Statistical analysis with MSstats

Skyline User Meeting 2014 Baltimore

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

Outline

1. MSstats : statistical tool for quantitative MS proteomics
    1. Workflow of MSstats
    2. MSstats as an external tool

2. Integration of Skyline improves analysis workflow
    1. User interface
    2. Checking quality of features

3. Different workflows require different
    1. statistical models
    2. normalization

4. How to access MSstats

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.

Technological Innovation and Resources

Protein Significance Analysis in Selected Reaction Monitoring (SRM) Measurements*

Ching-Yun Chang‡, Paola Picotti§, Ruth Hüttenhain§‡‡, Viola Heinzelmann-Schwarz¶‖, Marko Jovanovic**, Ruedi Aebersold§‡‡§§, and Olga Vitek‡¶¶‖‖

research articles Journal of proteome research

Protein Quantification in Label-Free LC-MS Experiments

Timothy Clough,[†,#] Melissa Key,[†,#] Ilka Ott,[‡] Susanne Ragg,[§] Gunther Schadow,[‖] and Olga Vitek*[,†,⊥]

MSstats 2.0

Bioinformatics

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRE

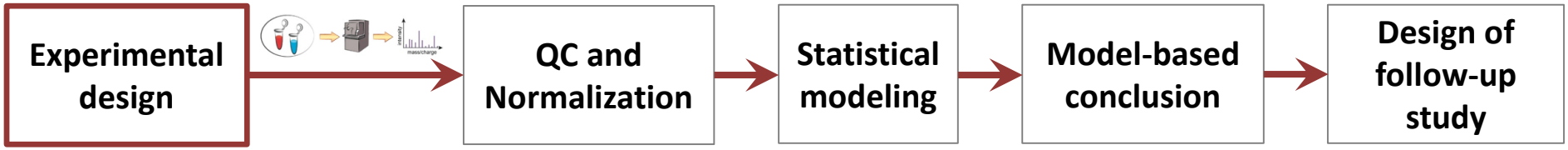
Meena Choi Change Password View/Change User Information CiteTrack Personal Alert

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]

MSstats workflow : Experimental design



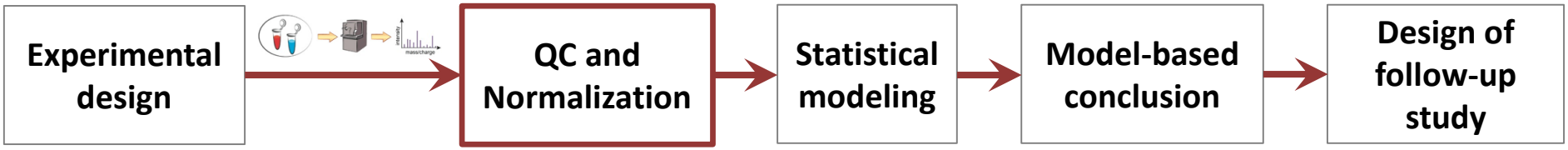
Type of experimental design

- 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

Input format

	Protein	Peptide	Precursor charge	Fragment	Product charge	Label	Condition	Subject	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

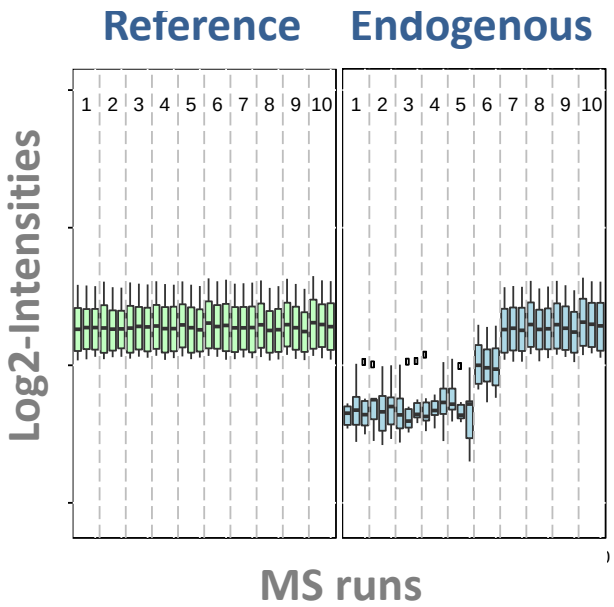
MSstats workflow : QC and normalization



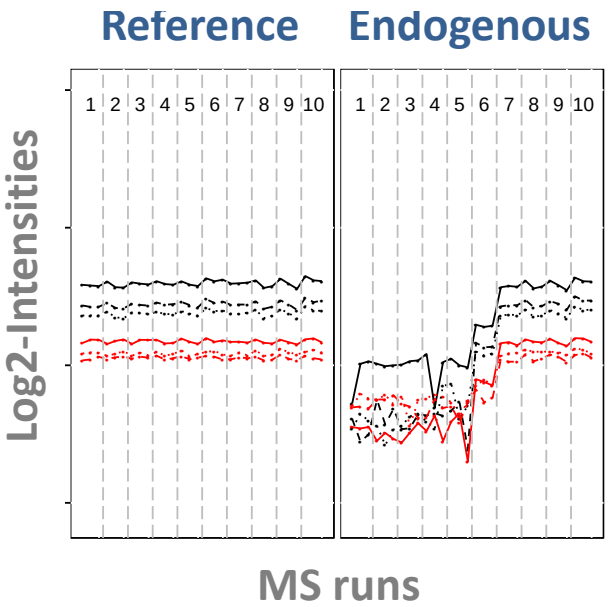
Data preparation

- Formatting
- Visualization : 2 plots
- Normalization : equalize medians, quantile, with standard protein

QC plot : All proteins



Profile plot : One protein



MSstats workflow : Statistical modeling

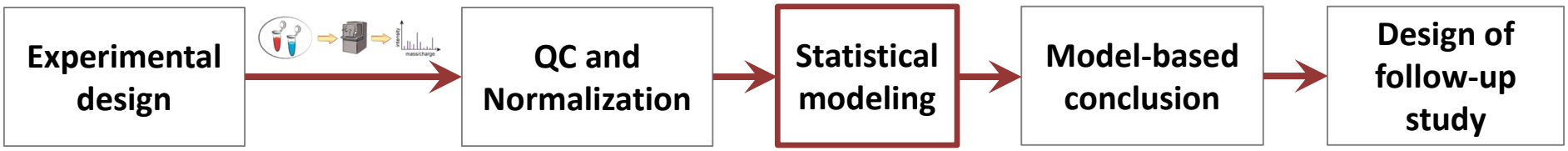
Experimental design → **QC and Normalization** → **Statistical modeling** → **Model-based conclusion** → **Design of follow-up study**

- Account for
  - different design of experiment
  - technical replicates
  - pattern of missing values
  - any special aspect

- 55 unique linear mixed models are used
  - 16 fixed effects models
  - 16 fixed effects models for unequal variance
  - 23 random effects models

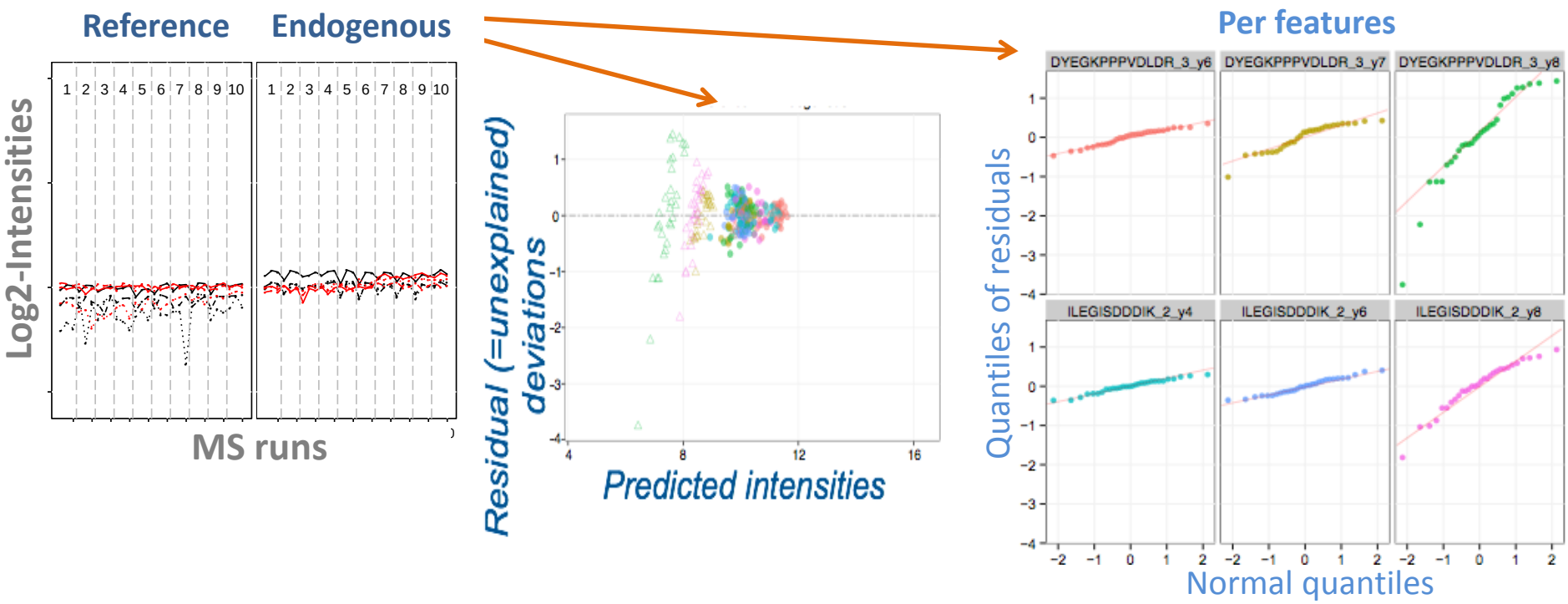
				Deviation from the reference due to								
	log(peak intensity)	=	Expected reference abundance	+ feature	+ condition or time	+ between-condition interference	+ biol. replicate	+ between-subject interference	+ run	+ between-run interference	+ Random meas. error
General case	**Group comparison:**										
	y_{ijkl}	$=$	μ_{1001}	$+ F_i$	$+ C_j$	$+ (F \times C)_{ij}$	$+ S(C)_k$	$+$	R_l	$+ (F \times R)_{il}$	$+ \varepsilon_{ijkl}$
	Time course:										
	y_{ijkl}	$=$	μ_{1001}	$+ F_i$	$+ T_j$	$+ (F \times T)_{ij}$	$+ S_k$	$+ (T \times S)_{jk}$	$+ R_l$	$+ (F \times R)_{il}$	$+ \varepsilon_{ijkl}$
	Paired design:										
	y_{ijkl}	$=$	μ_{1001}	$+ F_i$	$+ C_j$	$+ (F \times C)_{ij}$	$+ S_k$	$+ (C \times S)_{jk}$	$+ R_l$	$+ (F \times R)_{il}$	$+ \varepsilon_{ijkl}$
Single feature with technical replicates	**Group comparison:**										
	y_{1jkl}	$=$	μ_{1001}	$+$	C_j	$+$	$S(C)_k$	$+$	R_l	$+$	ε_{1jkl}
	Time course:										
	y_{1jkl}	$=$	μ_{1001}	$+$	T_j	$+$	S_k	$+ (T \times S)_{jk}$	$+ R_l$	$+$	ε_{1jkl}
	Paired design:										
	y_{1jkl}	$=$	μ_{1001}	$+$	C_j	$+$	S_k	$+ (C \times S)_{jk}$	$+ R_l$	$+$	ε_{1jkl}
Single feature, no technical replicates	**Group comparison:**										
	y_{1jkl}	$=$	μ_{1001}	$+$	C_j	$+$			R_l	$+$	ε_{1jkl}
	Time course:										
	y_{1jkl}	$=$	μ_{1001}	$+$	T_j	$+$	S_k	$+$	R_l	$+$	ε_{1jkl}
	Paired design:										
	y_{1jkl}	$=$	μ_{1001}	$+$	C_j	$+$	S_k	$+$	R_l	$+$	ε_{1jkl}

MSstats workflow : Statistical modeling

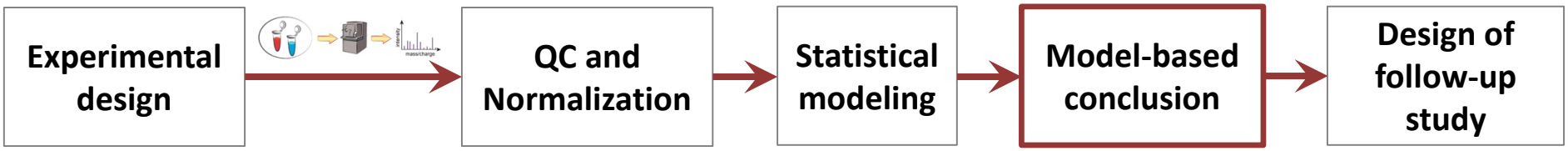


Verify the assumptions

Deviations from independence or from constant variance are often mistaken for deviations from Normality



MSstats workflow : Model-based conclusion

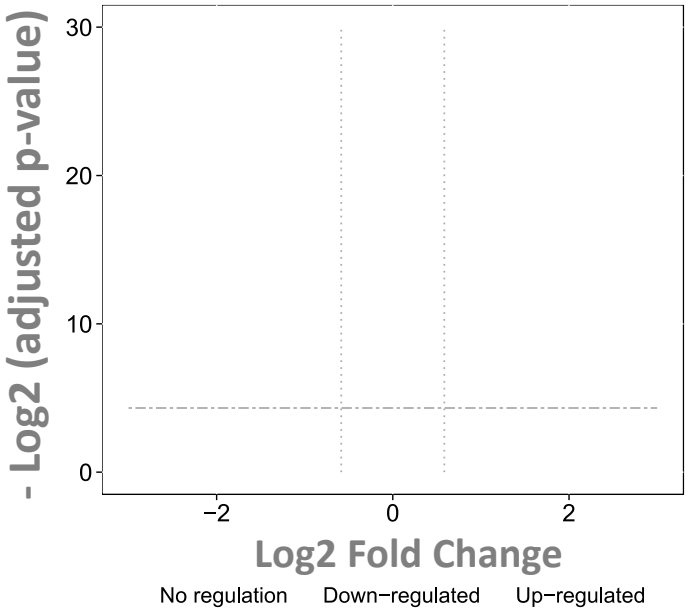


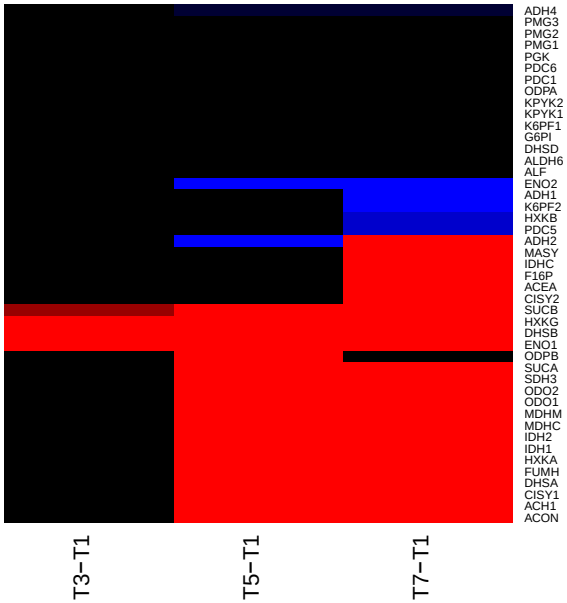
Model-based group comparison

- Quantify the uncertainty
- Adjust p-values to control FDR

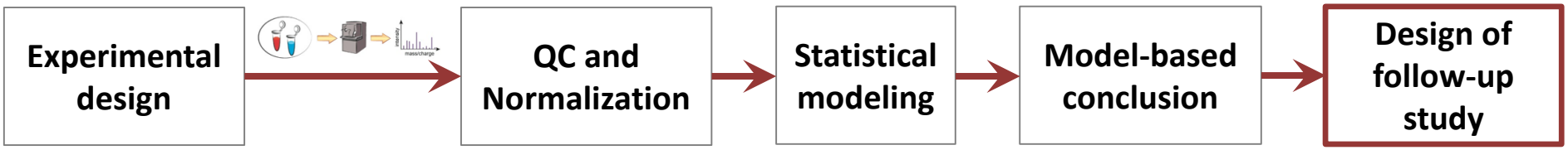
Relative protein quantification

- by sample
- by condition



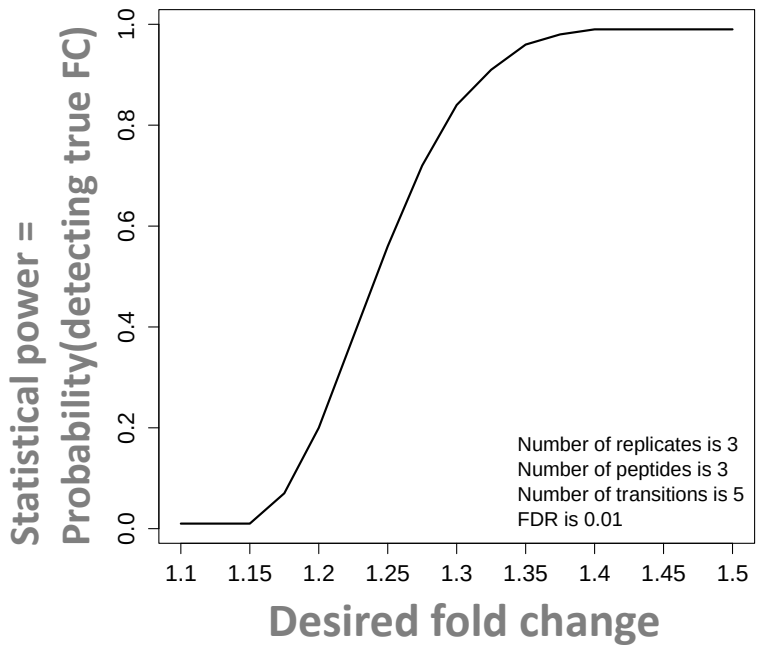


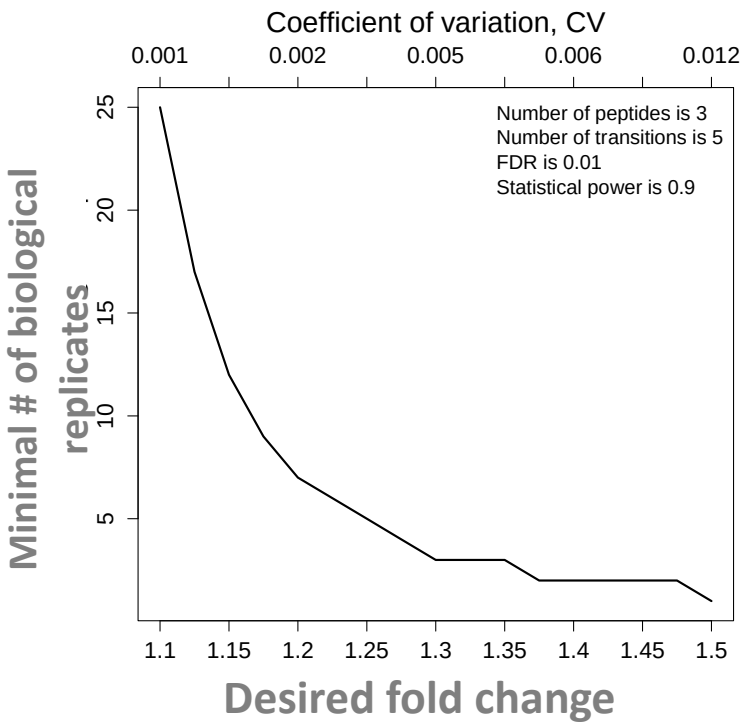
MSstats workflow : Design of follow-up studies



Use the dataset to improve :

- subject selection : matching
- resource allocation : blocking
- calculation of sample size
- use information from current study





Outline

1. MSstats : statistical tool for quantitative MS proteomics
    1. Workflow of MSstats
    2. MSstats as an external tool

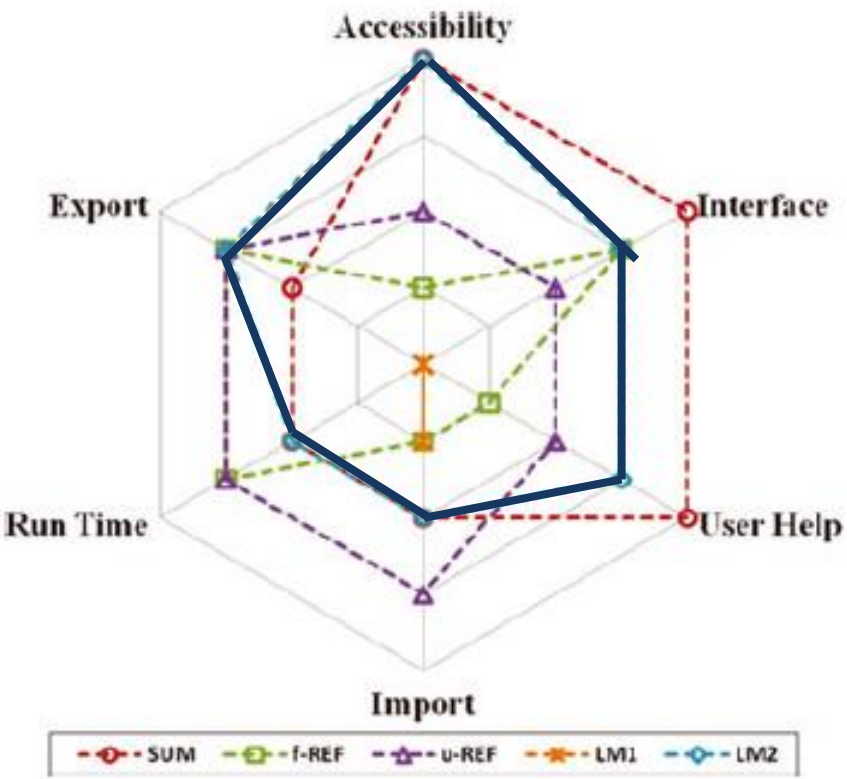
2. Integration of Skyline improves analysis workflow
    1. User interface
    2. Checking quality of features

3. Different workflows require different
    1. statistical models
    2. normalization

4. How to access MSstats

Motivation : many users are not familiar with R

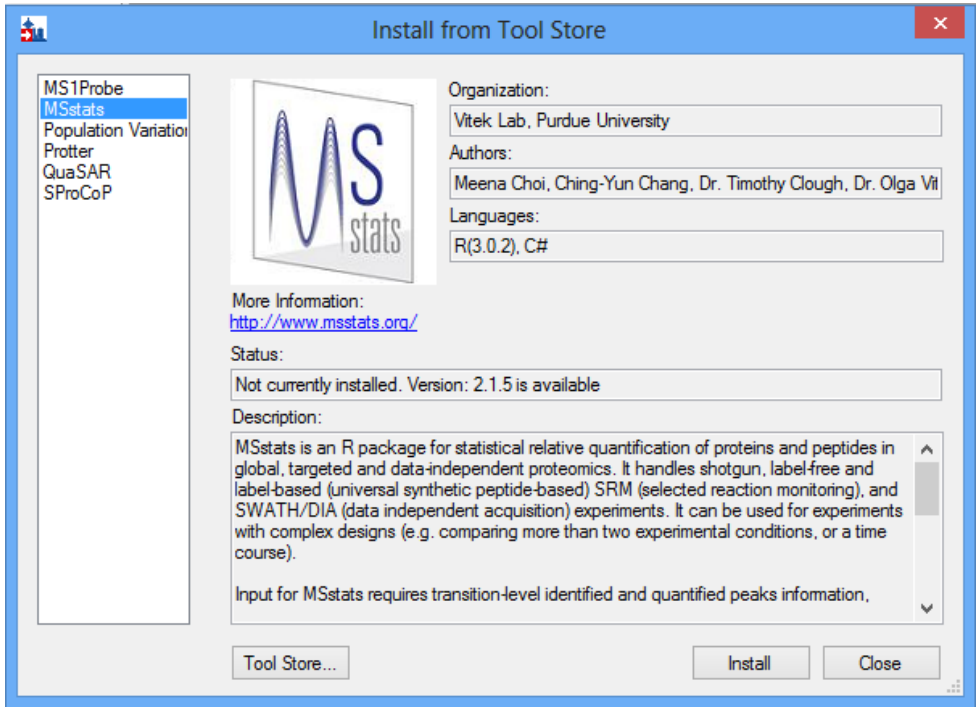
- Many requests from users
- Improve usability issue
  - User interface from installation to analysis
  - User help



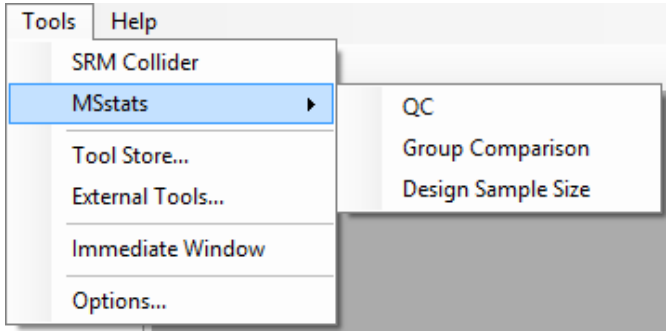
M.M. Matzke et al. Proteomics 2013

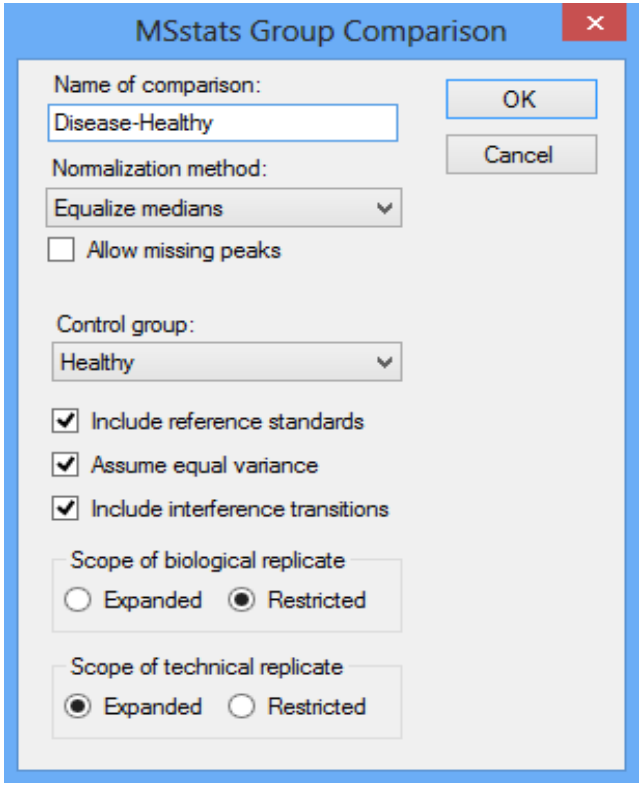
MSstats as an external tool

1. Automatic installation in skyline



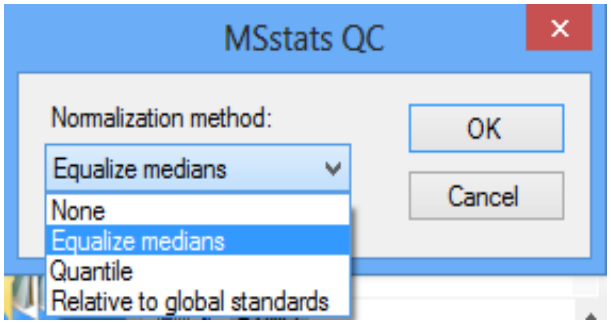
2. New option 'MSstats' under Tools





3. MSstats functions

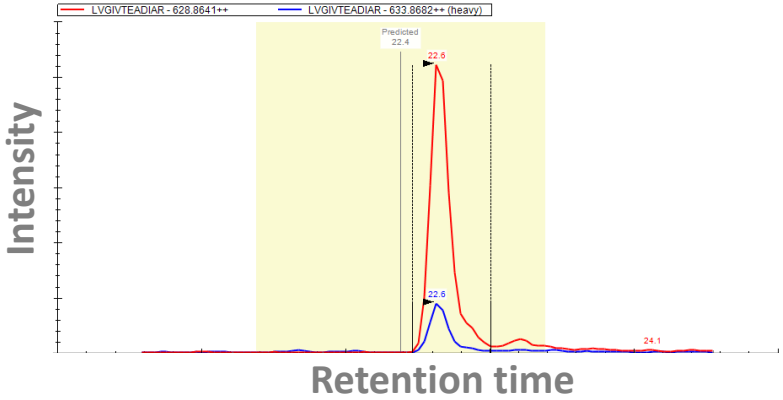
- GUI
- Output and results will be saved automatically



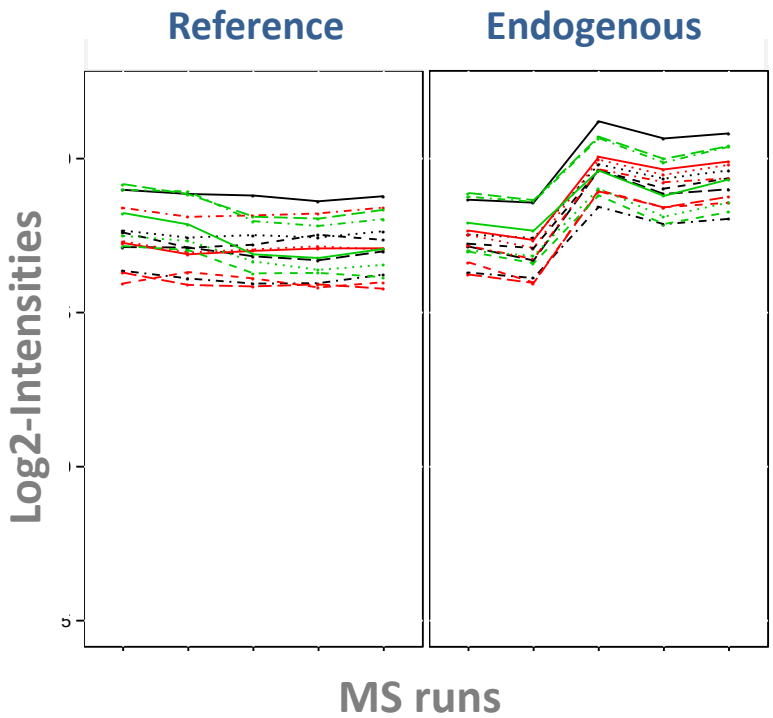
Skyline and MSstats provide complementary QC

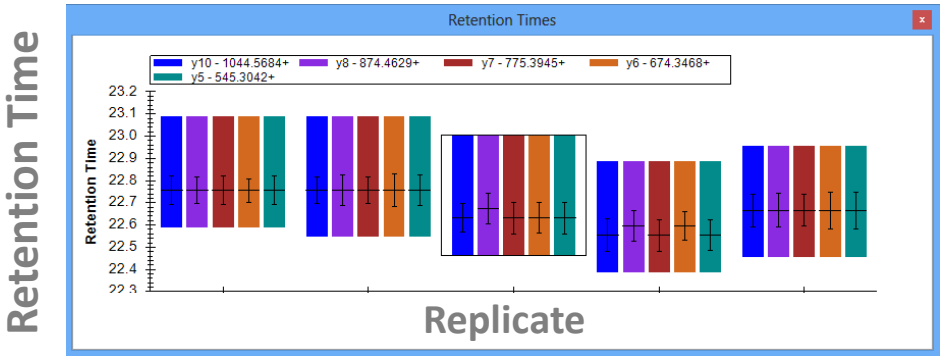
- Example dataset in 'Advanced peak picking models' Skyline tutorial.
- SRM with stable isotope labeled reference peptides

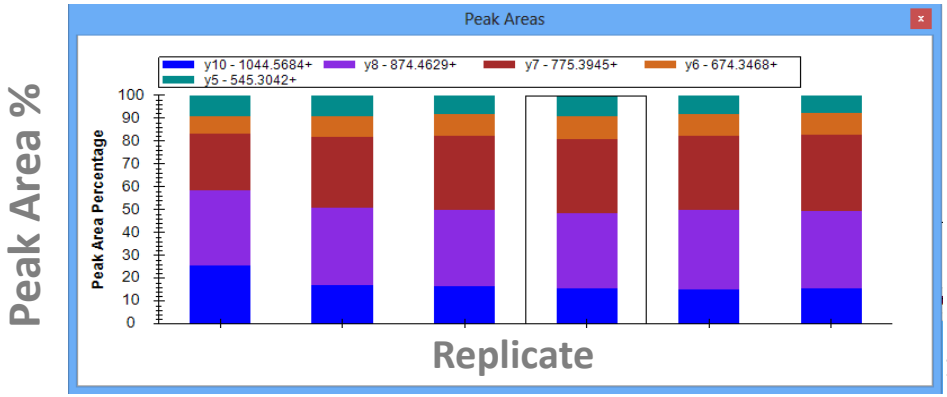
Visualization in Skyline



Visualization in MSstats

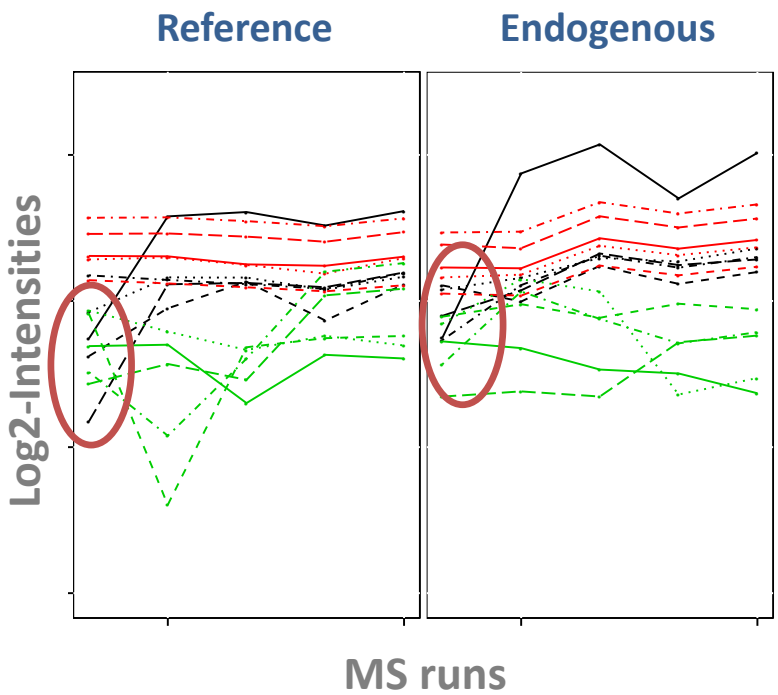




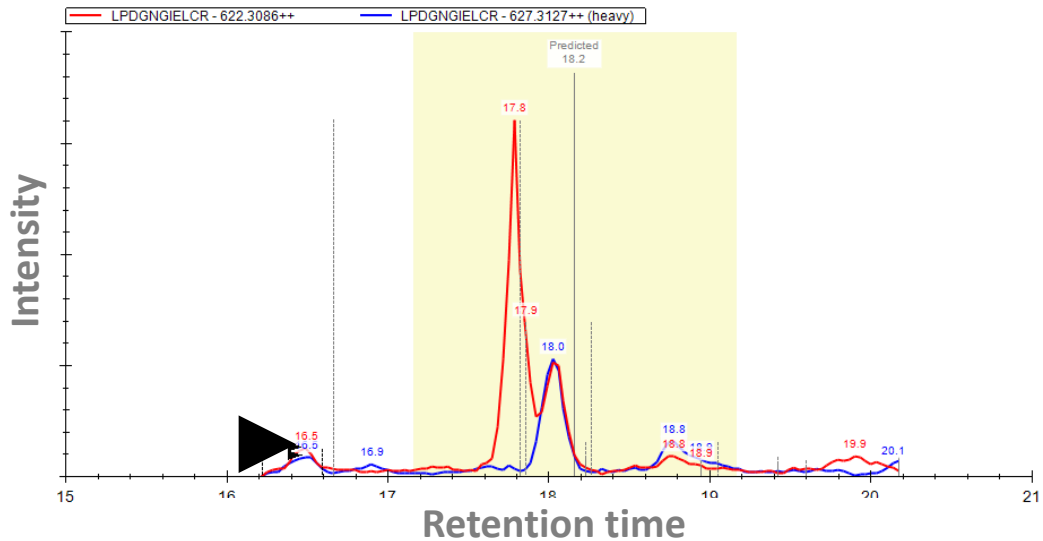


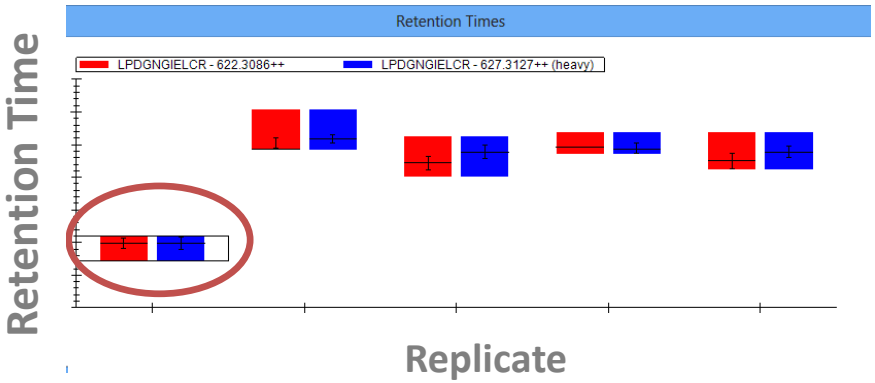
Between run variation points to poor quality peaks

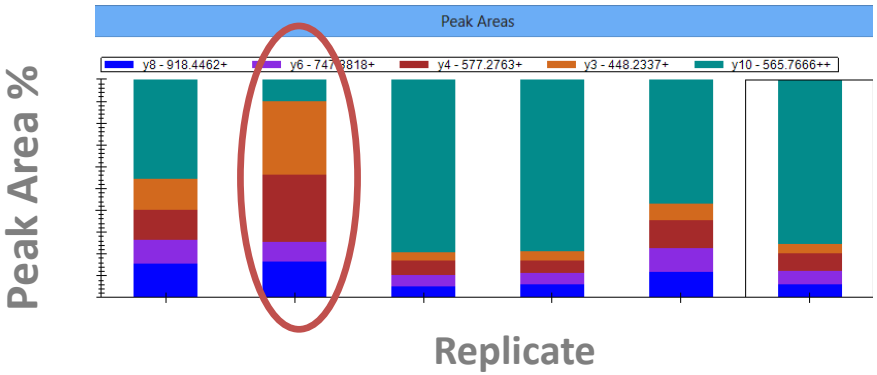
Profile plot from MSstats



Chromatography from Skyline

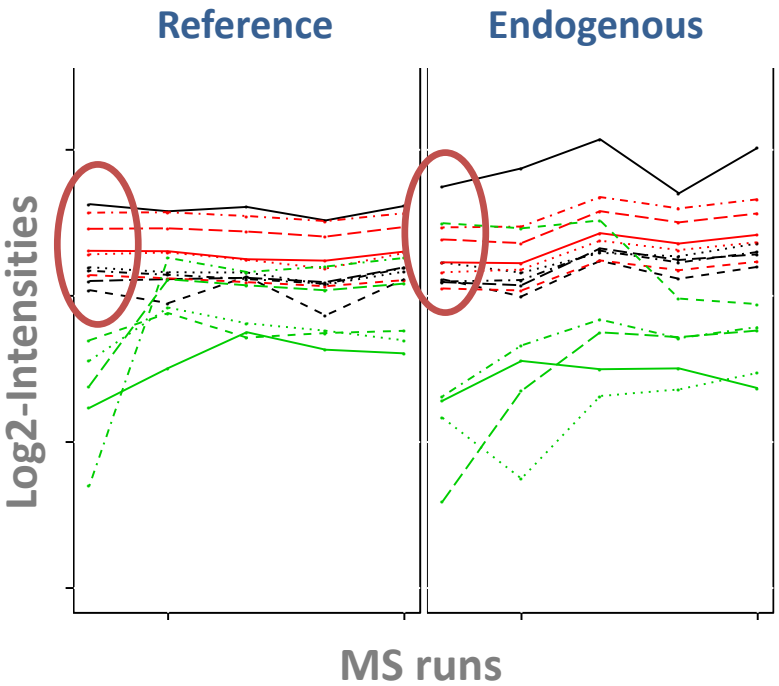




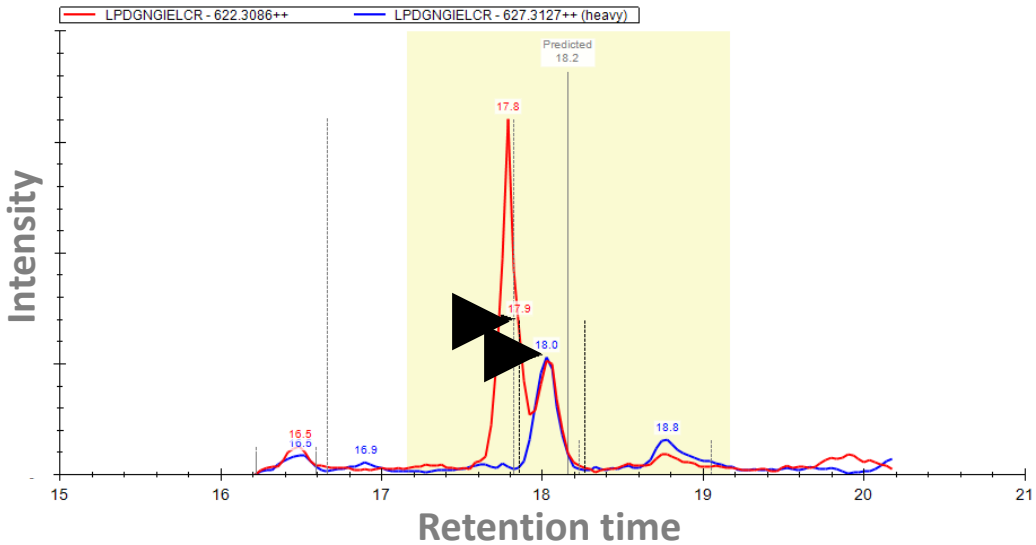


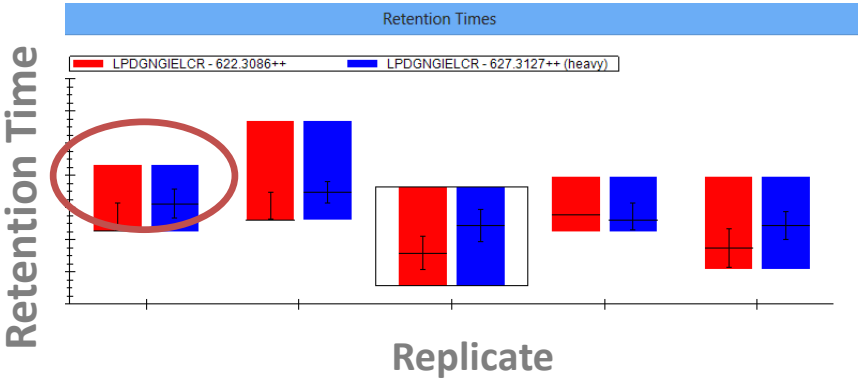
Refinement for picking peaks improves the quality

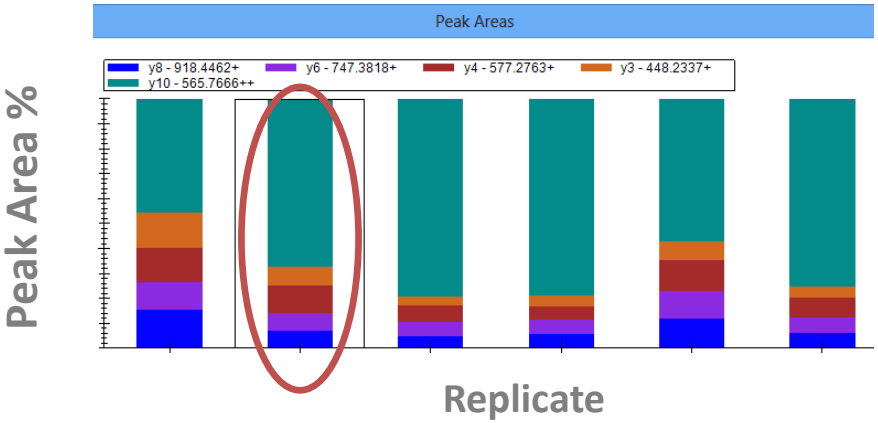
Profile plot from MSstats



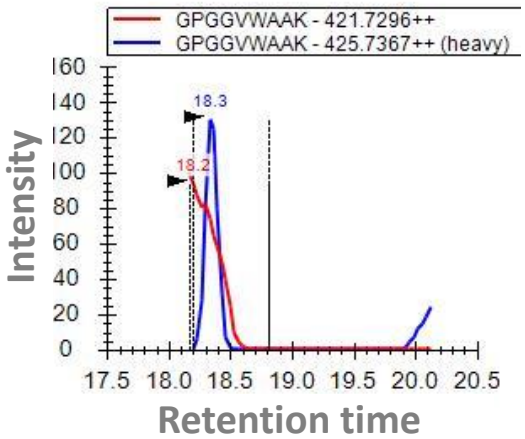
Chromatography from Skyline





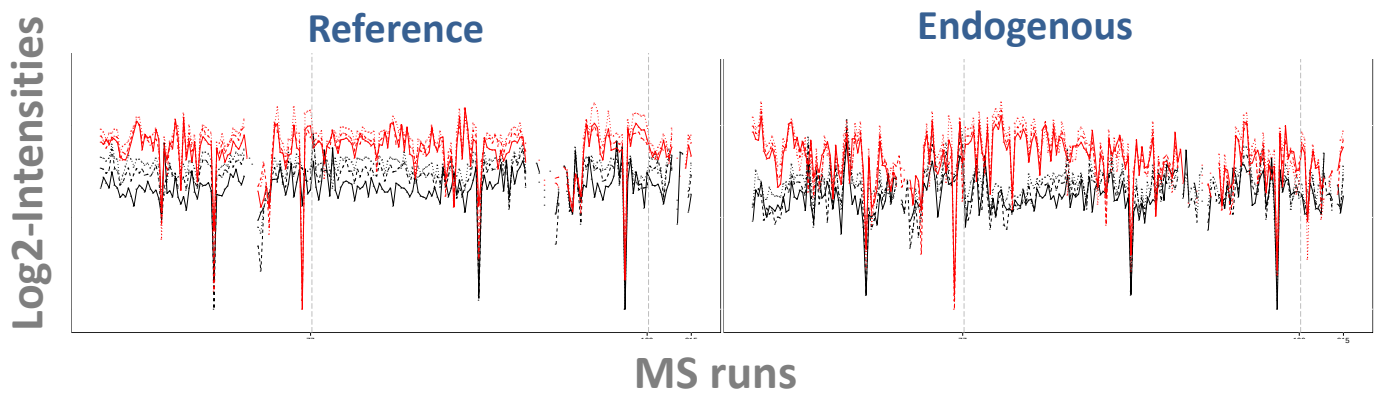


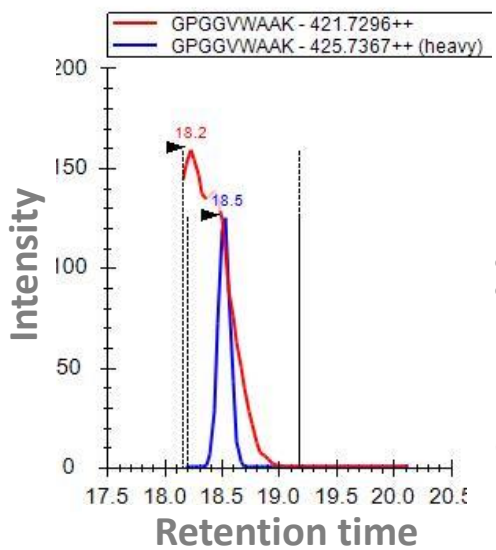
Truncated peaks also introduce between run variation



with truncated peaks :

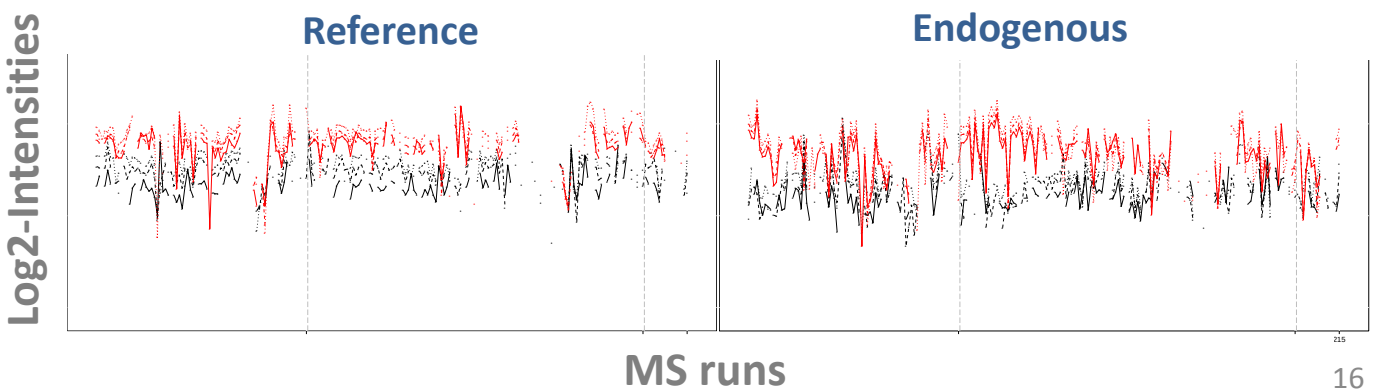
log2FC=-0.342, adjust p-value=0.0277





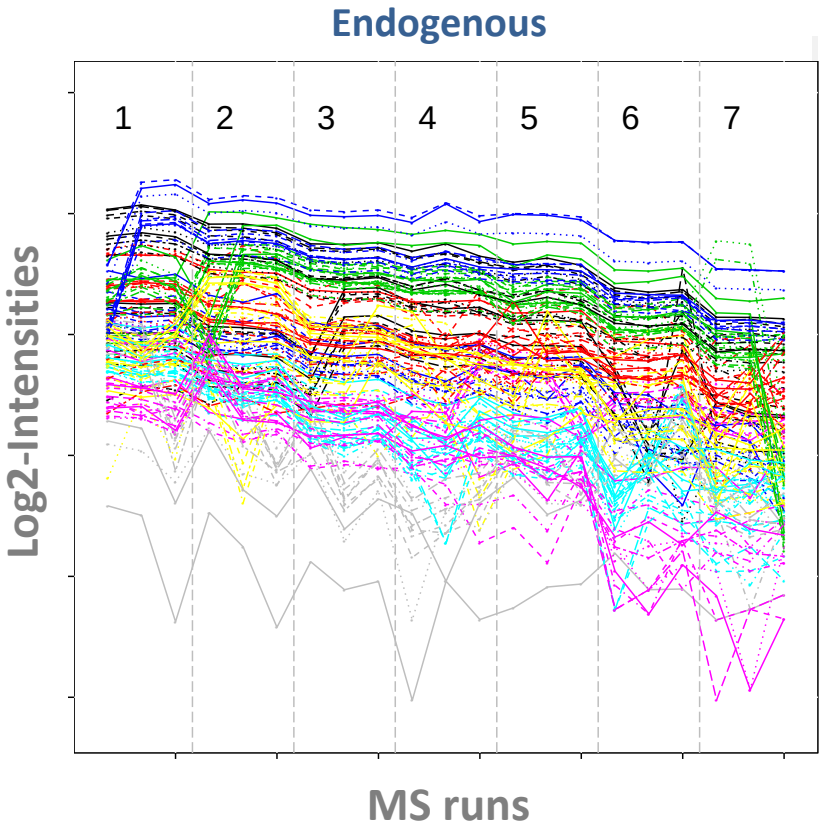
without truncated peaks :

log2FC=-0.155, adjust p-value=0.4999

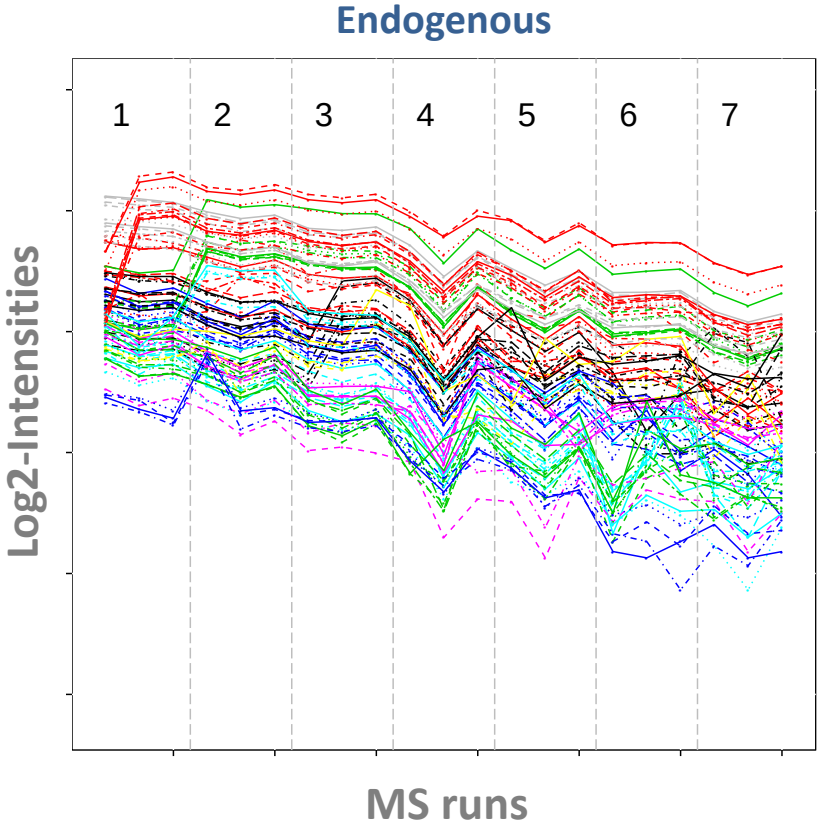


Feature selection is available in MSstats

Original spike-in DIA data

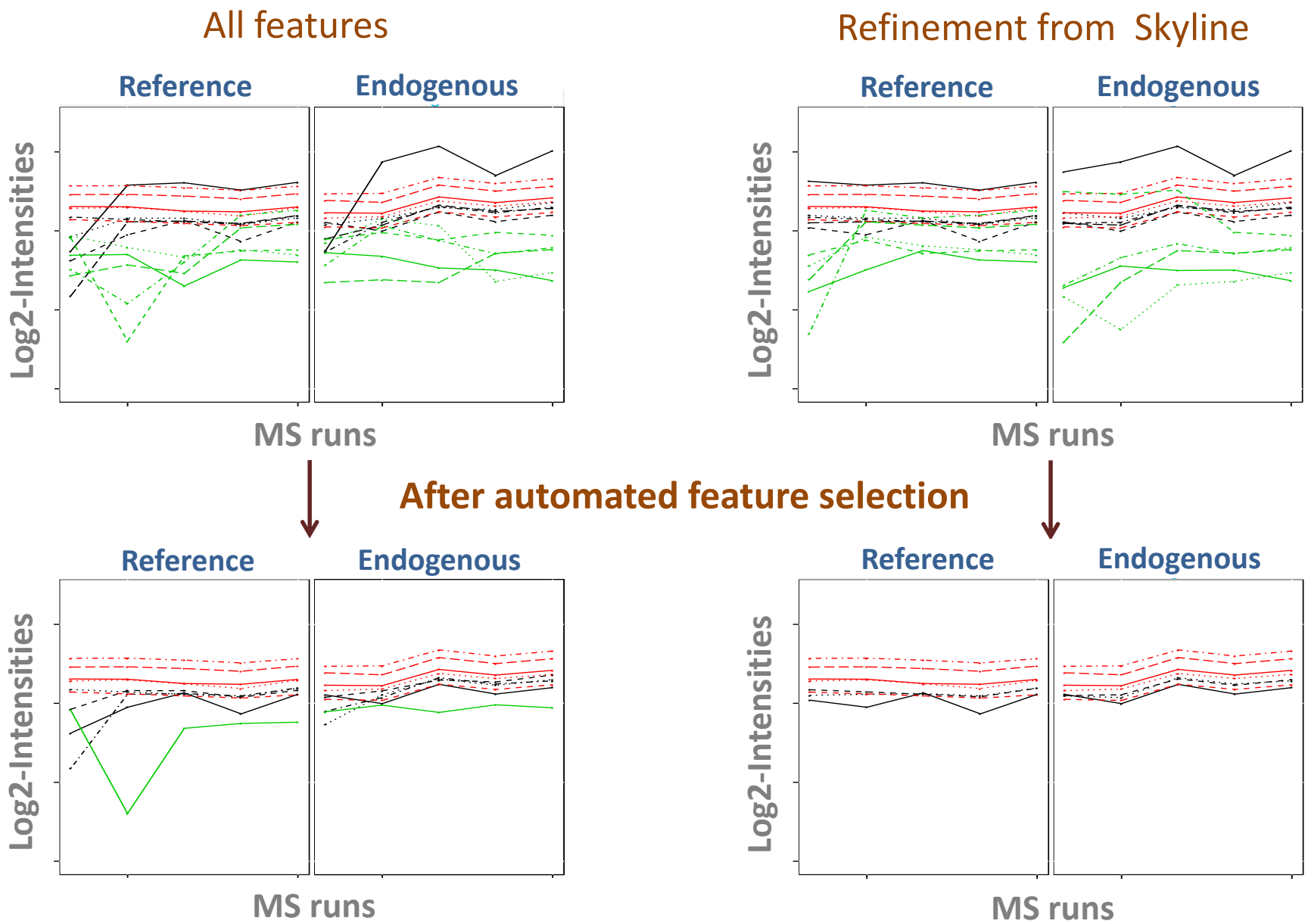


After feature selection,
informative features are selected.



Oral session : MOG 3:50pm, Statistical Elimination of Spectral Features with Large Between-Run Variation Enhances Quantitative Protein-Level Conclusions in Experiments with Data-Independent Spectral Acquisition (Lin-Yang Cheng)

New in MSstats (beta) : automated feature selection



Outline

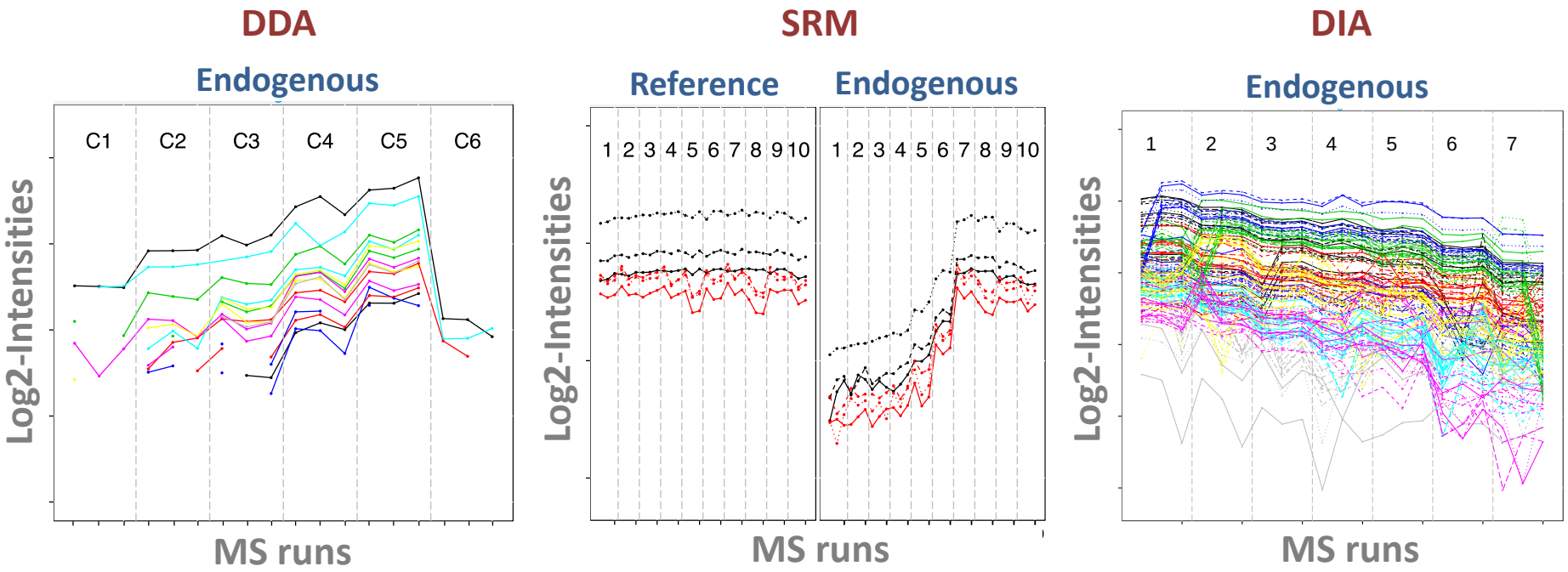
1. MSstats : statistical tool for quantitative MS proteomics
    1. Workflow of MSstats
    2. MSstats as an external tool

2. Integration of Skyline improves analysis workflow
    1. User interface
    2. Checking quality of features

3. Different workflows require different
    1. statistical models
    2. normalization

4. How to access MSstats

Different workflows require different statistical models



	DDA	SRM	DIA
Model options	Interference=FALSE equalFeatureVar=FALSE	Interference=TRUE equalFeatureVar=TRUE	Interference=FALSE equalFeatureVar=FALSE FeatureSelection=TRUE
Normalization	Normalization =globalStandard	Normalization =equalMedians or =globalStandard	Normalization =Quantile or =globalStandard

Internal standards are always good for quality control

Outline

1. MSstats : statistical tool for quantitative MS proteomics
    1. Workflow of MSstats
    2. MSstats as an external tool

2. Integration of Skyline improves analysis workflow
    1. User interface
    2. Checking quality of features

3. Different workflows require different
    1. statistical models
    2. normalization

4. How to access MSstats

External tool in Skyline

MSstats

MSstats

Version 2.1.6 [View All]

Uploaded Mar 28, 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: 1543

Documentation

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

- **Downloads from Feb 2014 : 1543**
- From MSstats external tool webpage or ‘Tool store’
- Automatic installations for all related software and packages
- One-click analysis

Bioconductor

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help

Home » Bioconductor 3.0 » Software Packages » MSstats

MSstats

- MSstats v2.3.0 under BioC 3.0
- Unique downloads from Oct 2013 : 1073

Protein Significance Analysis in DDA, SRM and DIA for Label-free or Label-based Proteomics Experiments

Bioconductor version: Development (3.0)

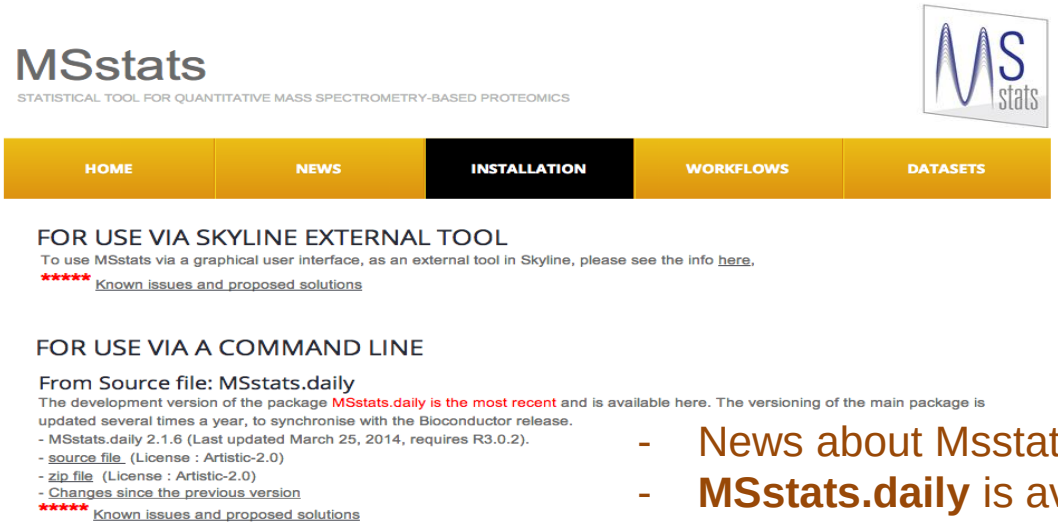
A set of tools for statistical relative protein significance analysis in DDA, SRM and DIA experiments.

Take advantage of options and modify the data easily.

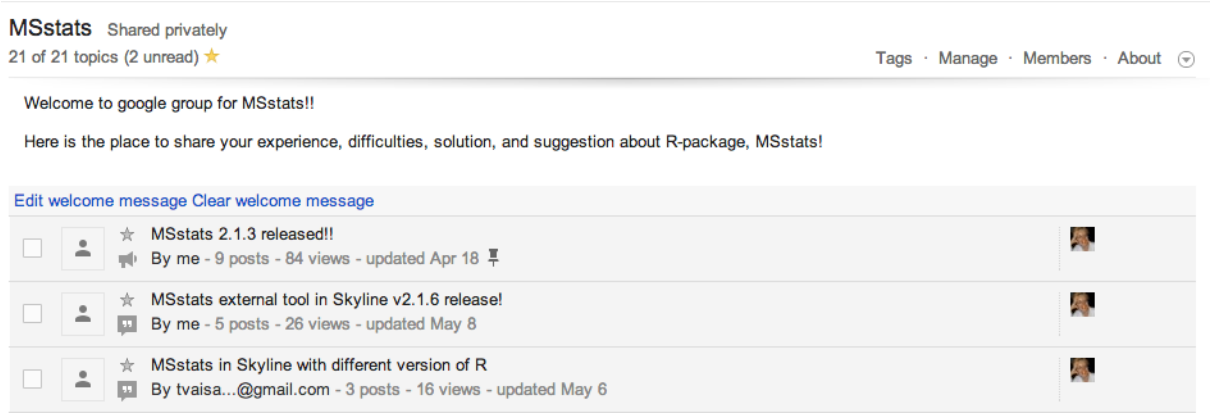
- New functionality : Quantification for sample, feature selection
- Detailed options for all plots
- More options for design sample size function

```
dataProcessPlots(data=data,type=type,featureName="Transition",
        ylimUp=FALSE,ylimDown=FALSE,scale=FALSE,interval="CI",
        x.axis.size=10,y.axis.size=10,text.size=4,text.angle=0,legend.size=7,dot.size=3,
        width=10, height=10, which.Protein="all",address="")
```

msstats.org and MSstats google group



- News about Msstats
- **MSstats.daily** is available : development version available
- 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

Biomedical investigations using MSstats

- Ruth Hüttenhain et al. "Reproducible Quantification of Cancer-Associated Proteins in Body Fluids Using Targeted Proteomics". Sci Transl Med 4, 142ra94 (2012);
- Ruth Hüttenhain et al. "N-Glycoprotein SRMAtlas: A resource of mass-spectrometric assays for N-glycosites enabling consistent and multiplexed protein quantification for clinical applications". MCP
- Surinova et al. "Automated SRM data analysis workflow for large scale proteomic studies." Nature Methods
- Sabido et al., "Targeted proteomics of the eicosanoid biosynthetic pathway completes an integrated genomics-proteomics-metabolomics picture of cellular metabolism." MCP
- Sabido et al., "Targeted proteomics analysis of the insulin-pathway and central metabolism reveals substantial, strain-specific proteome changes after sustained high-fat diet in C57B6/J and 129Sv mice." Molecular Systems Biology

Short courses to learn more about MSstats

- SRM course, Zurich (July 2013, February 2014)
- Targeted Quantitative Proteomics Course, Seattle (April 2014)
- US HUPO (Baltimore 2013, Seattle 2014)

- 8th European Summer School in Proteomics, Brixen, Italy (August 2014)
- EMBO practice course on targeted proteomics, Barcelona, Spain (September 2014)
- Workshop on Targeted Proteomics, Korea (October 2014)
- Workshop on Targeted Proteomics, India (December 2014)

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
- Olga Schubert
- Hannes Röst
- Ruth Hüttenhain
- Silvia Surinova
- Eduard Sabido
- Matondo Mariette

University of Zürich

- Meliana Riwanto