstats Shared privately

21 of 21 topics (2 unread) ★

Welcome to google group for MSstats!!

Here is the place to share your experience, difficulties, solution, and suggestion about R-package, MSstats!

Edit welcome message Clear welcome message

□ ☆ MSstats 2.1.3 released!!
By me - 9 posts - 84 views - updated Apr 18

□ ☆ MSstats external tool in Skyline v2.1.6 release!
By me - 5 posts - 26 views - updated May 8

□ ☆ MSstats in Skyline with different version of R
By tvaisa...@gmail.com - 3 posts - 16 views - updated May 6

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

Acknowledgements

Purdue University

- Prof. Olga Vitek
- Mike Cheng
- Veavi Chang
- Tim Clough
- Danni Yu
- Kyle Bemis
- April Harry
- Robert Ness

University of Washington

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

ETH Zürich

- Prof. Ruedi Aebersold
- Ruth Hüttenhain
- Silvia Surinova
- Eduard Sabido
- Olga Schubert
- Hannes Röst

Tutorial

Tutorial : ‘choi-shortCourse-MSstatsTutorial.pdf’

New version of MSstats

Poster 069:
Statistical Elimination of Spectral Features with Large Between-Run Variation Enhances Quantitative Protein-Level Conclusions in Experiments with Data-Independent Spectral Acquisition

Lin-Yang(Mike) Cheng[1], Ching-Yun Chang[1], Yansheng Liu[2], Hannes Rost[2], Meena Choi[1], Ruedi Aebersold[2], Olga Vitek[1] *;[1] Purdue University, West Lafayette, Indiana;[2] Department of Biology, ETH Zurich, Switzerland, Faculty of Science, University of Zurich, Zurich, Switzerland*