thermoscientific

Integration of MSstatsTMT into Proteome Discoverer Using the Scripting Node

David M. Horn,[1] Ting Huang,[2] Meena Choi,[2] Olga Vitek,[2] Rosa I. Viner,[1] Frank Berg[3], Kai Fritzemeier,[3] and Carmen Paschke[3]
[1]Thermo Fisher Scientific, San Jose, CA, 95134, [2]Northeastern University, Boston, MA, [3]Thermo Scientific GmbH, Bremen, Germany

ABSTRACT

Purpose: Integration of the MSstatsTMT tools directly into Thermo Scientific™ Proteome Discoverer™ 2.5 software using the Scripting Node.

Methods: An R script was created to send the PSM and study information to the MSstatsTMT R library for pairwise ratio generation. The Scripting Node subsequently re-imports the results for presentation and visualization in the Proteome Discoverer software.

Results: This script was evaluated with 3 datasets: the yeast triple knockout Thermo Scientific™ Pierce™ TMT11plex standard, two replicates of the same yeast triple knockout standard, and a three replicates TMT11plex experiment showing the effect of SARS-CoV-2 infection of human cells. We also demonstrate the integration of the MSstatsTMT results with the enrichment analysis function, aiding the downstream bioinformatic analysis of each of these datasets.

INTRODUCTION

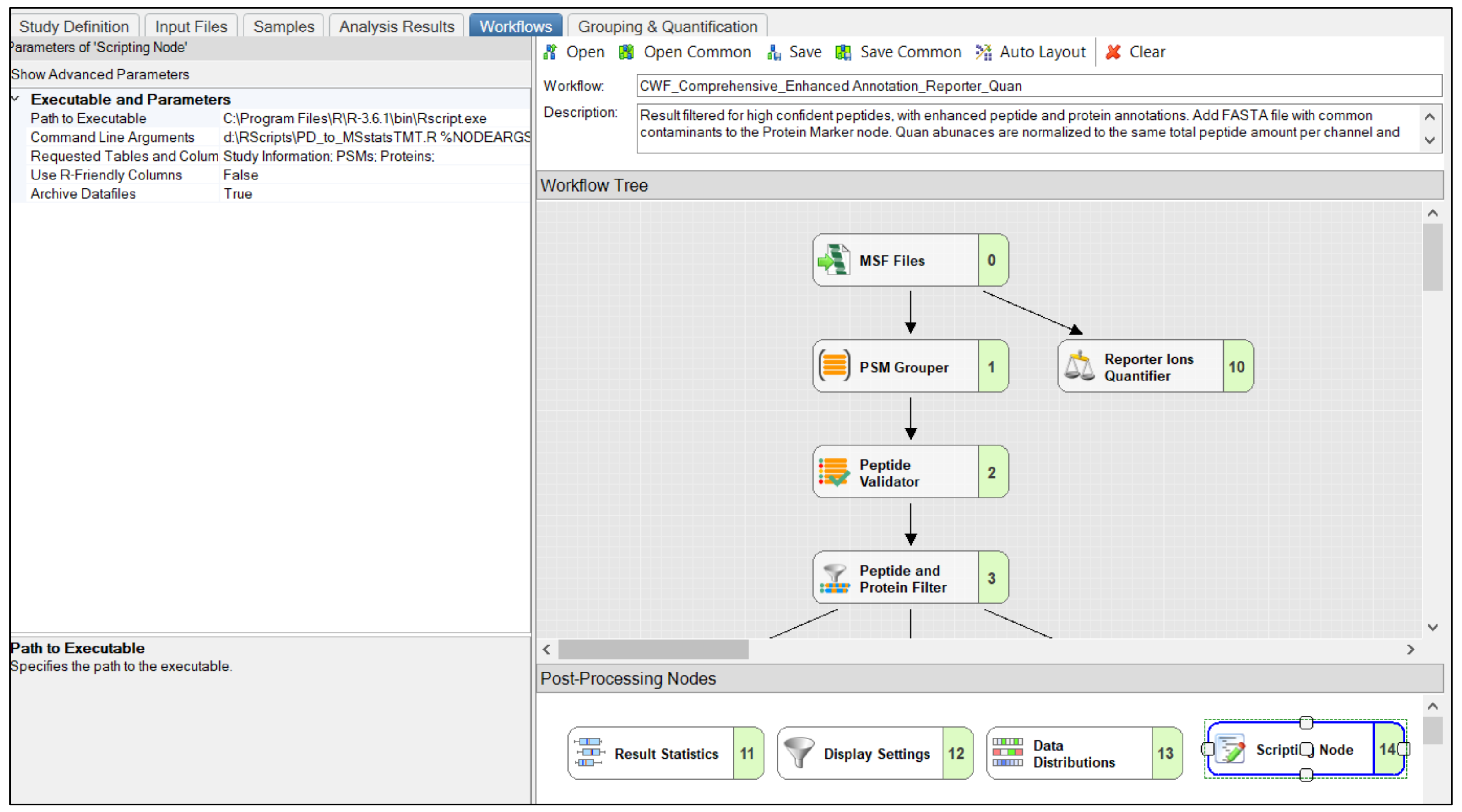
The MSstats software package has become a standard tool for DDA, SRM and DIA label-free quantification and recently the MSstatsTMT[1] libraries have been released for multiplexing TMT experiments. These R libraries support direct import of data from Proteome Discoverer, but they require a manual step to export the list of peptide-spectral matches and a separate file that maps quan channels and the various study factors. Here we show how to use the Proteome Discoverer 2.5 Scripting Node to automatically call a script to send study factor annotations and peptide-spectral match tables to the MSstatsTMT/R Bioconductor library. The Proteome Discoverer software subsequently reimports the results for visualization and bioinformatic analysis. We will demonstrate this workflow on multiple TMT datasets of increasing complexity.

MATERIALS AND METHODS

Creation of MSstatsTMT script

TMT datasets were analyzed by the Proteome Discoverer software using the standard analysis template for TMT SPS MS3 data, adding a Scripting Node to the Post Processing section in the Consensus workflow (see Figure 1).

Figure 1. Consensus workflow with Scripting Node that calls the script for MSstatsTMT analysis.



Parameters of 'Scripting Node'

Show Advanced Parameters

Executable and Parameters

Path to Executable	C:\Program Files\R\R-3.6.1\bin\Rscript.exe
Command Line Arguments	d:\RScripts\PD_to_MSstatsTMT.R %NODEARGS
Requested Tables and Colum	Study Information; PSMs; Proteins;
Use R-Friendly Columns	False
Archive Datafiles	True

Path to Executable
Specifies the path to the executable.

Workflow: CWF_Comprehensive_Enhanced Annotation_Reporter_Quan

Description: Result filtered for high confident peptides, with enhanced peptide and protein annotations. Add FASTA file with common contaminants to the Protein Marker node. Quan abunaces are normalized to the same total peptide amount per channel and

The Scripting Node calls custom R script, exporting the Proteins, PSM, and Sample Information tables for input to the script. This Sample Information table is a new feature for Proteome Discoverer 2.5 that maps the quantification channels for the analyzed dataset to the study factors defined by the user. The R script first loads the PSM and Sample Information tables and subsequently converts the tables into the formats recognized by MSstatsTMT. Subsequently, the MSstatsTMT functions are called in the following order: PDtoMSStatsTMT, proteinSummarization, and groupComparisonTMT. The resulting test.pairwise table is exported for input back into PD. The script subsequently maps the accession numbers between the test.pairwise table and the existing Proteins table. Finally, the script instructs Proteome Discoverer where to find and import the test.pairwise and Proteins-test.pairwise association tables to integrate with the rest of the search results.

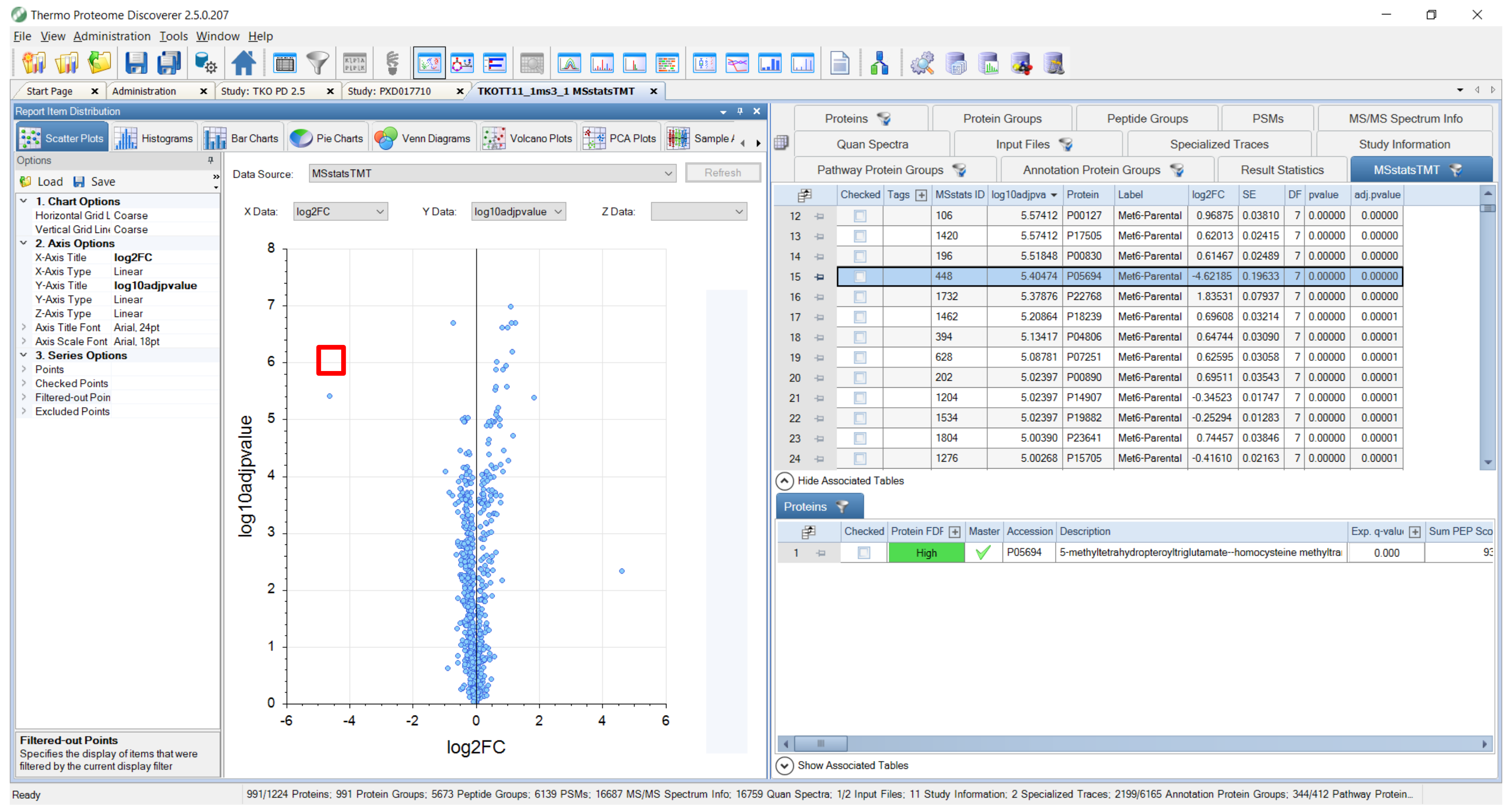
RESULTS

Example 1: Yeast triple knockout – single replicate

The yeast triple knockout standard is a TMT11plex dataset with 3 replicates each of the MET6, HIS4, and URA2 gene knockouts plus 2 replicates of the parental strain as the final two channels. For each of the gene knockout samples, the protein expressed by that gene should have little to no abundance in the results relative to the parental strains and other proteins in the amino acid biosynthetic pathways will be upregulated to synthesize the amino acids by alternative pathways. This specific dataset was acquired on a Fusion Lumos instrument using the SPS MS^3 approach. In Proteome Discoverer, the default workflows for TMT SPS MS^3 were used to analyze the dataset and a Scripting Node with the customized script was used to produce the MSstatsTMT information.

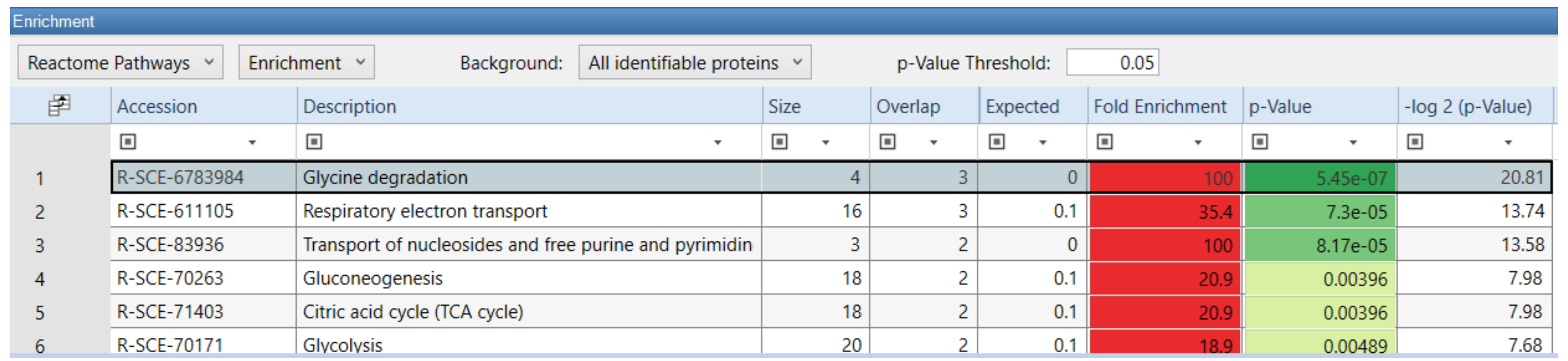
Figure 2 shows a plot of the log$_2$ fold change versus the –log10 (adjusted p-value) for the MSstatsTMT results using the Proteome Discoverer scatter plot function. The highlighted point is for the MET6 protein, which is highly downregulated as expected for such a sample.

Figure 2. Scatter plot of log$_2$ fold change versus –log$_{10}$ adjusted p-value. The highlighted point corresponds to the MET6 protein, which is down-regulated for the MET6 knockout sample. The table on the right is the pairwise ratio results returned from MSstatsTMT and the protein in the table below is associated with the selected entry on the table.



Proteome Discoverer 2.5 also includes a new "enrichment" analysis feature, which highlights gene ontology terms, pathways or protein family terms that are overrepresented in a given set of selected proteins. The enriched Reactome pathways for the 35 proteins with the MET6/parental fold change >1.5 and adjusted p-value <0.01 are shown in Figure 3. These indicate that up-regulated proteins participate in basic metabolism pathways as well as amino acid biosynthetic and degradation pathways.

Figure 3. Enriched Reactome pathways for up-regulated proteins for the MET6/parental ratio.



Enrichment

Reactome Pathways | Enrichment | Background: All identifiable proteins | p-Value Threshold: 0.05

	Accession	Description	Size	Overlap	Expected	Fold Enrichment	p-Value	-log 2 (p-Value)
1	R-SCE-6783984	Glycine degradation	4	3	0	100	5.45e-07	20.81
2	R-SCE-611105	Respiratory electron transport	16	3	0.1	35.4	7.3e-05	13.74
3	R-SCE-83936	Transport of nucleosides and free purine and pyrimidin	3	2	0	100	8.17e-05	13.58
4	R-SCE-70263	Gluconeogenesis	18	2	0.1	20.9	0.00396	7.98
5	R-SCE-71403	Citric acid cycle (TCA cycle)	18	2	0.1	20.9	0.00396	7.98
6	R-SCE-70171	Glycolysis	20	2	0.1	18.9	0.00489	7.68

Example 2: Yeast triple knockout – 2 replicates

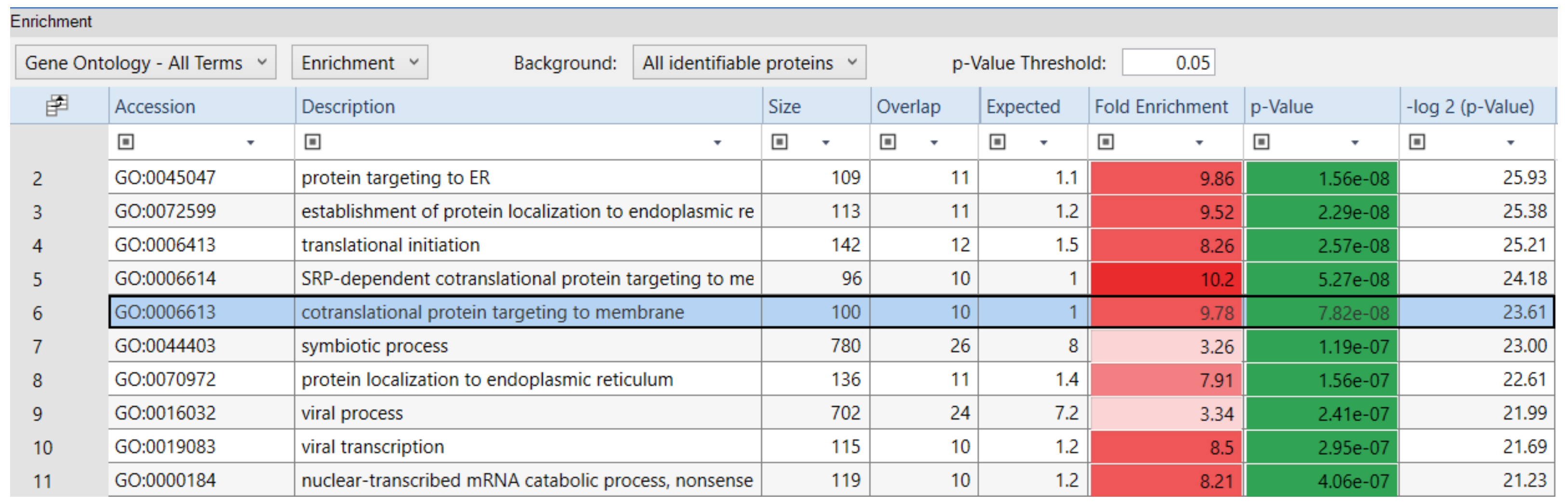
For this example, two technical replicate datasets of the triple knockout were run in the same analysis. The increased number of replicates should lead to an increase in the number of significantly changed proteins for each gene knockout species compared to the parental strain. However, with multiple files, there are a subset of peptides with missing quantification values where the peptide was identified in only one of the two datasets. For the MET6/parental ratio, the new –log adjusted p-value is 14.7 versus 5.4 for the single dataset, showing the improvement in confidence in differential expression. There were 36 proteins significantly up-regulated using the MET6/parental ratio compared to 35 for the single dataset. In total, there were 82 proteins up-regulated across all ratios versus 74 for the single dataset. This shows the additional statistical power of combining multiple runs even with the potential for missing values.

Example 3: SARS-CoV-2 host cell proteomics

Bojkova *et al*[1] recently released a preprint that used TMT quantification to study the effect of the SARS-CoV-2 virus on human host cell proteins over a 24 hour period. The raw data for this study was published in the PRIDE archive with accession PXD017710. These data were analyzed in Proteome Discoverer 2.5 using a similar data analysis strategy as described in the preprint. These data included three biological replicates and uses a common bridge channel to scale abundances across samples. For this poster, only the TMT quantification of the pulsed heavy-labeled peptides were quantified and the ratios for the viral-infected versus control samples were compared.

To determine which proteins and biological processes that are down-regulated with viral infection, the MSstatsTMT table was filtered to show only the Control 24 hr timepoint versus the 24 hr virus infection with a log2 fold change of 0.5 or higher and an adjusted p-value less than 0.01. The resulting 472 proteins were involved in a large number of biological processes and pathways, including apoptosis, lipid metabolism, IGF transport, and MTORC1-mediated signaling. For the 107 proteins that were up-regulated with viral infection, the enrichment analysis returned overrepresented terms such as viral-mRNA translation, protein targeting to membrane, translation initiation, and interleukin-12 mediated signaling pathway. (Figure 4)

Figure 4. Enrichment analysis for proteins up-regulated after 24 hours of viral infection.



Enrichment

Gene Ontology - All Terms | Enrichment | Background: All identifiable proteins | p-Value Threshold: 0.05

	Accession	Description	Size	Overlap	Expected	Fold Enrichment	p-Value	-log 2 (p-Value)
2	GO:0045047	protein targeting to ER	109	11	1.1	9.86	1.56e-08	25.93
3	GO:0072599	establishment of protein localization to endoplasmic re	113	11	1.2	9.52	2.29e-08	25.38
4	GO:0006413	translational initiation	142	12	1.5	8.26	2.57e-08	25.21
5	GO:0006614	SRP-dependent cotranslational protein targeting to me	96	10	1	10.2	5.27e-08	24.18
6	GO:0006613	cotranslational protein targeting to membrane	100	10	1	9.78	7.82e-08	23.61
7	GO:0044403	symbiotic process	780	26	8	3.26	1.19e-07	23.00
8	GO:0070972	protein localization to endoplasmic reticulum	136	11	1.4	7.91	1.56e-07	22.61
9	GO:0016032	viral process	702	24	7.2	3.34	2.41e-07	21.99
10	GO:0019083	viral transcription	115	10	1.2	8.5	2.95e-07	21.69
11	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense	119	10	1.2	8.21	4.06e-07	21.23

This indicates that the MSstatsTMT algorithms are indeed detecting the effect of viral infection, showing that further virus is being produced and apoptotic and anti-inflammatory responses are being suppressed. In comparison, the built-in Proteome Discoverer statistics produced 30 proteins that were up-regulated and 51 proteins that were down-regulated using the same filtering criteria.

CONCLUSIONS

- The MSstatsTMT algorithms were successfully integrated into the Proteome Discoverer software using the Scripting Node
- MSstatsTMT can handle simple to complex datasets and adeptly handles datasets with missing values
- MSstatsTMT results integrate well with other features in the platform, including the new enrichment service to be released in the Proteome Discoverer 2.5 software.

REFERENCES

1. Huang, T., Choi, M., Tzouros, M., Pandya, N.J., Banfai, B., Dunkley, T., Vitek, O., MSstatsTMT: Statistical detection of differentially abundance proteins in mass spectrometry experiments with isobaric labeling. ASMS 2019 poster WP386. Manuscript in preparation.

2. Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., Jindrich, C., Münch, C., SARS-CoV-2 infected host cell proteomics reveal potential therapy targets. Preprint: https://www.researchsquare.com/article/rs-17218/v1.

TRADEMARKS/LICENSING

ThermoFisher
SCIENTIFIC