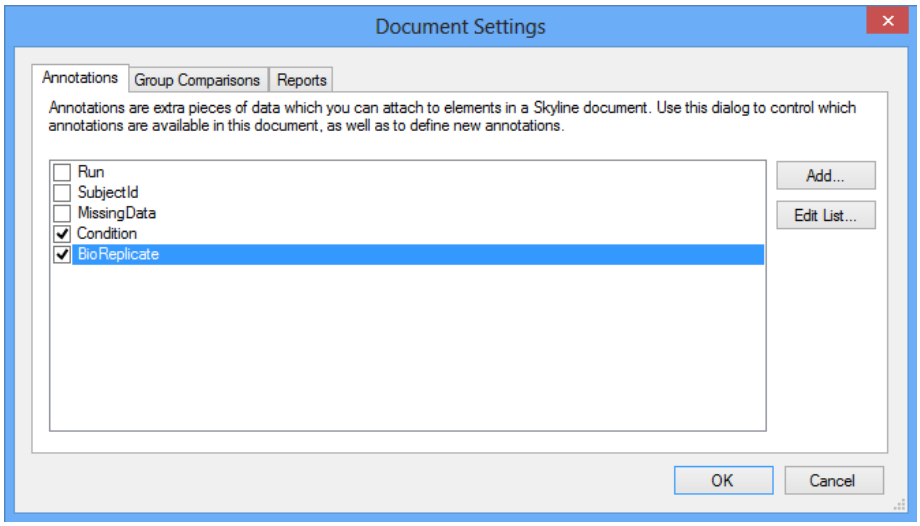
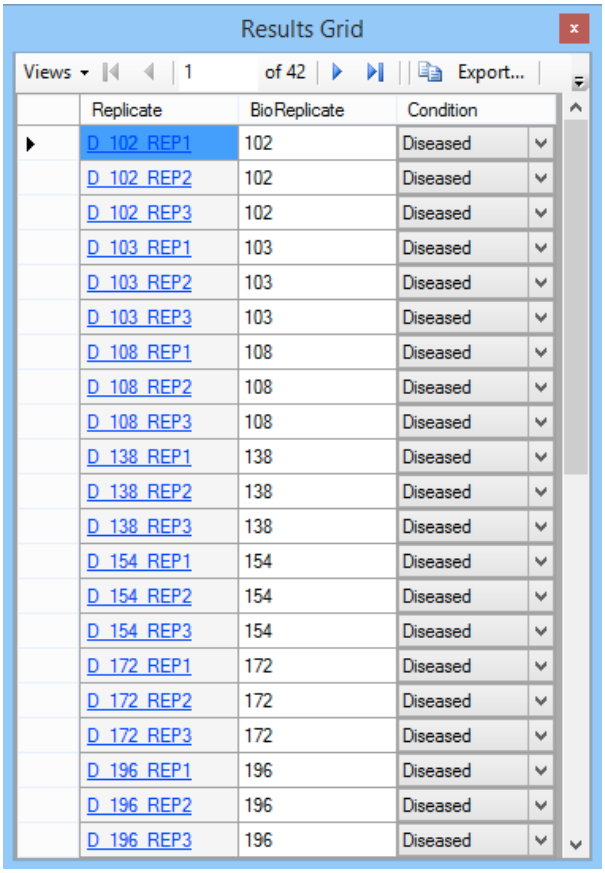
MSstats as an External Tool

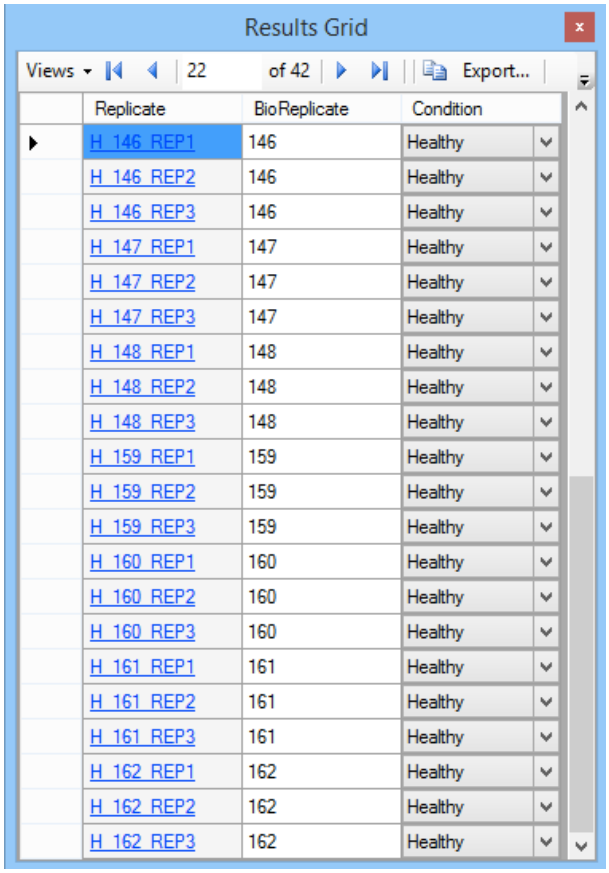
1. Annotating samples with group information

- Example data: rat plasma for heart failure
  - 7 control samples and 7 disease samples
  - 3 technical replicates
  - Total 42 injections (runs)
  - Label-free SRM
- Open the zipped Skyline file *Rat_plasma.sky.zip* in the folder *Tutorials/GroupedStudies/Heart Failure* of Flash drive.
- Before we can run MSstats, we need to annotate which samples are replicates. Skyline allows you to associate additional information with the runs in the document by defining custom annotations. To view the Annotation Settings form, perform the following steps:
  - On the "Settings" menu, click "Document Settings…" → "Annotations" tab → "Edit List…" → Select "Condition" and click "Edit…":
    - Name: Condition
      - Type: Value List
      - Values: "Diseased", "Healthy", each on a new line (without " ")
      - Applies To: Replicates
      - Click twice "OK" to confirm.
  - After confirming the new annotation type, activate the 2 required Annotations for MSstats by checking their checkboxes: Condition and BioReplicate (with default settings).
  - Click the "OK" button.



- Edit annotation values in Skyline using the Results Grid.
  - To bring up the Results Grid, go to the "View" menu and click "Results Grid". The Results Grid will show you chromatogram peak areas and other measured results for the currently selected peptide, or transition.
  - Annotate "Condition", "BioReplicate" as shown in the table below.
    Caution! This table view will only be visible when a peptide is selected in the target window. If a protein, a precursor or a transition is selected the columns will change accordingly.
    Caution! If you see little pencil icon at the left bottom, move the cursor to any other cell to save all the information.



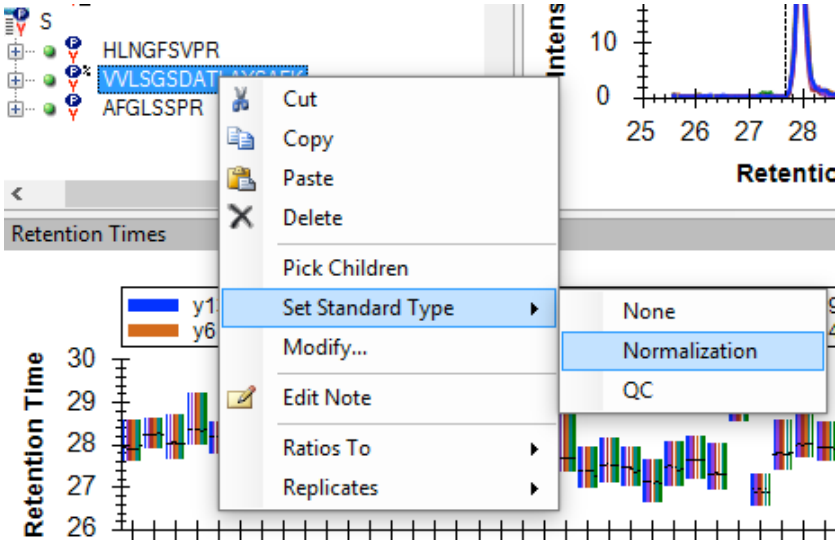


- Close the Results Grid.
- Save the Skyline file as *Rat_plasma.sky* in the folder.

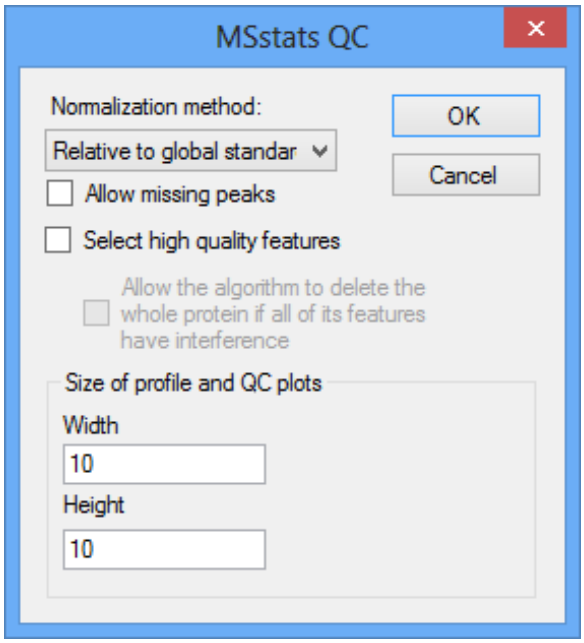
2. Running MSstats

- MSstats is composed of three individual tools:
  - **Quality Control (QC)**: Provides quality control statistics for MS runs and performs run-level summarization of protein abundance to be used in subsequent analysis. Logarithm transformation with base 2 and then normalization to remove systematic bias between MS runs is applied.
  - **Group Comparison**: Tests for significant changes in protein abundance across conditions based on linear mixed-effects models.
  - **Design Sample Size**: Calculates sample size for future experiments using an intensity-based linear model.
  - **Tip!** For more detailed information about functionality and options, visit **msstats.org**.

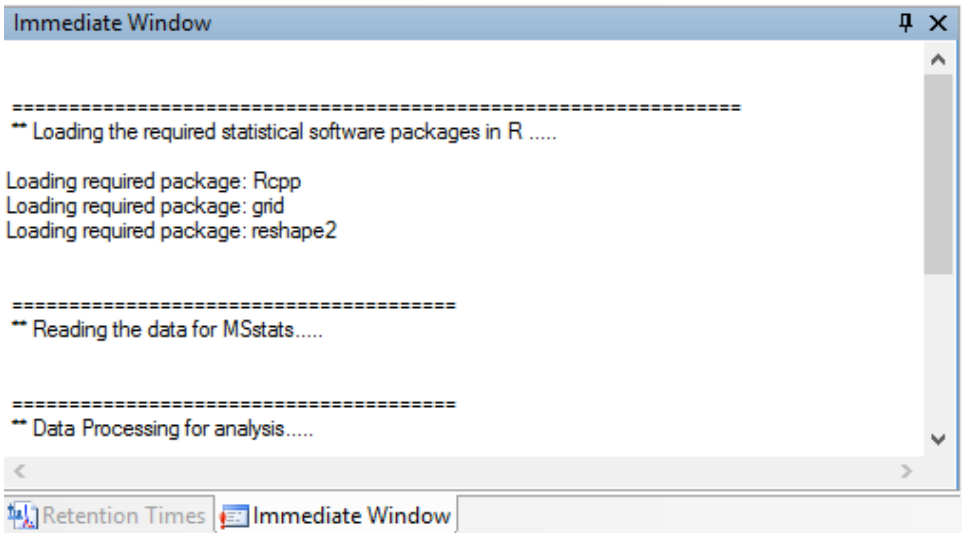
- To run a QC analysis, perform the following steps:
  - For normalization with global standards, choose global standard proteins or peptides.
    - Click peptide 'VVLSGSDATLAYSAFK'
    - Right click and select 'Set Standard Type'
    - Choose 'Normalization'
    - Then '%' is shown on the left side of the peptide name



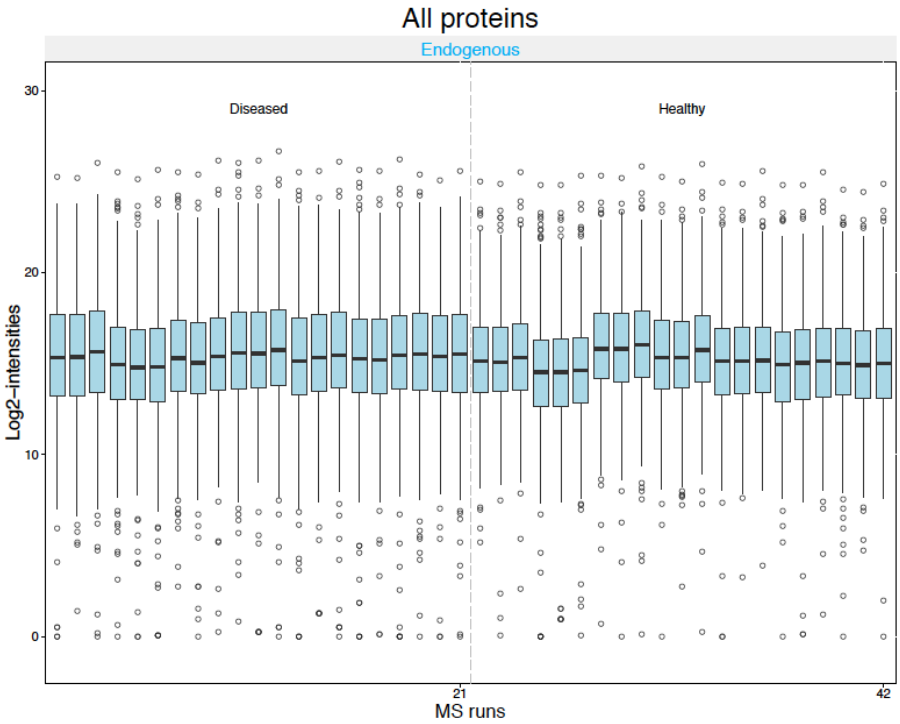
- On the “Tools” menu, choose “MSstats” and click “QC”.
- Choose ‘Relative to global standards’ for normalization.
- Uncheck ‘Allow missing peaks’: whether adding incomplete rows in the input or not.
  - Option: To remove features with interference not agreeing with the pattern of the average features across runs, check ‘Select high quality features’.



- By clicking “OK”, Skyline will begin to export the MSstats Input report. Once the report is exported, Skyline will output the results of the tool run in the Immediate Window docked at the bottom of the main Skyline Window.

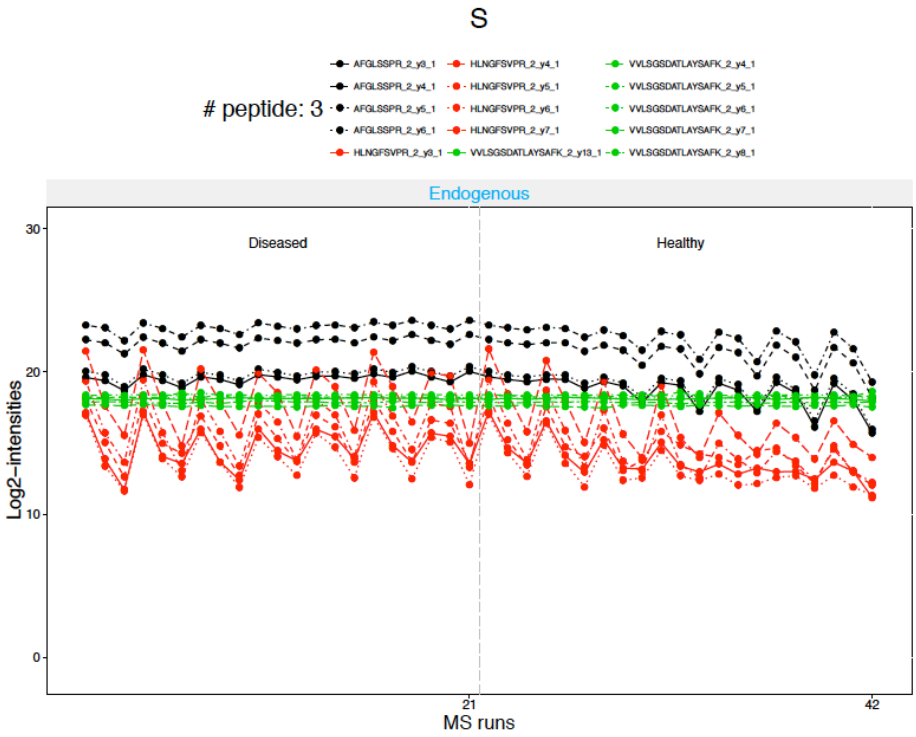


- When the tool is completed ("Finished." In the Immediate Window), navigate to the directory containing *Rat_plasma.sky* under the folder (*Tutorials/GroupedStudies/Heart Failure)*. This directory should now contain the outputs of the dataProcess function: dataProcessedData.csv, ProfilePlot.pdf, QCPlot.pdf, ConditionPlot.pdf. There are also 2 text files containing information about the software/package versions: msstats.log and sessioninfo.txt.
- The three output plots help examine the data quality in the experiment.
  - QCPlot shows the boxplots of peak intensities (on log scale) in all runs, where the bottom and top of a box represent the first and third quartiles of the log-intensities and the band inside the box is the median. It provides a quick way to examine and compare distributions between runs, and to detect systematic bias.

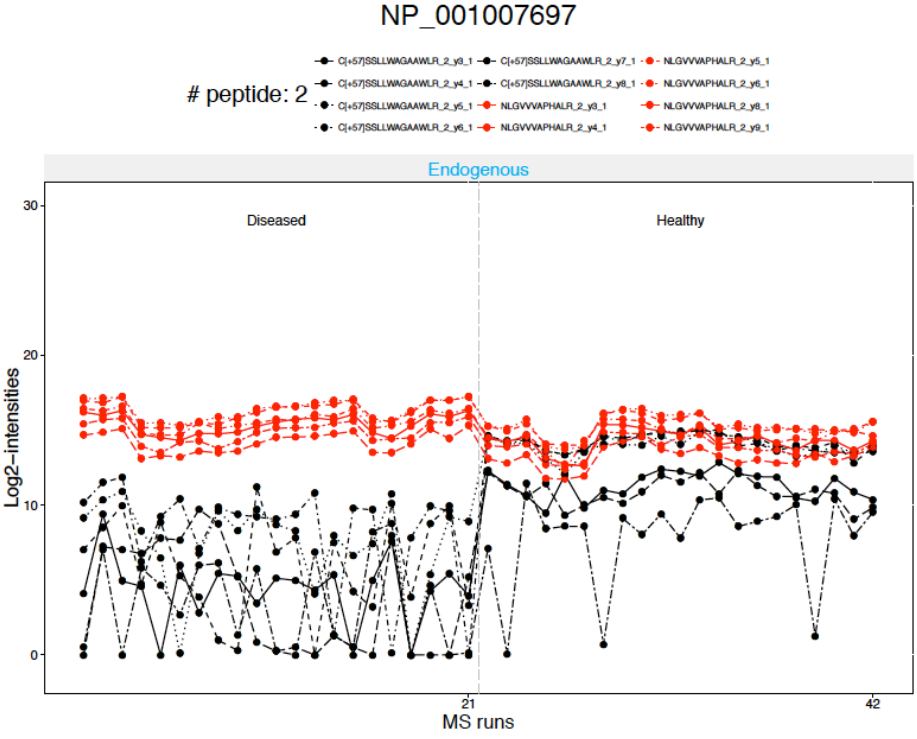


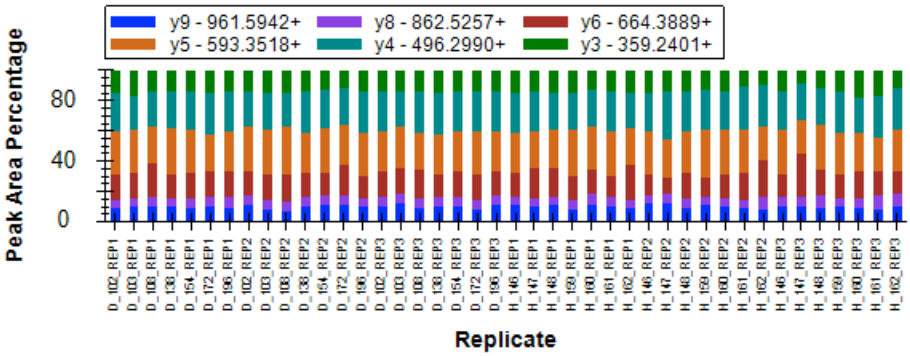
  - ProfilePlot shows individual observations for each protein. It is useful to examine the consistency of measurements in transition, run and condition, and to detect potential source of variation and missingness in the data.

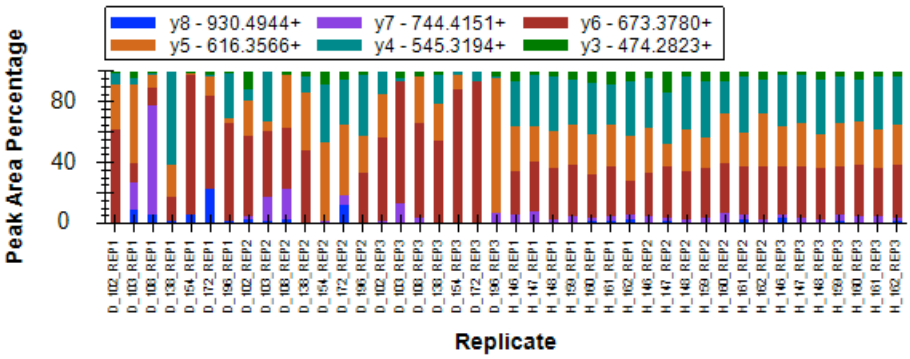
- Peptide 'VVLSGSDATLAYSAFK' was used for normalization, as confirmed in the profile plot.



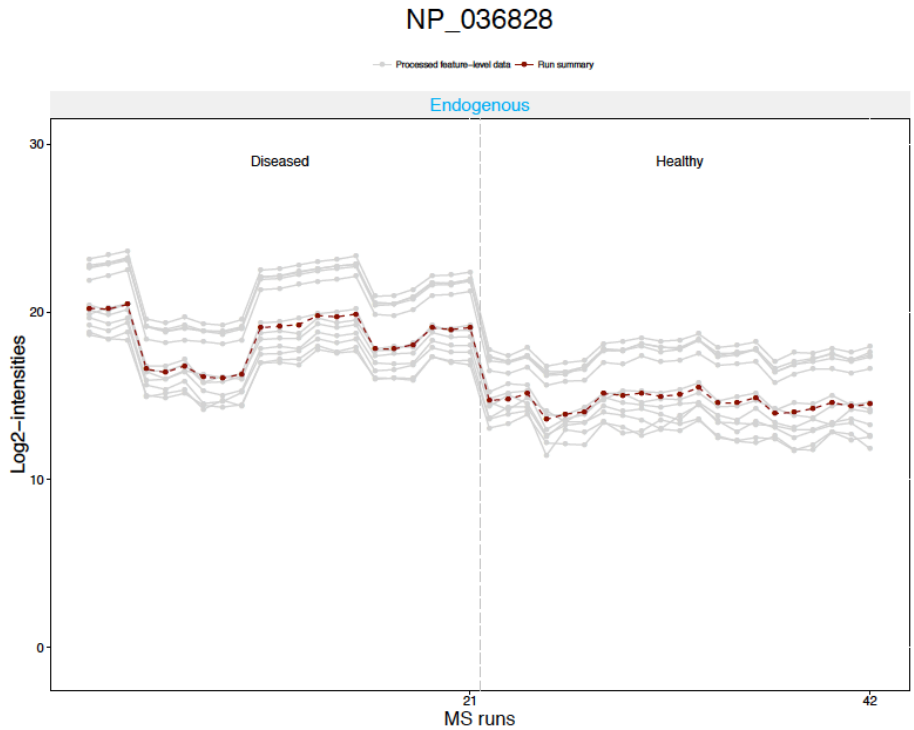
- Parallel profiles on log scale correspond to consistent peak area percentage (for example, peptide 'NLGVVVAPHALR'), from which we gain confidence in the integration of the peptide. When any inconsistency is observed, we should look into the data before conducting subsequent analysis.



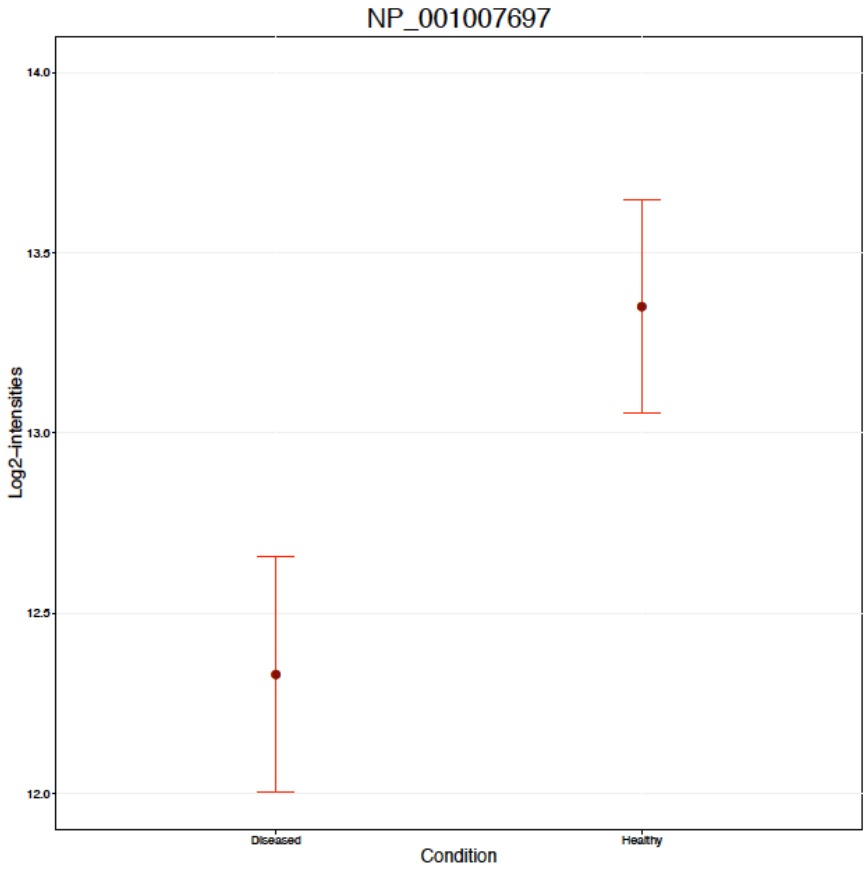




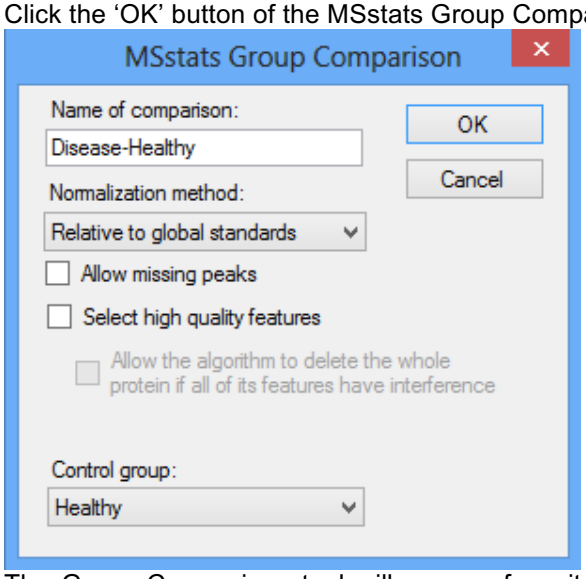
- ProfilePlot_wSummarization overlays the summarized protein abundance on top of the feature profiles.



- ConditionPlot shows the mean of log-intensity and 95% confidence interval for each condition. To answer whether a protein is differentially abundant between conditions, however, this plot is not sufficient and group comparison analysis needs to be conducted.



- To run a Group Comparison analysis, perform the following steps:
  - On the “Tools” menu, choose “MSstats” and click “Group Comparison”.
  - In the Name of comparison textbox enter “Disease-Healthy”.
  - Choose ‘Relative to global standards’ in Normalization method.
  - Click ‘Healthy’ in Control group.
  - Uncheck ‘Allow missing peaks’.
    - Option: To select the most informative features which agree with the pattern of the average features across the runs, check ‘Select high quality features’.
  - Click the ‘OK’ button of the MSstats Group Comparison form.



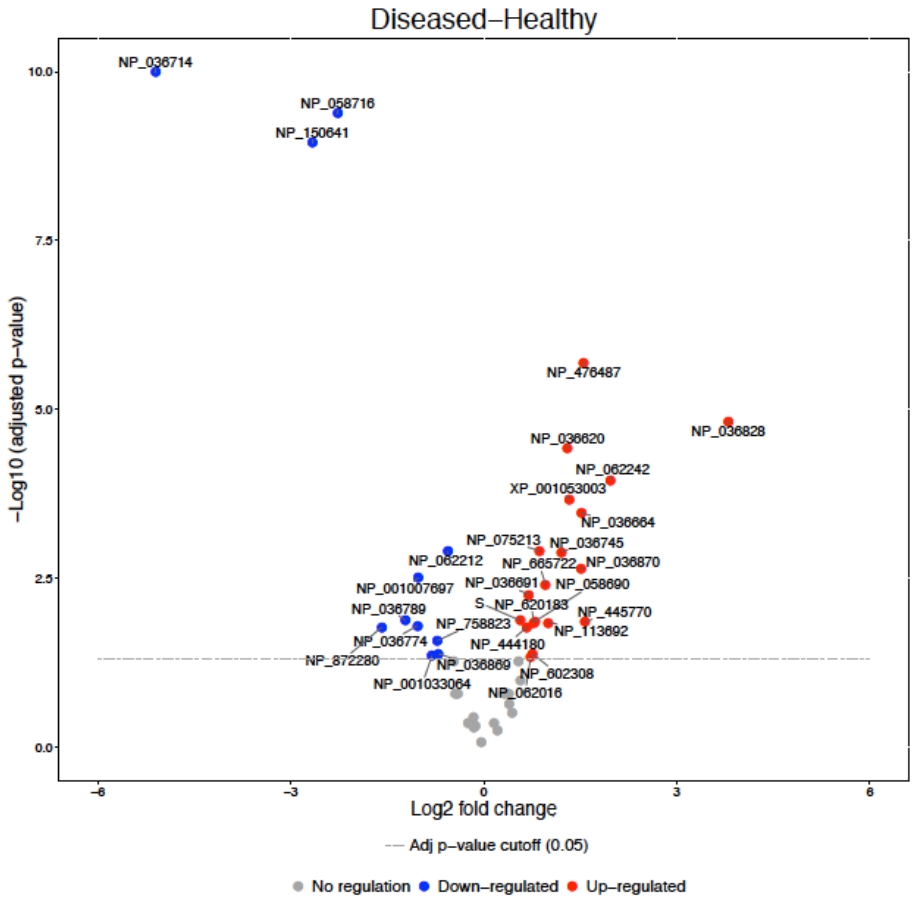
  - The Group Comparison tool will now perform its analysis. When the group comparison run is completed, switch back to the directory containing the

Rat_plasma.sky file. The directory should now contain VolcanoPlot, ComparisonPlot PDF files, TestingResult.csv file, and msstats-1.log file and dataProcessedData-1.csv generated by the tool.

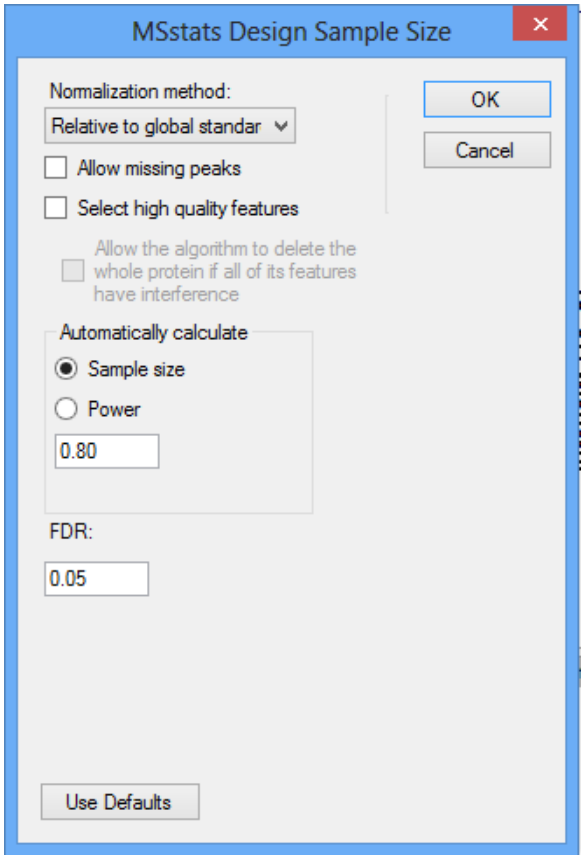
- Results of the group comparison including log2 fold change, standard error, and adjusted p-value are shown in TestingResult.csv.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue	issue	MissingPerce	ImputationPercentage	
2	1	NP_0010076	Diseased-Hea	-1.0226015	0.27588304	-3.7066487	28	0.0009173	0.00314504	NA	0.0218254	0.0218254	
3	2	NP_0010087	Diseased-Hea	0.39338814	0.28846953	1.36370778	28	0.18352357	0.23181925	NA	0	0	
4	3	NP_0010109	Diseased-Hea	0.31582353	0.19613331	1.61024933	28	0.11856115	0.16071834	NA	0.04285714	0	
5	4	NP_0010119	Diseased-Hea	-0.4626898	0.20845312	-2.2196346	28	0.03471625	0.05237893	NA	0.00621118	0	
6	5	NP_0010120	Diseased-Hea	0.56940559	0.30114834	1.89078111	28	0.06904307	0.10042629	NA	0.0014881	0.0014881	

- Volcano plot summarizes all the proteins with respect to their practical significance (log2 [fold change]) and statistical significance (-log10 [adjusted p-value]). Proteins with greater y values are more statistically significant. Changes with an adjusted p-value less than a significance level (default of 0.05) are considered as statistically significant. Up-regulated and down-regulated proteins are shown in red and blue, respectively.



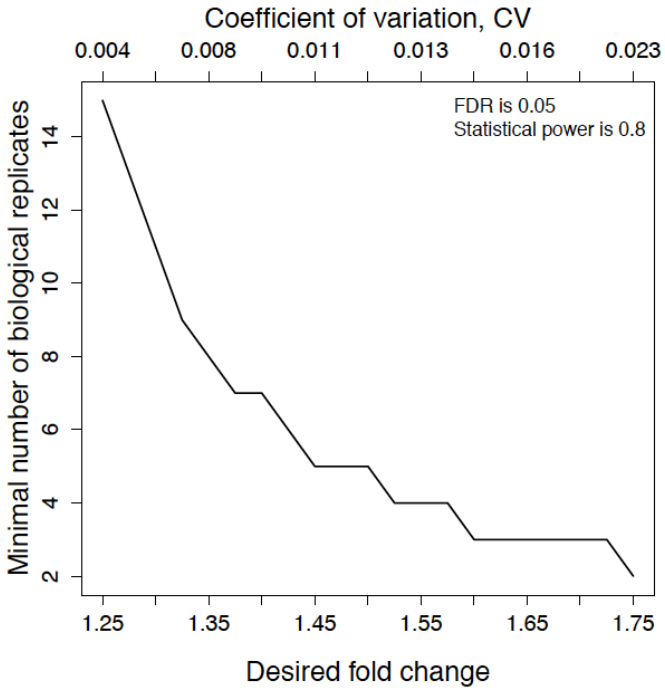
- To run a sample size calculation analysis, perform the following steps:
  - On the Tools menu, choose MSstats and click Design Sample Size.
  - Choose ‘Relative to global standards’ in Normalization method.
  - Uncheck ‘Allow missing peaks’
    - Option: To select the most informative features which agree with the pattern of the average features across the runs, check ‘Select high quality features’.
  - For this tutorial you will leave the options in their default settings.
    - The default setting will calculate sample size in order to get power=0.8 with FDR=0.05 between 1.25 and 1.75 fold change.
    - One of “Sample size” or “Power” needs to be selected for calculation. The other value needs to be provided.
  - Click the OK button of the MSstats Design Sample Size form.



- The Design Sample Size tool will now perform its analysis. When the analysis is completed, switch to the directory *Rat_plasma.sky* file again. The directory should now contain a SampleSizePlot PDF file, SampleSizeCalculation CSV file, and msstats-2.log and dataProcessedData-2.csv file generated by the tool.
- Required sample size corresponding to each of the expected fold changes (effect size) is reported in SampleSizeCalculation.csv. For example, to detect a fold change of 1.3 with power 0.8 and false discovery rate 0.05, the minimum required sample size (per condition) is 11.

B	C	D	E	F
desiredFC	numSample	FDR	power	CV
1.25	15	0.05	0.8	0.004
1.275	13	0.05	0.8	0.005
1.3	11	0.05	0.8	0.006
1.325	9	0.05	0.8	0.007
1.35	8	0.05	0.8	0.008
1.375	7	0.05	0.8	0.008
1.4	7	0.05	0.8	0.008
1.425	6	0.05	0.8	0.01
1.45	5	0.05	0.8	0.011
1.475	5	0.05	0.8	0.011
1.5	5	0.05	0.8	0.011
1.525	4	0.05	0.8	0.013
1.55	4	0.05	0.8	0.013
1.575	4	0.05	0.8	0.013
1.6	3	0.05	0.8	0.017
1.625	3	0.05	0.8	0.017
1.65	3	0.05	0.8	0.016
1.675	3	0.05	0.8	0.016
1.7	3	0.05	0.8	0.016
1.725	3	0.05	0.8	0.016
1.75	2	0.05	0.8	0.023

- SampleSizePlot visualizes the results in SampleSizeCalculation.csv



- To help troubleshoot potential problems with installation or functionalities of MSstats, a progress report is generated in a log file msstats.log. The file includes information on the R session (R version, loaded software libraries), options selected by the user, checks of successful completion of intermediate analysis steps, and warning messages. If the analysis produces an error, the file contains suggestions for possible reasons for the errors. If a file with this name already exists in working directory, a suffix with a number will be appended to the file name.