# May 11 2021, 12:45pm-2:15pm Hands-on, Meena Choi and Brendan MacLean: Analysis of data in Skyline Part 1 and Part 2 in MSstats, and comparison of the results

**Speaker**
Brendan MacLean - email: brendanx .at. uw.edu
Meena Choi

This session will focus on data-independent acquisition (DIA) and will use experimental datasets processed in Skyline Part 1 from the Bruderer, MCP 2015 paper, "*Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues*" and Skyline Part 2 from Selevsek, MCP 2015 paper "*Reproducible and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-mass spectrometry*"; Ideally participants have attended the May 4 session and May 5 session presented by Brendan. We will perform the downstream statistical analysis with MSstats, open-source R package and visualization in R, and evaluate the impact of various choices made at the data processing stage on the downstream conclusions.

Presentation Slides - MacLean

**Preparation**
In order to get the most out of this session, it is recommended that you:

1) Review the May 4 session
2) Review the May 5 session

**Materials**
If you would like to follow the tutorial hands-on during this workshop, please:

1) Install R Studio
2) Install R 4.0.3 (or R 4.0.5)
3) Participants that attended the May 4 & 5 sessions should already have the above, plus Skyline-daily and Skyline Batch-daily
4) Install MSstats, download R script for installing the required packages

If you were unable to complete the Skyline Batch runs from the May 4 & 5 sessions or new for week 2:

1) Download the resulting reports for MSstats analysis
    a) Place the ZIP file in a folder named NEU21_Skyline
    (either create it or use the one you have already from the prior sessions)
    b) Unzip the ZIP file directly into this folder
2) Or Download the results reports+R scripts+outputs (MSstats_tutorial_Selevsek.zip) for MSstats analysis

More information is at https://computationalproteomics.khoury.northeastern.edu/

---

**Q&A**

How would you compare MSSTats with Perseus for TMT statistics?

For statistical comparison, MSstats fits proper linear mixed effects models based on the underlying experimental designs, instead of fitting one single model for all the data like Perseus.

Perseus will work with protein-level data, MSstats/MSstatsTMT will start to work with feature-level data(PSM for TMT). Protein-level summarization in MSstats/MSstatsTMT is different. Perseus might requires that user define the stat model. MSstats/MSstatsTMT automatically fits the best linear model for experimental design.

Why is the distribution shifted so much to the left between both for DIA?

Absolute intensity is not important here. The tools are calculating different numbers as "intensity". We saw this (and specifically noramalized for it) in Navarro, Nature Biotech. 2016. What is important and true is that the numbers are strongly linearly correlated. The numbers are not _calibrated_ and not strictly quantitative on their own.
Assuming you mean why are the numbers for Spectronaut smaller than for Skyline.
e.g. Spectronaut may be using apex peak height in the MS/MS spectrum (minus background), while Skyline is using summed intensity.

Hi, to data processing, do you plan a direct DIA-NN support?

Yes we have. after the next release

Maybe a naive question, but can the workflow of Msstats be 'modified'/'implemented'/'utilized' to anaylse RNASeq (for example differential gene exprerssion) data?

MSstats is design to reflect the structure of mass spectrometry data

Hi does MSSTAT Has a vignette in the R package? name of that? I used R package called proteous. I have never used msstats

https://msstats.org/wp-content/uploads/2020/02/MSstats_v3.18.1_manual_2020Feb26-v2.pdf

Is MSstatPTM going to be added to galaxy?

At the moment we are not in the process of adding it, but it may be added in the future once v2.0 is released.
MSstatsTMT is a recent addition, so there's a good chance

How do you know how many biological replicates (n's per group) are needed?

We use variance component and caldulate it . I will briefly talk about it later.

Is it possible to use MSstats for SILAC data analysis? Thanks!

No, in experimental design perspective, MSstatsTMT is better for SILAC.

What is the best way to do blocking for different biological replicates?

This question has been answered live

why Do you use the log2 to normalize the data?

This question has been answered live

If a Sub-cellular proteome analysed in DIA and if we also have whole cell lysate, Spiking iRT peptides could help as being explained in this examples?

Yes, some controls could help.

How many features can I consider statistical important to take in consideration (like a filter) when I need to check what are my up or down regulated proteins

There is no fixed number of features required. Depending on the acquisition methods and biological study, it might be different. For DIA, which has quite many features, top 100 would be enough in my experience.

Does MSstats input require peptide/protein identification or could be statistical analysis done on ions with names assigned in more generic manner (for instance based on just peak ID: 9, 123, 2245, etc)?

Generally it is best to include fully identified peptides/proteins, but some processing tools include proteins which could not be uniquely identified (ie they include multiple potential proteins). In this case you can chose whether to use only uniquely identified proteins or not.

Can you explain the advantage of the MSstats linear model over the limma package? Given that limma uses Bayesian theorem, is it true that it can handle data with higher variability?

This question has been answered live

are you reporting the fold change of the pvalue in the heatmap pairwise comparison?

Currently, no, color in heatmap is based on adj p-value

Could you please explain why 10% (0.1) FDR cutoff is used? Also is it different from Bonferroni-Hochberg correction? Thanks!

MSstats uses BH correction, MSstatsTMT can use other corrections as well.
The cutoff choice is made after the analysis with MSstats

In our research we focus on endogenous peptides and usually do statistical analysis prior to peptide/protein identification to decrease search space thus I was wondering if MSstats could be utilized in that case when ions don't yet have assigned peptide IDs

MSstats assumes that we observe a set of features across runs. Most of the analysis is done per protein, so the measurements need to be assigned to specific proteins and be identifiable across runs
The specific identity is not that important, but we need to be able to match them to proteins and know that it's the same entity in different runs

a slightly vague question - instead of t-tests, can we do something like regression analysis, especially when we do not have categories like disease/control but samples from a general population

Yes, you can use regression and even multiple regression. You just need to chose the statistical variables about your samples that you will perform regression on, e.g. Age, Sex, etc. T-test is essentially a simplified regression with 2 conditions.

and when our data do not have a normal distribution?

The mean will still likely be normally distributed (central limit theorem). This is an important difference. The distribution of the data is not required to be normal. This can be done with true/false data.
Let me add that MSstats has a visualization function to check the normality assumption. Also, logarithmic transformation and robust methods used make the impact of outliers etc smaller

In a cohort study, is there a way to know with MSstat how many biological replicate do we need regarding the statistical power of the statistical that we use for ou study ?

Yes, MSstats provides a designSampleSize() function to estimate sample size for future experiments

Can MSStats be used for DDA data processed in Proteome Discoverer?

yes you can

if you have several groups (five) to compare and we used aN ANOVA test, I got several proteins that are modulated. Whar kind of test do i need to use to see in which conditon the proteins are modulated?

This question has been answered live

About multiple comparisons; waht if I have a prior assumption about something? For wxample, I'm doing proteomics on KO vs WT. The statstical testing to ask whether my KO target is really Knocked out should be different than statistics for the "discoveries", which are the proteins affected by my manipulation.

This question has been answered live

When you were discussing normalcy and independence... can you discuss more how you ensure independence?

This question has been answered live

For the FDR, Perseus uses permutation based FDR, what is the difference between perseus method and MSstats method?

Permutation based method in Perseus controls FWER. It also has option to control FDR via B-H correction, which is what MSstats does

If the protein is not detected in one condition, how to calculate the p-value in this case?

The protein needs to be detected in both conditions in order to make a comparison

if the data are from multiple age groups, can MSstats adjust for age?

No, MSstats can't adjust the numeric variables.

Could you comment on power i.e number of minimum replicates needed to consider 2 Fold and number of replicates for 1.6 Fold (1.6 fold may be relevant in case of PTMs)

We will talk shortly during msstats handson

Assuming 4 BioReps each measured once, can it make sense to group proteins by the number of occurences (2, 3, 4) or better keep all the data together? Thx for a great session so far!

not much makes sense for me, groupoing proteins by the number of occurences. There are some testing method based on the counting of measurement. This make sense more.

How important are technical replicates in a disease/heathy experiment. Is there some intuition around when it is better to add more biological replicates vs technical replicates?

It is always better to add more biological replicates if you are time restricted. Your statistics will always benefit more from adding biological replicates than adding technical replicates. Technical replicates will give you added visibility into your technical variance and possibly expose a confounding variable, but it will not give you as much added power as adding biological replicates.

Could you not only calculate the means on fully measured value pairs? Why would imputation be able to reveal information when it essentially "cheats the test" by looking at other data?

This question has been answered live

Sorry could you please explaine the difference between perseus and MSstat imputation?
At the moment we primarly use Perseus for all analysis.

MSstats imputation is performed on feature-level data, before summarizing to protein-level. As my understanding, Perseus imputation is for protein-level.

What is your opinion about imputation by assigning small random values from the normal distribution?

We tested around 40 different datasets across different acquisitions and tools. The conclusion is the AFT model was better than random value from the normal distribution. Actually, AFT model has normal distribution assumptions. But different parameter though..

CV calculation has to be done before or after normalization?

After normalization. In the session last week on Selevsek data, we chose median normalization, which happens before the CV calculation.

Why T90 # of proteins is less than T60 #?

This question has been answered live

With MQ report would you start with the modification specific peptide?

Since MSstats is designed for global protein data analysis, it will takes all the identified and quantified peptides as input. Some peptides may be filtered due to low quality. But you can select modification specific peptides (filtering the modification column in evidence file) by yourself before running MSstats.

filtering to keep lowCV is basically the same as eliminating proteins with too many missing values?

This question has been answered live

as was anwer before MSstats new release will handle DIA-NN data, when actually new release is planned? thanks

new release, MSstats v4 will be available around at the end of May or early June. We expect DIA-NN converter around June.

when I am doing DIA analysis with skyline the decoys histogram apears overlapped with the sample histogram, why is it?

Best to submit a detailed support request to the Skyline support board (Help > Support in Skyline).

Would this line change to format MQ input-how?

This question has been answered live

can we get the full citation or doi for the Tsai publication?

https://pubmed.ncbi.nlm.nih.gov/32234965/

What's the current capabilities of MSstats with PTMs? For example can we use open searching (MSFragger, pFind, etc.) and then MSstats for downstream validation?

MSstatsPTM currently supports PTM output from MaxQuant.
For PTM output from other tools, you may need to do data data manipulation by yourself to generate the requried input format

Can you repeat witch files from MQ to start with I missed it

evidence.txt and proteinGroups.txt

How to select a subgroup of peptide features quantitatively that represent different regulational changes due to the isoform or  truncated forms of identified protein groups?

MSstats doesn't support the subgrouping based on the inference. lean on the pre-defined information

How does MSstats deal with peptide level outliers?

Would outlier value be removed for calculate protein abundance?

With default, we use TMP, robust summarization which is helpful for outliers. With feature selection option, we detect and remove outliers.

How do you impute for PTMs? or better not to?

We can impute for PTMs in the same way (ATF) as the base MSstats. It can lead to some inconsistent imputations due to low number of features so it needs to be analyzed on a case by case basis.

So the opinion here is that imputation is fine for summary-level statistics [Brendan's "Google Earth View" :)], but how does this affect protein/peptide-level conclusions? [The "Street View Level"]

This question has been answered live

can you comment on the reasons to use or not use permutation based FDR versus BH for multiple corrections? If this was already discussed, and I missed it, I'll go back and listen.

Based on Perseus tutorial, permutation method controls FWER (different criterion)

Will there be any more information on how to design your experiments in terms of number of replicates etc.?

There is the function to calculate sample size in MSstats. will go through it shortly.

What is the synthax for other normalization methods?

use "?dataProcess" to see the parameter options!

I installed MSstats, it went to all steps but when I search it in the "Packages" it is not found, can you help?

https://drive.google.com/file/d/10lkz9kwKc8mWoqww-wk_yVam4Itub7Eq/view Could you download this  r script?

Is there a way to quickly find one protein of interest?

there is a which.Protein parameter where you can list the proteins you want to plot

Why does the FDR inflate?

This question has been answered live

I have install MSstats in my mac. however i have the message : ackage ‘MSstats’ was built under R version 4.0.4 My R version is : 4.0.3. What do I need to do?

R version 4.0.3 is fine. You don’t need to do anything

Can the spectra library (DDA) cover all of detectable ions in DIA? I means maybe there could be some of missing there, right?

This question has been answered live

Is MSStats able to do batch corrections?

MSstats provides different normalization options to correct the technical bias between runs. But if you want to correct batch effects such as age or gender, MSstats is unable to do that

sorry I can not find this scrip, do we have it?

https://drive.google.com/file/d/1CBx8tSTquH7EjSHpSLbBYPpYUVkKqSut/view
download the zipped folder and unzip it

If you input two txt files for MQ- what is the right synthax to read both?
You need to read them into separate variables. One for evidence, one for proteinGroups (and one for annotation)

https://msstats.org/wp-content/uploads/2020/02/MSstats_v3.18.1_manual_2020Feb26-v2.pdf
There is the section for MaxQuant converter including the synthax

So suppose I randomize all my subjects but i have to ran in several days. in metabolomics we usually have a qc (pool of all samples) and we ran each 50 sample. During the processing we apply  QC-based signal correction (either QC based random forest signal correction (QC-RFSC) or QC LOESS signal correction). How exactly MSStat does this corrections?

This question has been answered live

Also what imputation type MSStat does? RF, kNN, SVD, Median ?

Imputation is based on accelerated failure time model
Accelated failure model is a linear model-based imputation

How can I check all formats? We utilize an in-house software for data processing but could replicate the file format in order to load it to MSstats

Please check the MSstats vignette (describes the format) + MSstatsConvert package may help

What is the criteria to decide which best normalization we have to use?

This question has been answered live

Question regarding the converter from Spectronaut output to MSstat format: the searching setting in the Spectronaut will affect the Spectronaut output, for example, the filtering type could be Qvalue (only proteins that pass the qvalue threshold have numeric values, others listed as Filtered) , Qvalue sparse (all proteins in all runs have numeric values), in this case, which setting in the spectronaut would you recommend? Would different spectronaut search settings affect the MSstats anlaysis output? thanks!

Ting Huang is typing an answer...

Ting Huang 10:30 AM
We haven't done such evaluation to see which options generates best results. But for spectronaut output, MSstats also has the option to filter the peptides based on the Qvalue

As a newby to R, Skyline and MSstats !!! I am wondering if there is any repository of computed files + associated R scripts for different projects to be used as tempelate? (something like a repository)?

There are analysis on MassiveIVE.quant (we will be reviewing this on Thursday!)
https://massive.ucsd.edu/ProteoSAFe/static/massive-quant.jsp

Is there a MASCOT output converter?

Mateusz Staniak 10:29 AM
currently not, but most likely MASCOT outputs can be adapted easily to MSstats format with the MSstatsConvert package

Could you please repeat how did we specify the information provided in 'Summary of samples'?

This question has been answered live

I assume you will run into this by the end of this step as well, but just in case this is specific to me this is what the dataProcess call ends with for me:

```
 == the summarization per subplot is done.
Warning message:
In survreg.fit(X, Y, weights, offset, init = init, controlvals = control,  :
  Ran out of iterations and did not converge
```

It means you run the code successfully. The warning is fine and you can ignore it.

Can you use MA plots from microarray to test for normality of your data?

Yes, you can make MS plots to check the normality. But MSstats doesn't provide this plot

what is the advantage of feature level analysis over the protein level statistical analysis?

This question has been answered live

For the "Summary of Samples" table, can we include 3 technical replicates per biological sample?

This question has been answered live

I have a dia dateset with a spiked protein. Can I ran normalization based an internal standard?

yes, the normalization option is 'globalStandards' + need the input for 'nameStandards' as well.

the type of comparison you just described (repeated measures vs comparisons) is determined automatically, or can we change that?

it is determined automatically in the groupComparison function. User can't specific the model.

the input file come from skyline in the DIA analysis?

Yes

so if we use the same report format we can analyce other dia works by using this as input?

yes, if you have the same report format, you can reuse this R script your own analysis. You may change some parameters, such as normalization, according to the nature of your dataset

Is it possible to have profile plot header with gene names/proein names than Uniport ID?

MSstats uses the protein ID column provided by the data processing tools as proetin names. So if the processing tool uses gene names/proein name as protein IDs, then the profile plot will have gene names/proein name as head
With Skyline, you can include any of these values in your report and then rename the column header to "Protein Name".

when I run the code for the QC plot, a plot does not show up in the Plots window within R. How do i view the plot?

Use address = FALSE parameter for dataProcessPlots

which.protein  = c(prot1,prot2, prot3,..)  can be a vector ?

Yes,

For the comparisons: can I upload my own table with the samples info or does it have to match the levels of GROUP_ORIGINAL in the way its generated by the previous steps?

This question has been answered live

Is it ok to have varying number of technical reps in the comparison groups?  How does that affect the results?

This question has been answered live

Would you mind introduce some good sources (book, article, etc) on statistics for mass spec proteomics?

I started from this paper: “Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments” by Oberg and Vitek

if we have like two excel input to compare, example DIA proteome profile from wild type vs ko ... should we create an unique matrix?

Sorry, I don’t fully understand your question. The columns of contrast matrix should be the same as the number of groups in the data

In what cases we also need to set the log fold change cutoff?

You can set a FC cut off if you only want to accept significant protein changes that are above above a certain absolute FC
p value (adj.pvalue) means statistically significance. p-value is more meaningful. no required to cut off for fold change.

Can we change the colors of the volcano plot? Or display lable names of the colored points?

Unfortunately, can’t change the colors.
Fortunately, can display the protein name, ProteinName=TRUE

maybe I missed but could you comment on setting fold change cut-off for proteins that were significant after p-value correction? do use arbitrarry cut-off like 1.5?

By default MSstats does not set a FC cut-off. It is up to the user to define what cut off they need if any.

Can we get it to show only a few names? or is all or nothing?

it is all (display all the significant protein names) or nothing.

Is there a function that outputs only the discoveries / the significant results per comparison? Or do I have to filter the output files myslef?

If you want only the significant results, you have to filter it by yourself
Yes, how you filter could be quite different by people. We leave it for users. Something like test[test$adj.pvalue < 0.05, ] can be used.

What does power value mean? the scale?

power is a statistical value which means the probability of rejecting the null hypothesis when it is actually false.

The desired FC is linear or Log2 value?

it is original scale, not log2

This is a question for Brendan. Is manual validation of all features a requirement for analysis by Skyline (regardless of exp type)?

This question has been answered live

I ran MaxQtoMSstatsFormat() on MaxQuant results with useUniquePeptide = TRUE and removeProtein_with1Peptide=TRUE options, I found that the formatted results returned had removed proteins that had measurements in 2 runs out of total of 6 runs -- why were these proteins removed?

Did these measurements come from the same peptide? If so, the second option removed them

Hi, you answered that: MSstats provides different normalization options to correct the technical bias between runs. But if you want to correct batch effects such as age or gender, MSstats is unable to do that. Is this envisoned to be included in future releases? It is rather important for many of our studies. Thanks

This question has been answered live

can you point us to a link for additional MSstats tutorials and specifics for how to import data from other search programs (e.g., Open SWATH, MaxQuant, etc.)

msstats.org + vignettes / help files for MSstats, MSstatsConvert, MSstatsTMT etc

https://msstats.org/wp-content/uploads/2020/02/MSstats_v3.18.1_manual_2020Feb26-v2.pdf for MSstats

Does MSstats support further analysis such as GO anotation? If not what package would you recommend? A lot of transcriptomic packages don't correct for the smaller background in proteomics datasets.

MSstats doesn't support that. And in general this is a bit tricky for proteomics - exactly for the reasons that you mention. It is not clear what the background is. So this is really not as much about the software but about what is the right choice of the background for your study

is there a interactive (shiny based) implementation for MSstats available (similar to DEP for MQ)?

We have a prorotyp, not ready for making publically available. Keep tuned!

There is also a GUI at usegalaxy.eu

In which case do I use the processed.quant$ProcessedData (Feature level?) over the processed.quant$RunlevelData (Protein level?)?

ProcessedData is just for quality control / knowing how the dataset was processed. Protein level data is used for statistical analysis

+ From v4, ProcessedData has imputed values and so on. In case that you want to see what values are used..

the plots are sorte alphabetically  per accession.  woukd it be better to sort  on basis on significance of differences?

If there is only two groups, then we can do that. if there are multiple groups, it is hard to do that since for different comparison, one protein may ahve different significance of differences.