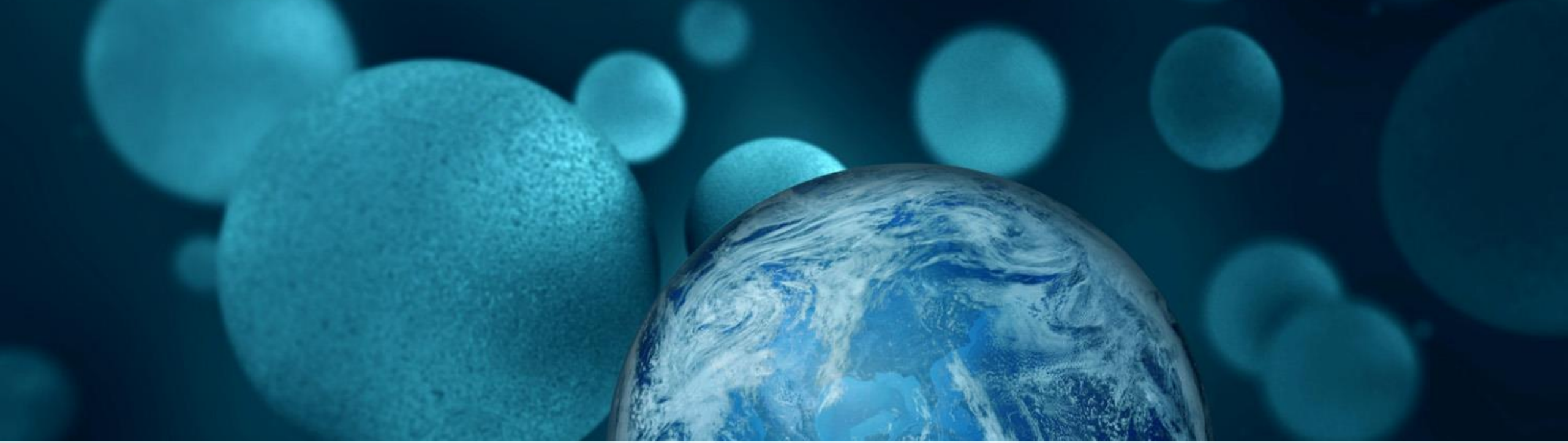


ThermoFisher SCIENTIFIC

Proteome Discoverer 2.4 overview

Proteome Discoverer History

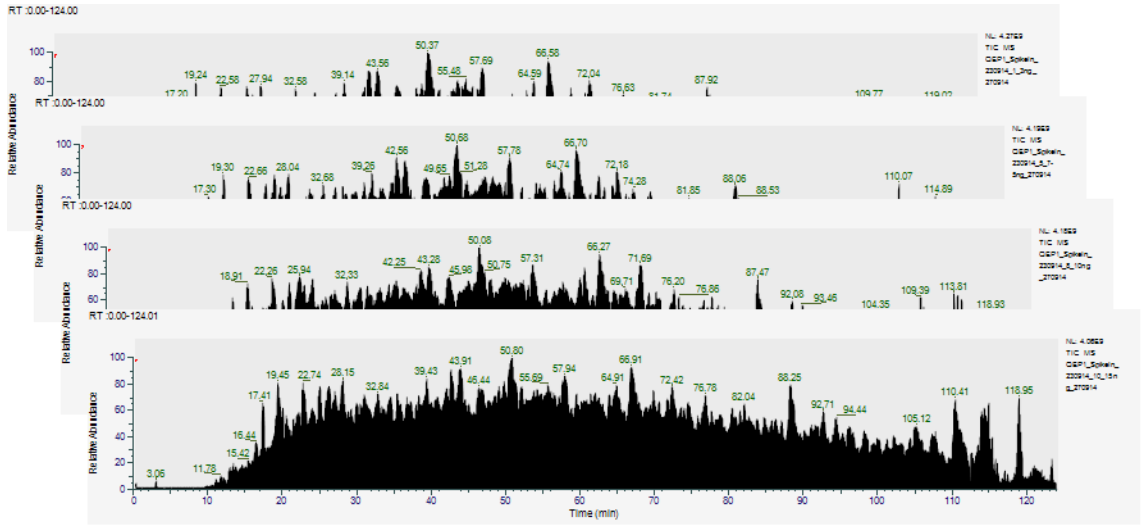
Software for identification and quantification of proteins in complex samples

1.0 - July, 2008	Workflow-based solution for protein identification
1.1 - Nov, 2009	Automation, batch processing, reporter ion quantification
1.2 - Apr, 2011	SILAC quantification
1.3 - Mar, 2012	Validation (Percolator, PhosphoRS), biological annotation, 64-bit
1.4 - May, 2013	Deep data mining (Sequest HT, library searching)
2.0 - Mar, 2015	Architectural changes, study management, large data sets
2.1 - Oct, 2015	Improved reporter ion quantification
2.2 – July, 2017	Label Free Quantification, Statistics, Cross-Linking
2.3 – Jan, 2019	Improved library search, heat maps, PTM site tables, cross-link quantification, annotation groups, ProSightPD
2.4 – Oct, 2019	Precursor detector/chimeric spectra, scripting node, TMTpro, new licensing

What is Proteome Discoverer?

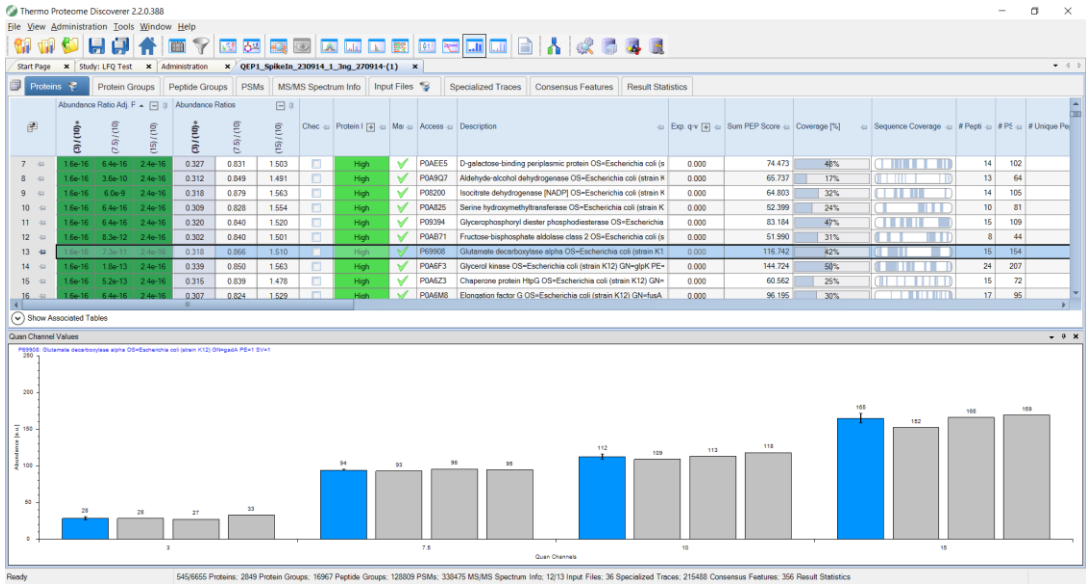
- Software for analysis of quantitative discovery proteomics data
- Supports hybrid Orbitrap, Q Exactive, and ion trap mass spectrometers
- Is software for simple conversion of raw files to protein ID lists (with quan) enough?

RAW files



?

Identified proteins and quantitative results

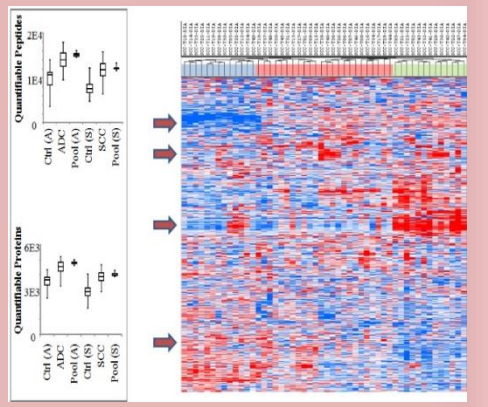


ThermoFisher SCIENTIFIC

Challenges for analysis of quantitative proteomics data

Complex studies/Large datasets

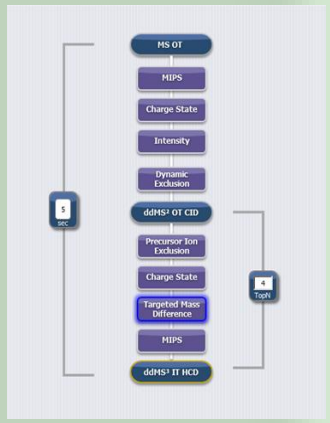
- 100's of raw files
- Results need to presented by sample, not raw file
- Statistics and proper study design are required



Requires study management

Complex acquisition methods

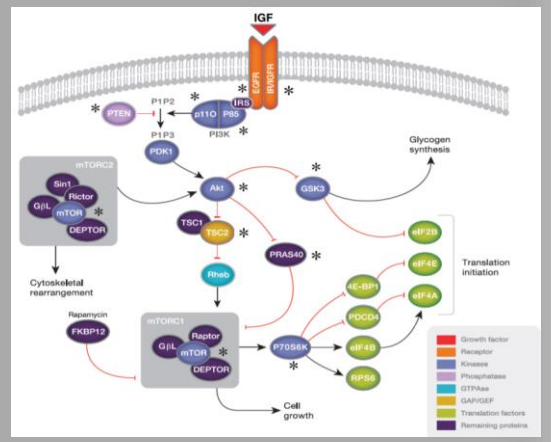
- TMT – SPS MS^3
- Glycopeptides – HCD-triggered-> CID->EThcD
- Cross-linking – MS2/MS2/MS3
- Top down – CID, ETD, HCD, EThcD, UVPD



Requires customizable workflows

Biological complexity

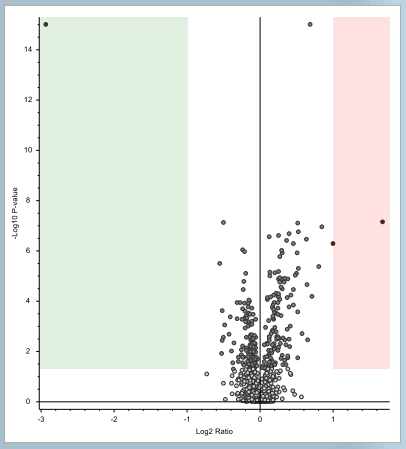
- >10,000 proteins
- PTMs
- Proteoforms
- Metaproteomics
- Pathway analysis
- Protein structure



Requires links to bioinformatic databases

Results interpretation

- How to denote significantly changing proteins/peptides?
- What is already known about proteins of interest?
- How do we make biological conclusions?



Requires statistics and visualization

ThermoFisher SCIENTIFIC

What is Proteome Discoverer?

Study Management

- Maps “study factors” to quantification channels
- Set up replicates, statistical analysis
- Manage files and search results

Client/server based workflow processing system

- Customizable data analysis pipelines for complex acquisition methods
- Extensible framework allows faster deployment of new algorithms
- Support for large datasets

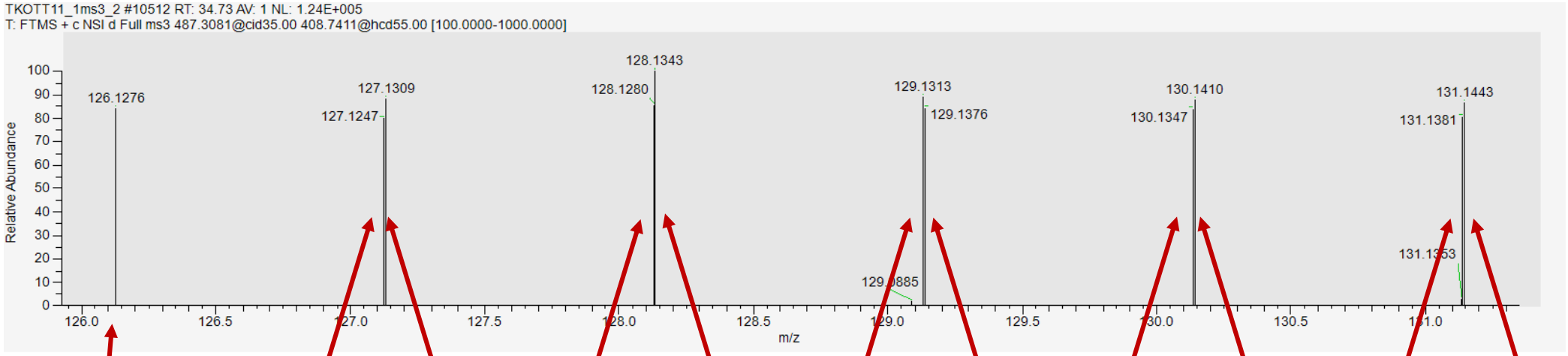
Biological Annotation

- Pathway, GO term, protein family annotation
- ProteinCard for summary of known information of selected proteins
- Links to KEGG, Wikipathway, Reactome maps

Data Interpretation

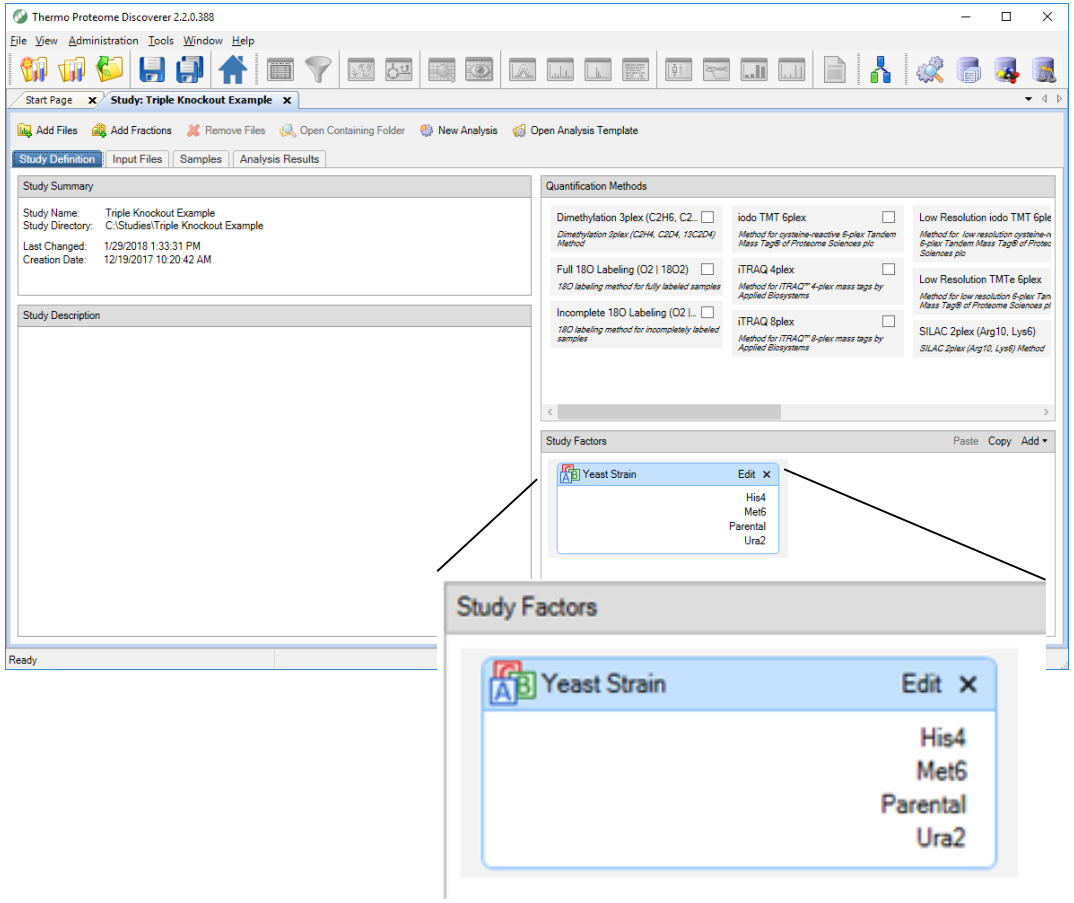
- Hierarchical views with links between proteins, peptides, PSMs
- Interactive graphical views for statistical analysis

Enables users to assign biological meaning (“study factors”) to quan channels:



Study Management in PD – Yeast Triple Knockout

- Create the 4 yeast strains as a study factor



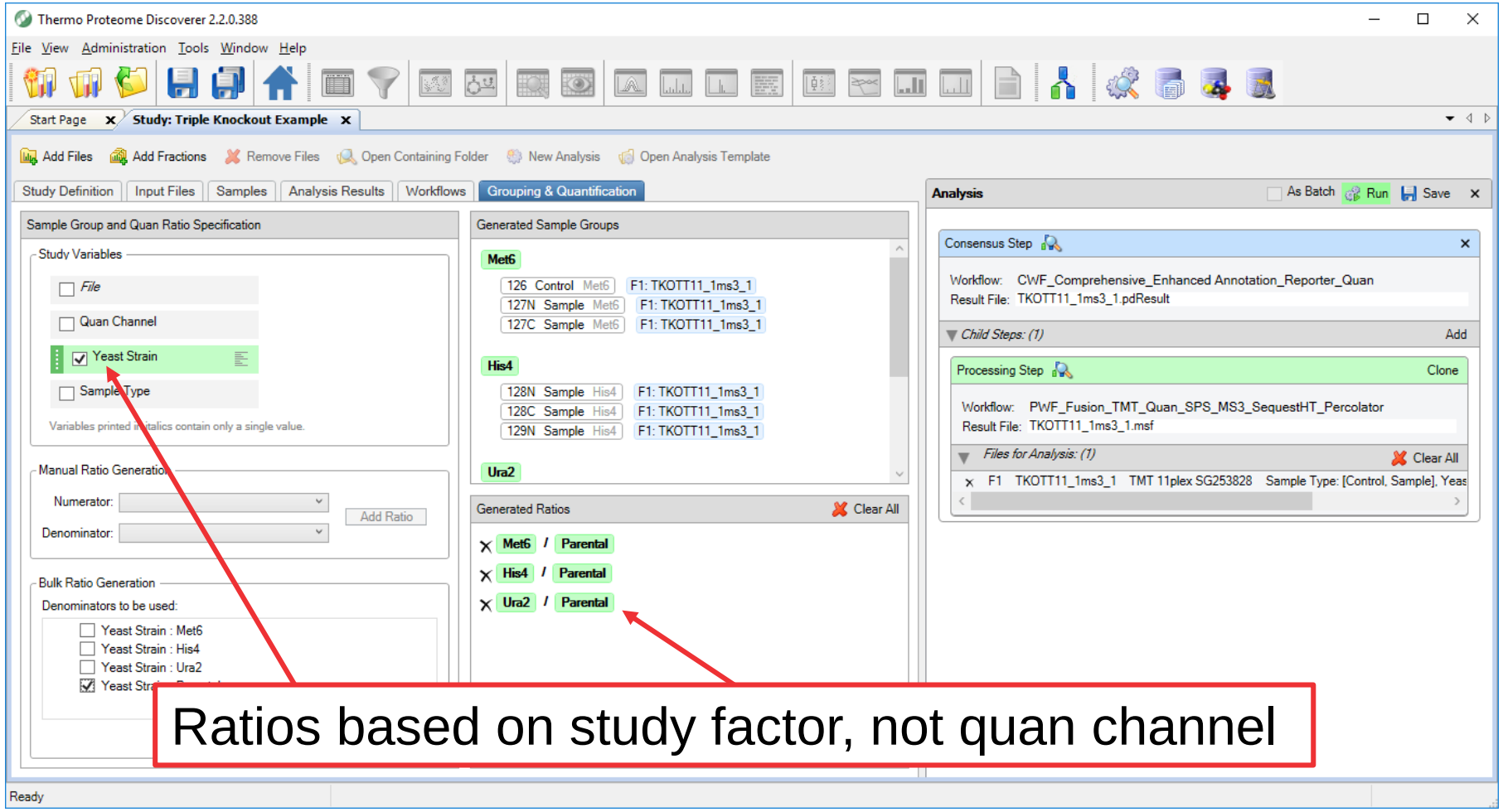
- Assign study factors to the quan channels

Study Definition | Input Files | Samples | Analysis Results

Sample	Sample Identifier	Sample Type	Yeast Strain
S1	TKOTT11_1ms3_1 - [126]	Control	Met6
S2	TKOTT11_1ms3_1 - [127N]	Sample	Met6
S3	TKOTT11_1ms3_1 - [127C]	Sample	Met6
S4	TKOTT11_1ms3_1 - [128N]	Sample	His4
S5	TKOTT11_1ms3_1 - [128C]	Sample	His4
S6	TKOTT11_1ms3_1 - [129N]	Sample	His4
S7	TKOTT11_1ms3_1 - [129C]	Sample	Ura2
S8	TKOTT11_1ms3_1 - [130N]	Sample	Ura2
S9	TKOTT11_1ms3_1 - [130C]	Sample	Ura2
S10	TKOTT11_1ms3_1 - [131N]	Sample	Parental
S11	TKOTT11_1ms3_1 - [131C]	Sample	Parental

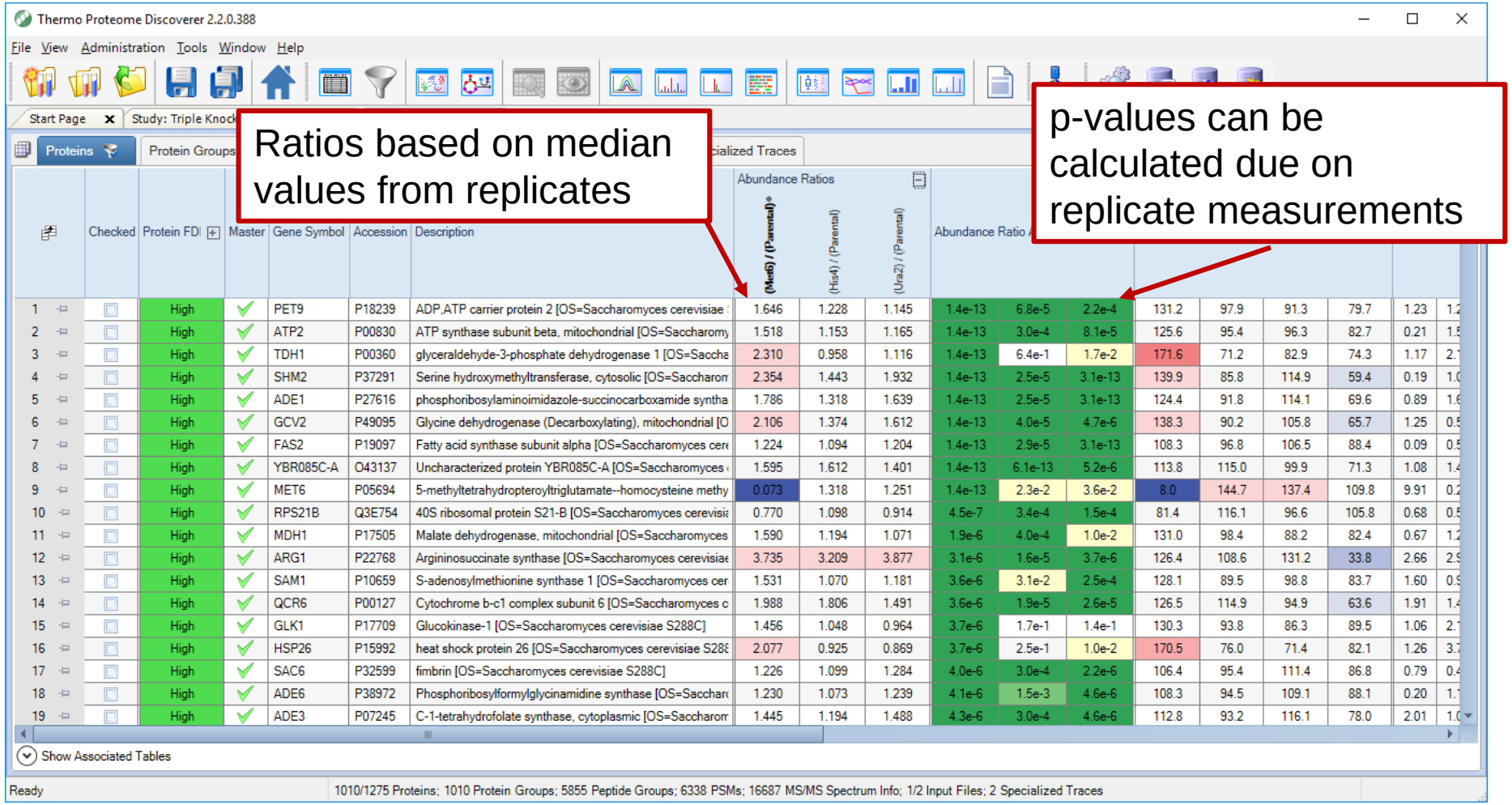
Study Management in PD – Yeast Triple Knockout

- Create quantitative ratios based on the Yeast Strain study factor:



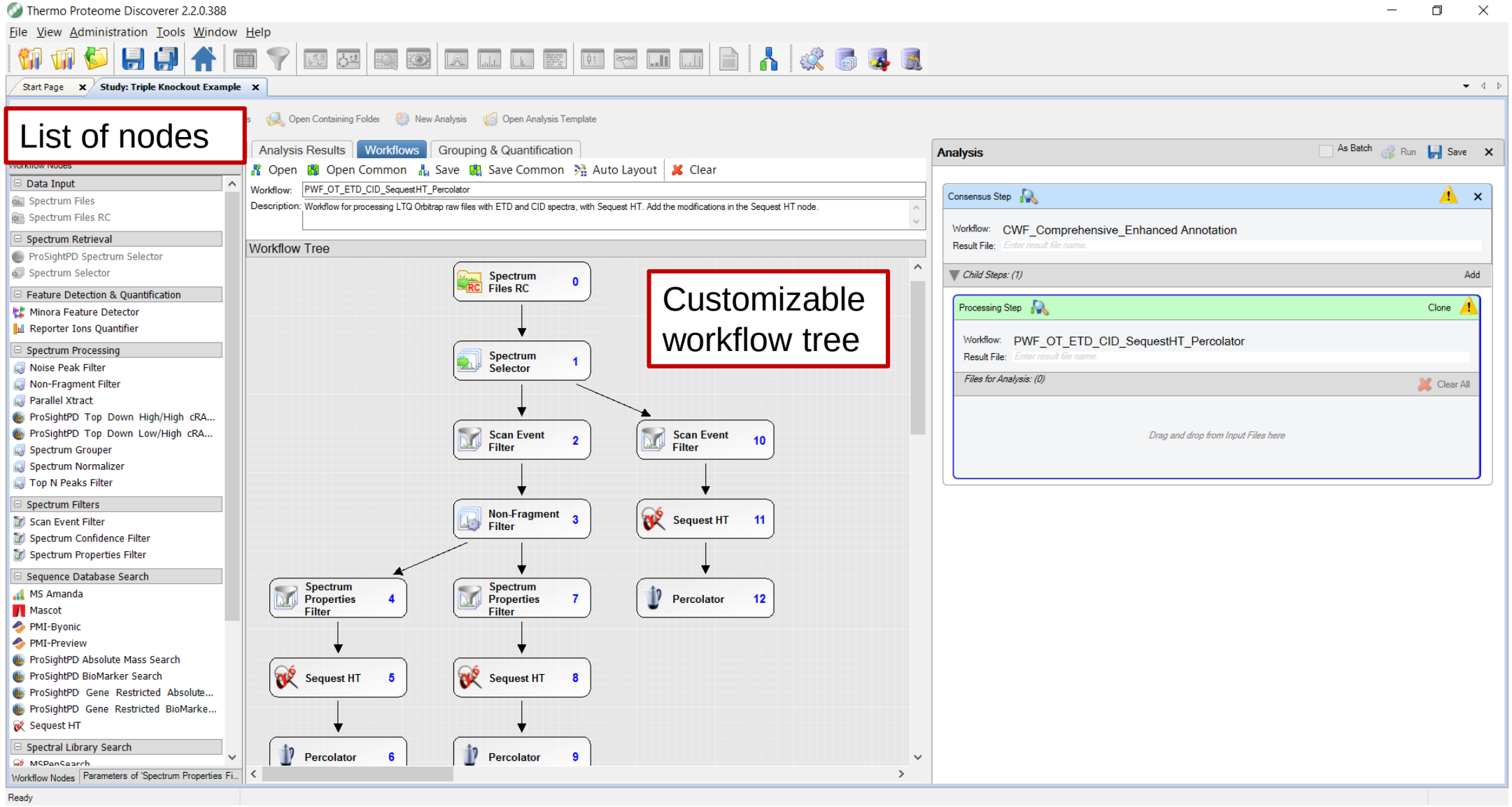
Study Management in PD – Yeast Triple Knockout

- View results based on study factor rather than quan channel



ThermoFisher SCIENTIFIC

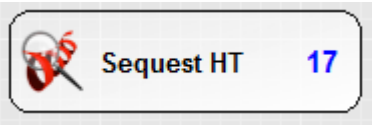
User customizable workflows for processing complex datasets



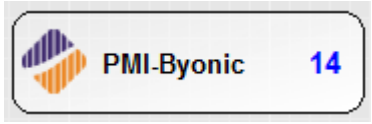
ThermoFisher SCIENTIFIC

Search engine nodes available in the PD framework

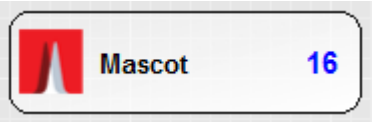
- Proteome Discoverer framework includes several search engines
- The nodes can be used in series or in parallel to identify more peptides than each individual search engine
- More search engines are currently being developed by 3rd parties
- Byonic, ProSightPD, Mascot require purchase of the standalone search engines



Default search engine



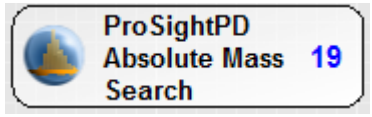
Glycosylation/PTM



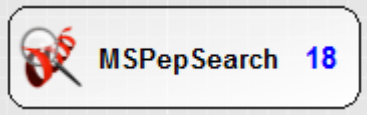
Connects to a local Mascot server

Quick preview search

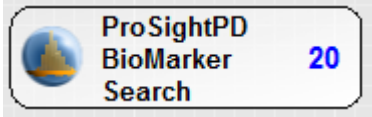
IMP Vienna (Mechtler group)



Top down protein identification



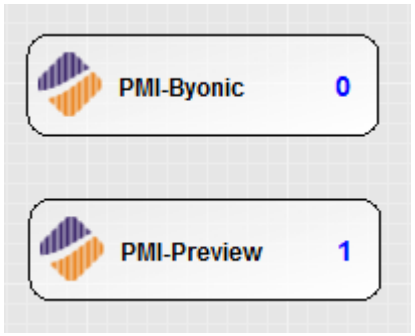
Library search engine from NIST



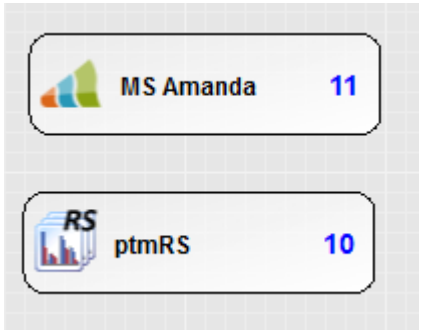
Truncated protein identification

Third party nodes (or nodes that encapsulate 3rd party algorithms)

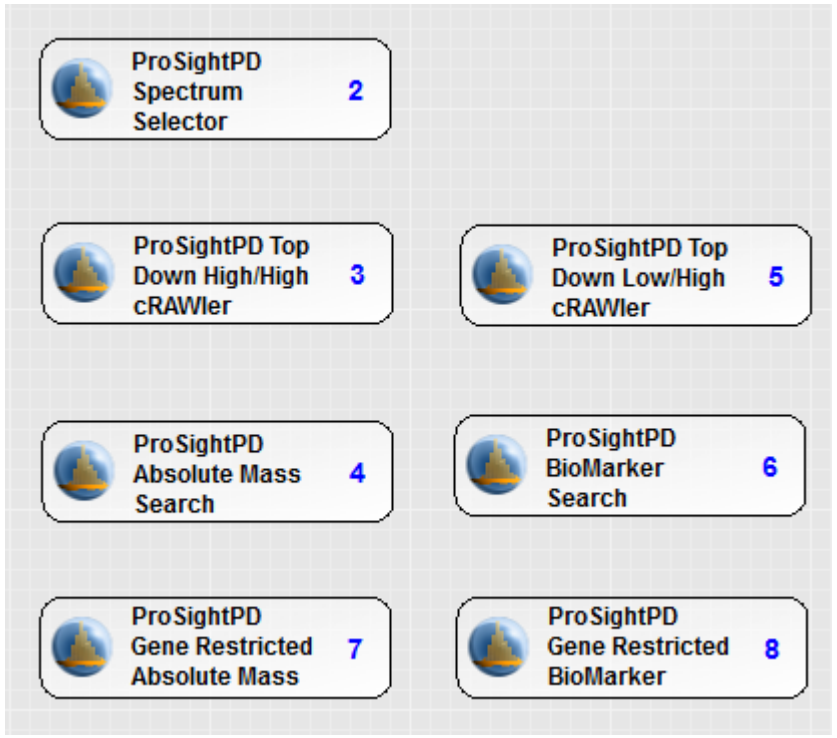
Protein Metrics



IMP Vienna



Proteinaceous/Northwestern

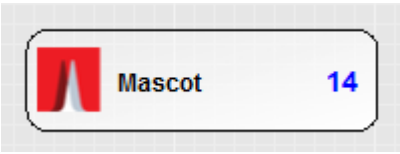


Nodes distributed only by external sites:

- pd-nodes.org (IMP Vienna)
- www.openms.de (PD 2.1)
- SuperQuant (PD 2.1)

NIST

Matrix Science

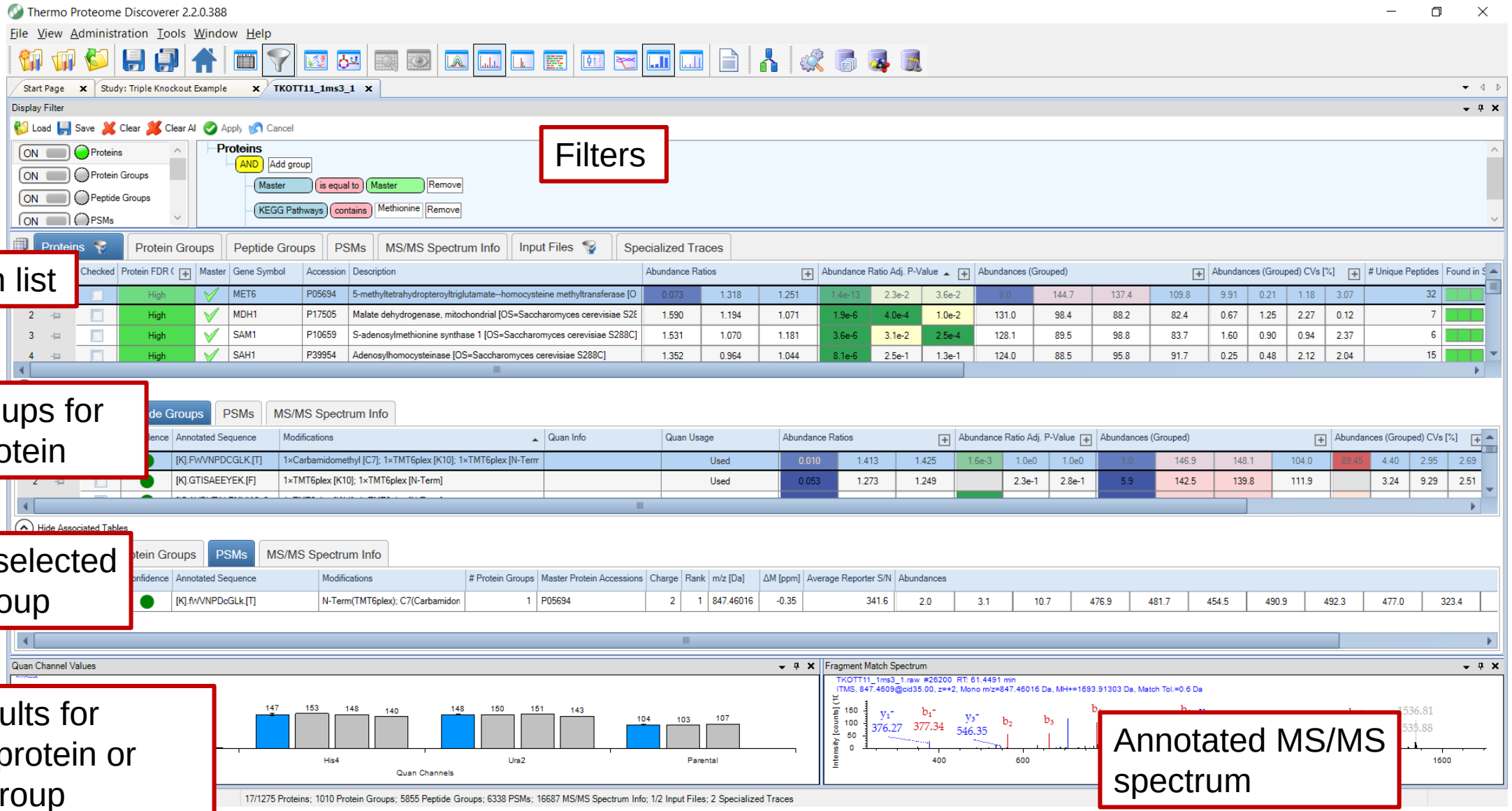


University of Utrecht

UW/KTH

ThermoFisher SCIENTIFIC

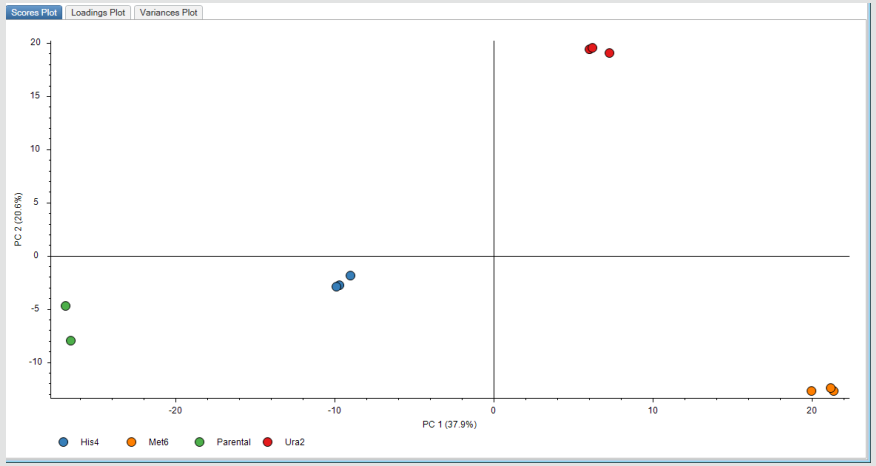
Proteome Discoverer Visualization – Hierarchical tables and row filters



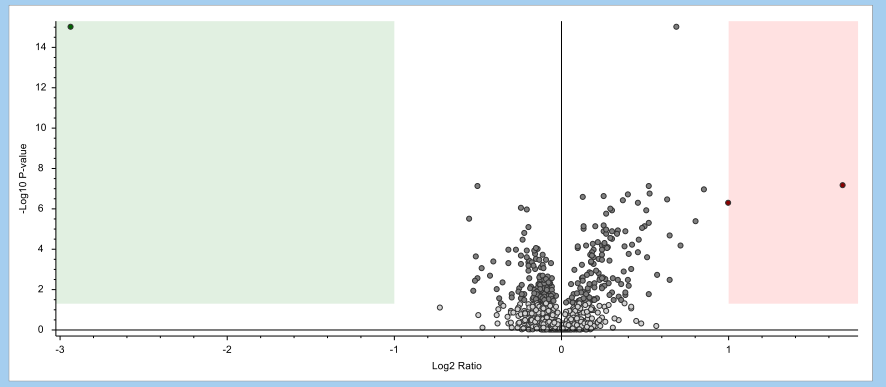
ThermoFisher SCIENTIFIC

Proteome Discoverer Tools for statistical and biological interpretation

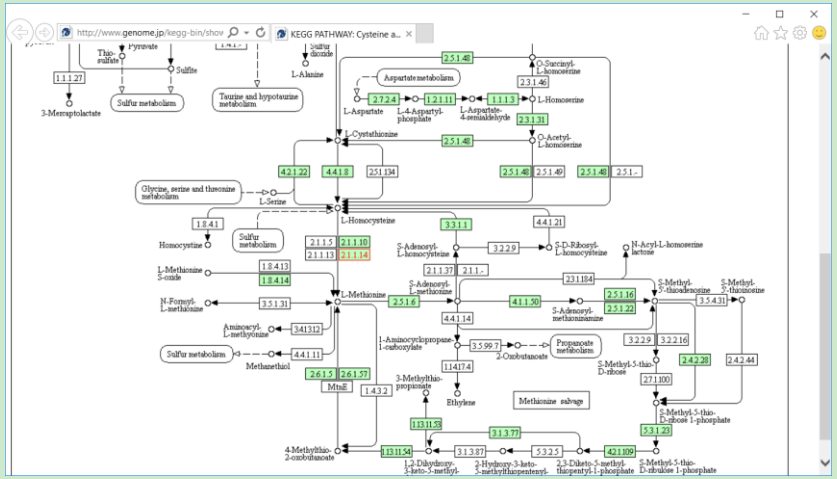
- PCA



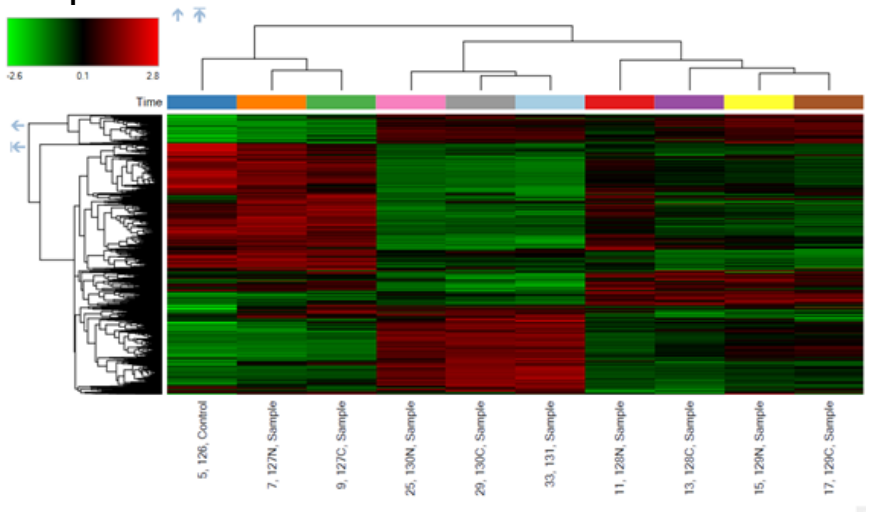
- Volcano plots



- Pathway maps



- Heat maps



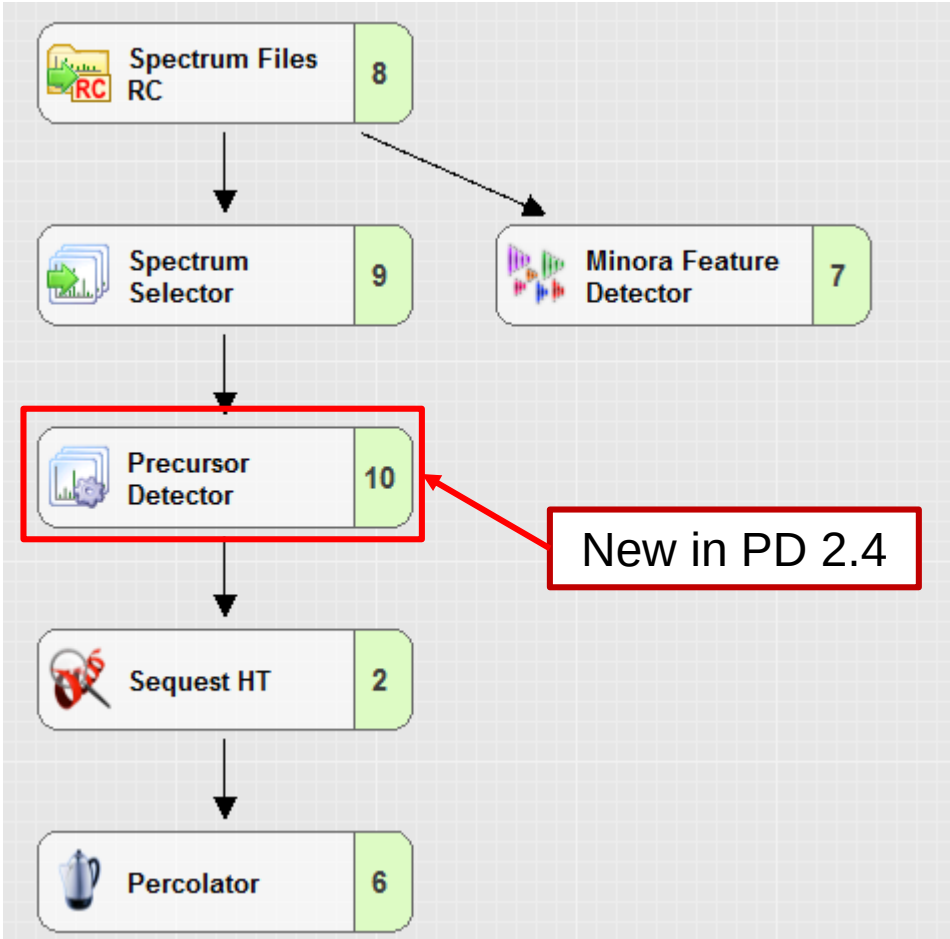
ThermoFisher SCIENTIFIC

Proteome Discoverer 2.4 Overview

- Mixed spectra/chimeric spectra support
  - Precursor Detector node
  - Works with Sequest HT and MSPepSearch search engines only
- Scripting node
  - Incorporate R, Python, or other scripts into Proteome Discoverer workflows
- New TMTpro 16plex method
- FAIMS LFQ – license no longer needed
- New Flexera licensing (like Thermo Scientific™ BioPharma Finder™ software)
- Updated Proteome Tools libraries coming soon
- Updated Familiarization Exercises

Proteome Discoverer 2.4 – Precursor Detector

- Spectrum Selector by default chooses isotopic cluster selected by instrument
- **Precursor Detector** node
  - Detects other isotopic clusters with peaks within the precursor isolation window
  - Duplicates MS/MS spectrum for each new precursor mass
  - Only input parameter: Input S/N
- Search engine can identify multiple peptides in the same MS/MS spectrum.
- Works **only** with Sequest HT and MSPepSearch
- Can increase proteins IDs up to 10-20%, peptide group IDs up to 15-30%



ThermoFisher SCIENTIFIC

Precursor Detector Node produces new Precursor Table

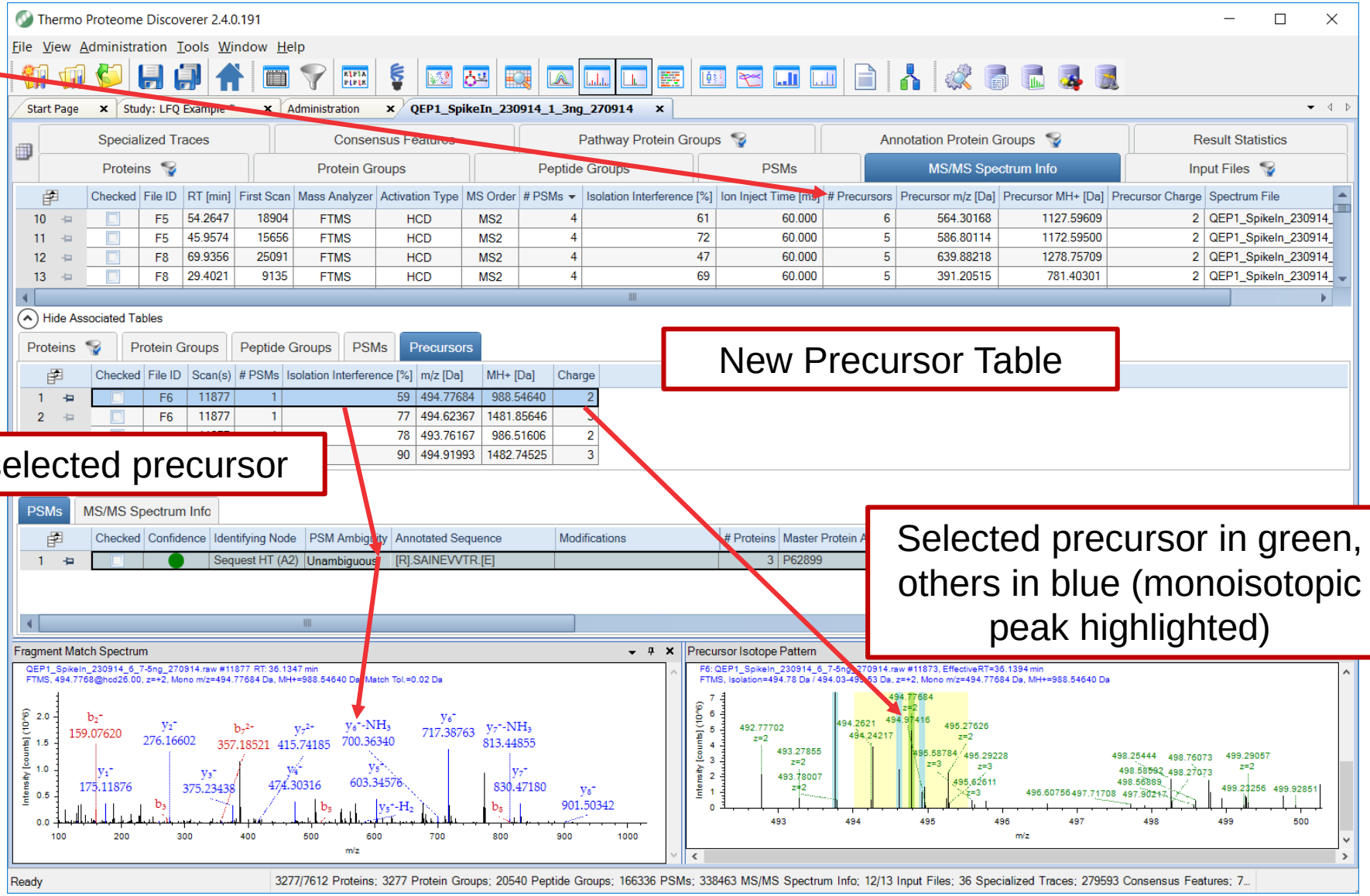
New # Precursors Column in MS/MS Spectrum Info table

Link to PSM for selected precursor

MS/MS spectrum with annotations for selected PSM

New Precursor Table

Selected precursor in green, others in blue (monoisotopic peak highlighted)

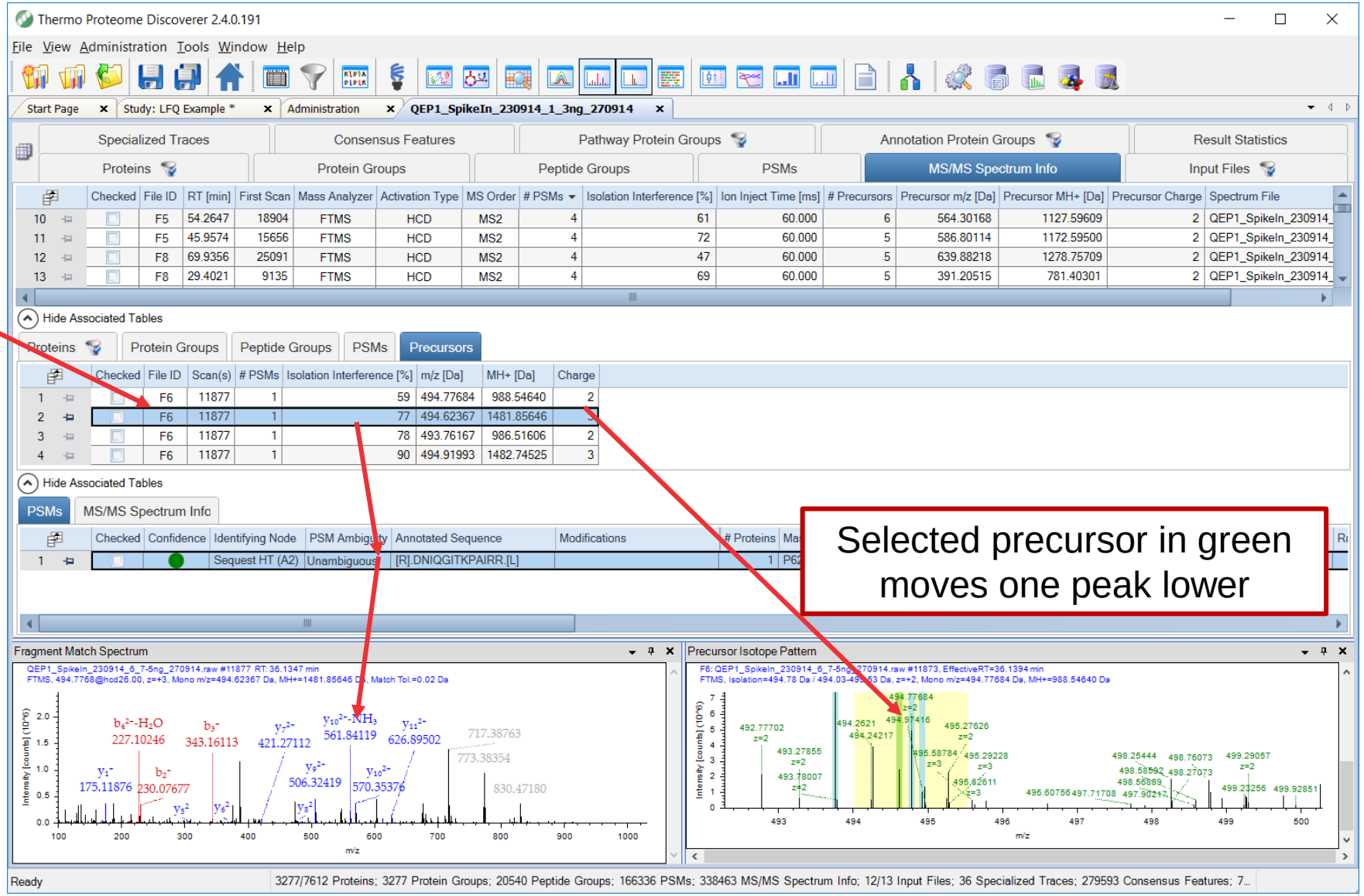


ThermoFisher SCIENTIFIC

New Precursor Table – Second Precursor Selected for same MS/MS spectrum

Second Precursor selected

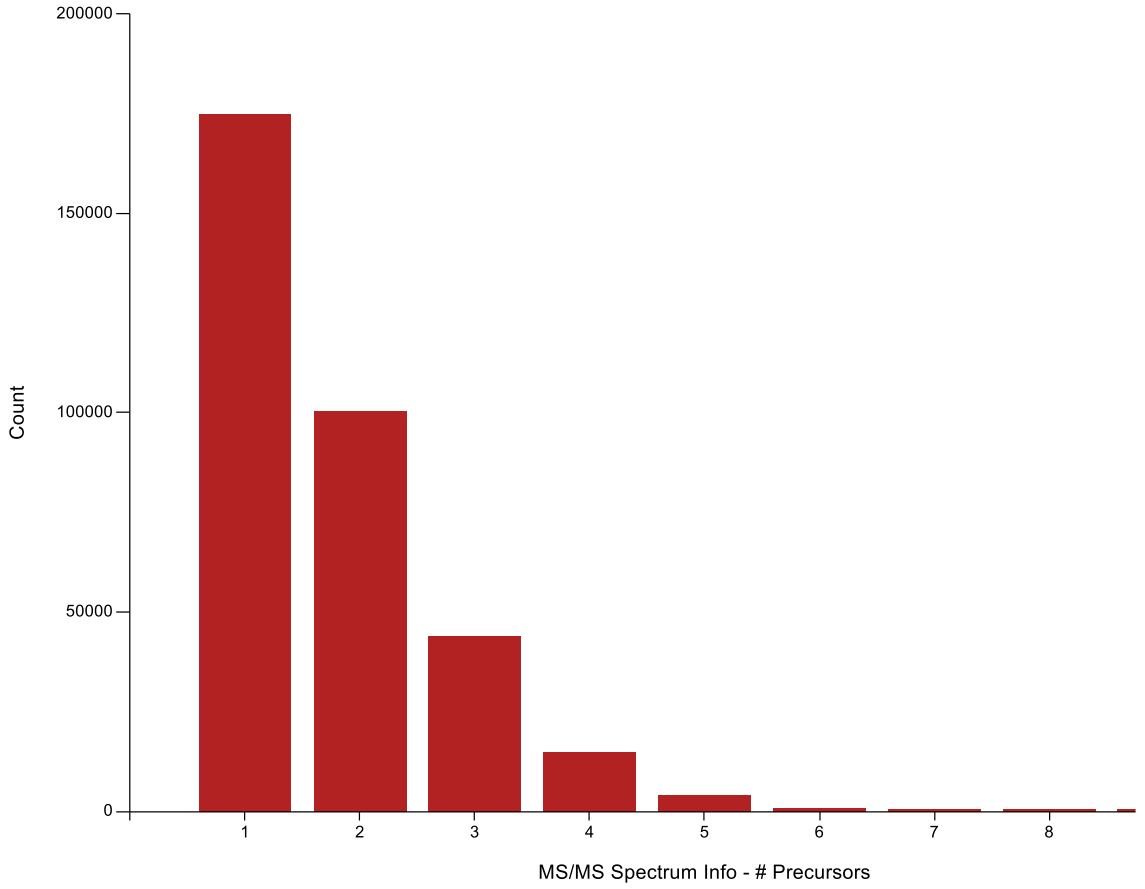
MS/MS spectrum annotated with peaks from selected PSM



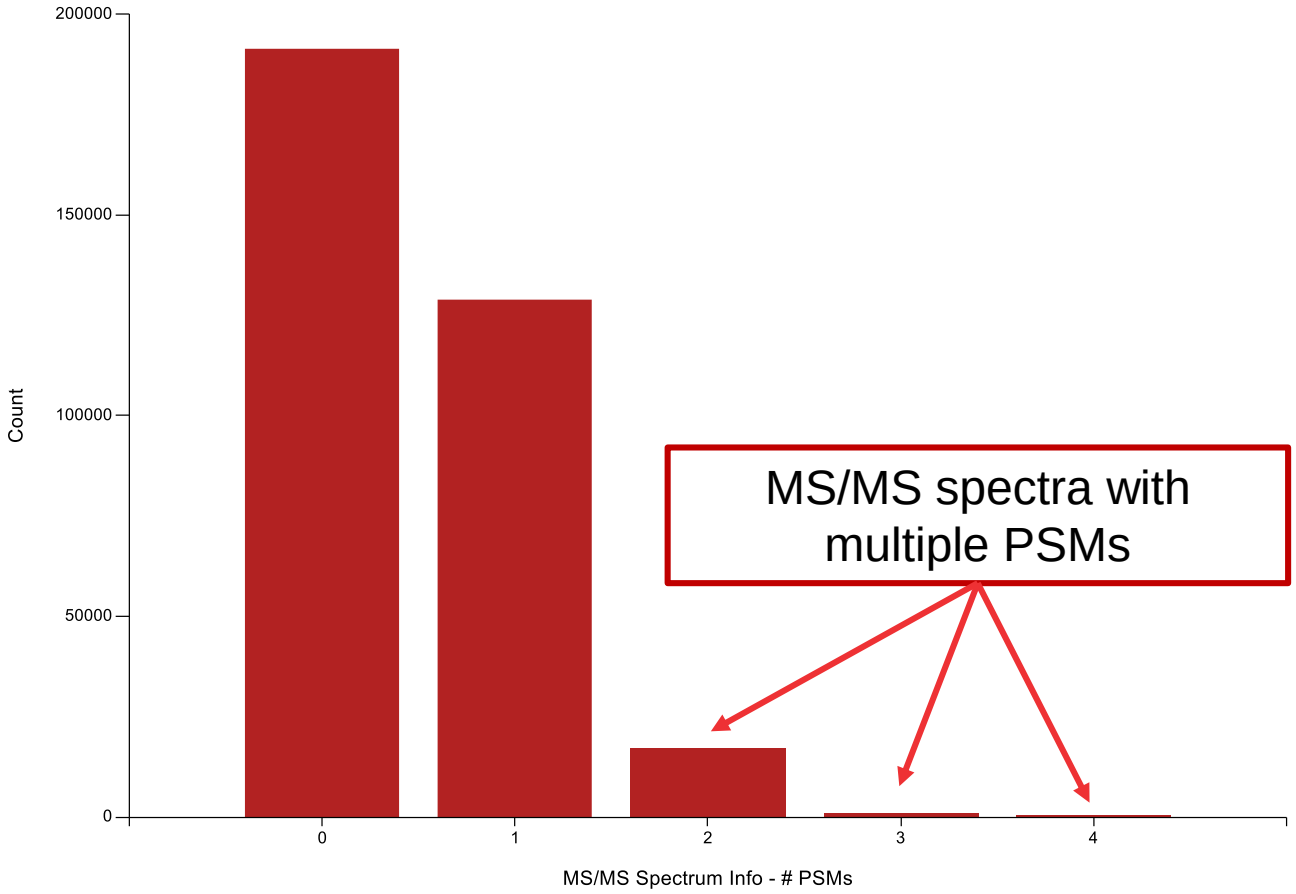
ThermoFisher SCIENTIFIC

Precursors and **Sequest HT** PSMs per MS/MS spectrum using Precursor Detector node

of Precursors Per Spectrum produced by Precursor Detector



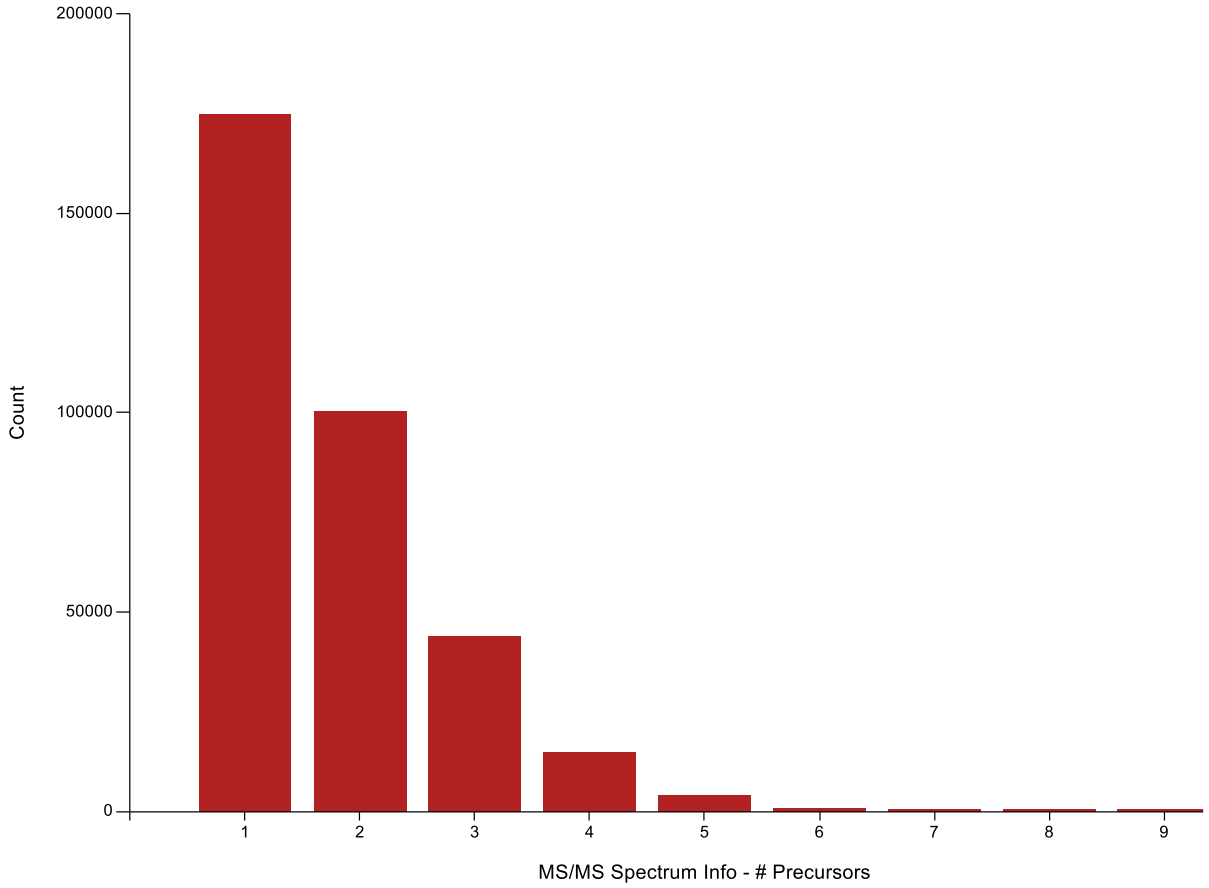
IDs per MS/MS spectrum for Sequest HT search



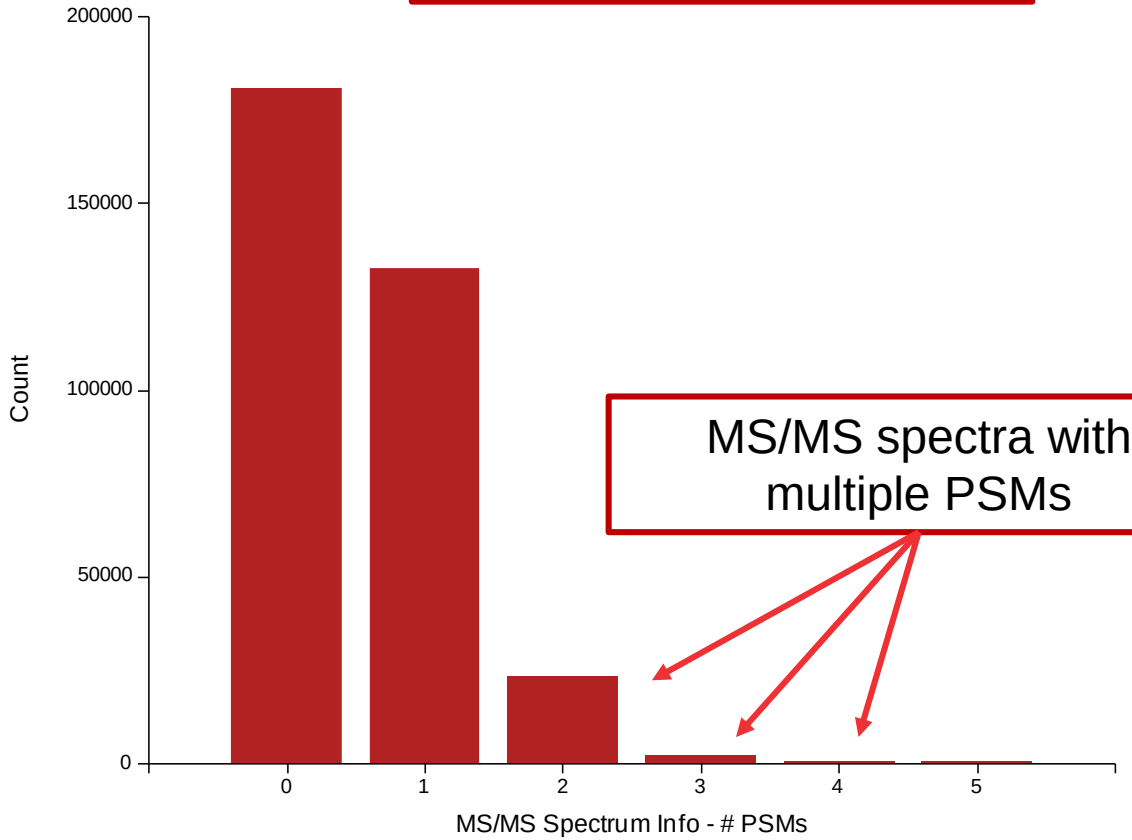
Data from Exercise 3 in Familiarization Guide

Precursors and **MSPepSearch** PSMs per MS/MS spectrum using Precursor Detector node

of Precursors Per Spectrum produced by Precursor Detector



IDs per MS/MS **MSPepSearch** search (NIST OT HCD library)



Data from Exercise 3 in Familiarization Guide

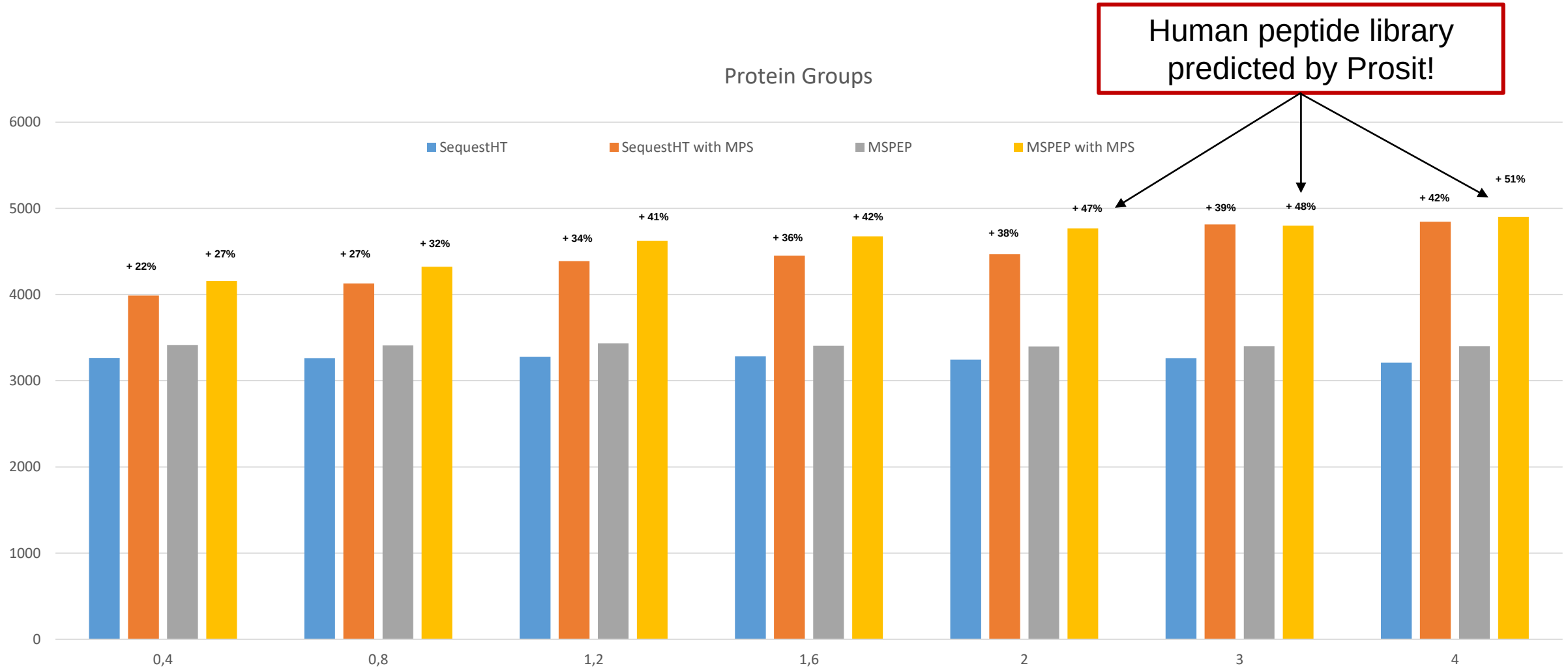
ThermoFisher SCIENTIFIC

LFQ Example data Familiarization Exercise 3 – With and Without Precursor Detector

	Sequest HT + Percolator			MSPepSearch (NIST OT HCD library) + Percolator		
	PD 2.3	PD 2.4 with Precursor Detector	Improvement	PD 2.3	PD 2.4 with Precursor Detector	Improvement
PSMs	129705	166336	**+28%**	135710	185457	**+37%**
Peptide Groups	17288	20540	**+18%**	17771	22500	**+26%**
Quantified Peptides	16250	19122	**+18%**	16659	20827	**+25%**
Proteins	2931	3277	**+12%**	2797	3233	**+16%**
Quantified Proteins	2757	3062	**+11%**	2607	3043	**+17%**

- Without Precursor Detector, Sequest HT and MSPepSearch perform similarly
- With Precursor Detector, MSPepSearch identifies ~10% more unique peptides than Sequest HT.
- MSPepSearch requires a comprehensive spectral library, which is currently only available for the unlabeled human proteome (NIST OT HCD)

Figure from Bernard Delanghe's ASMS 2019 poster (MP 414)

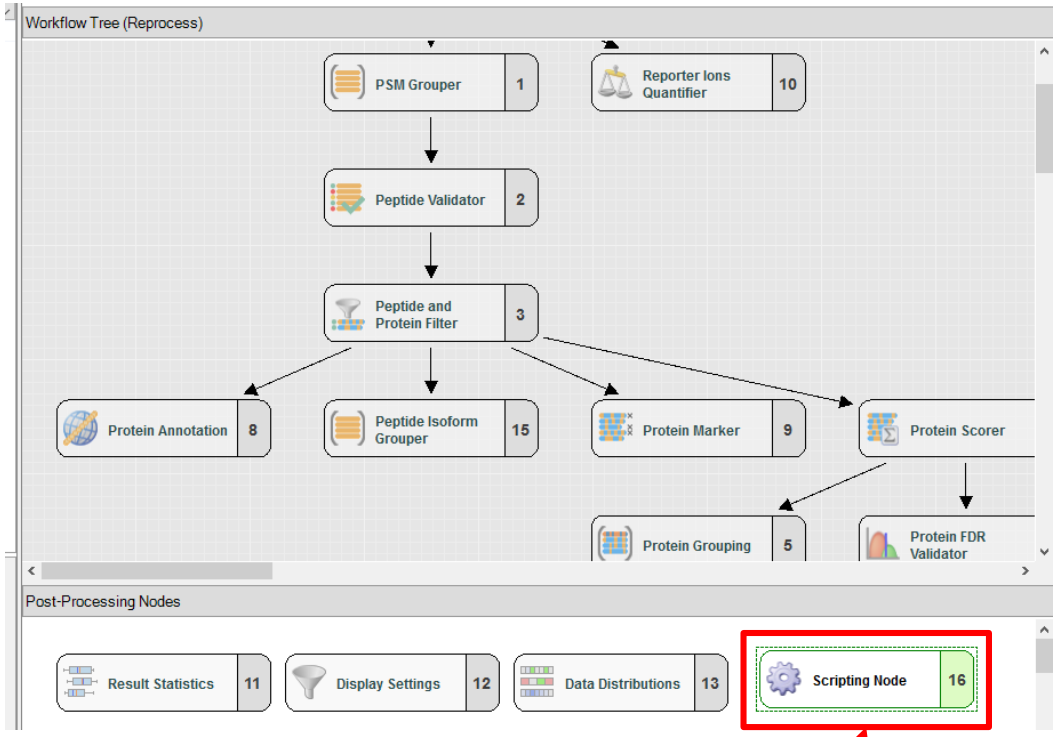


Use of Precursor Detector (MPS) leads to increased IDs with wider isolation windows!

ThermoFisher SCIENTIFIC

Proteome Discoverer 2.4 – Scripting Node

- **New Scripting Node** in Proteome Discoverer 2.4 and Thermo Scientific™ Compound Discoverer™ 3.1
  - Enable users to incorporate R and Python scripts (or any other executable) to perform custom data analysis
  - Installed as a Post-Processing node
  - Can also register custom scripting nodes as a Processing, Consensus, or Post-Processing node
  - Can access any information from any visible table
    - No access to MS/MS spectra or study information



New Scripting Node available by default as a Post Processing node

ThermoFisher SCIENTIFIC

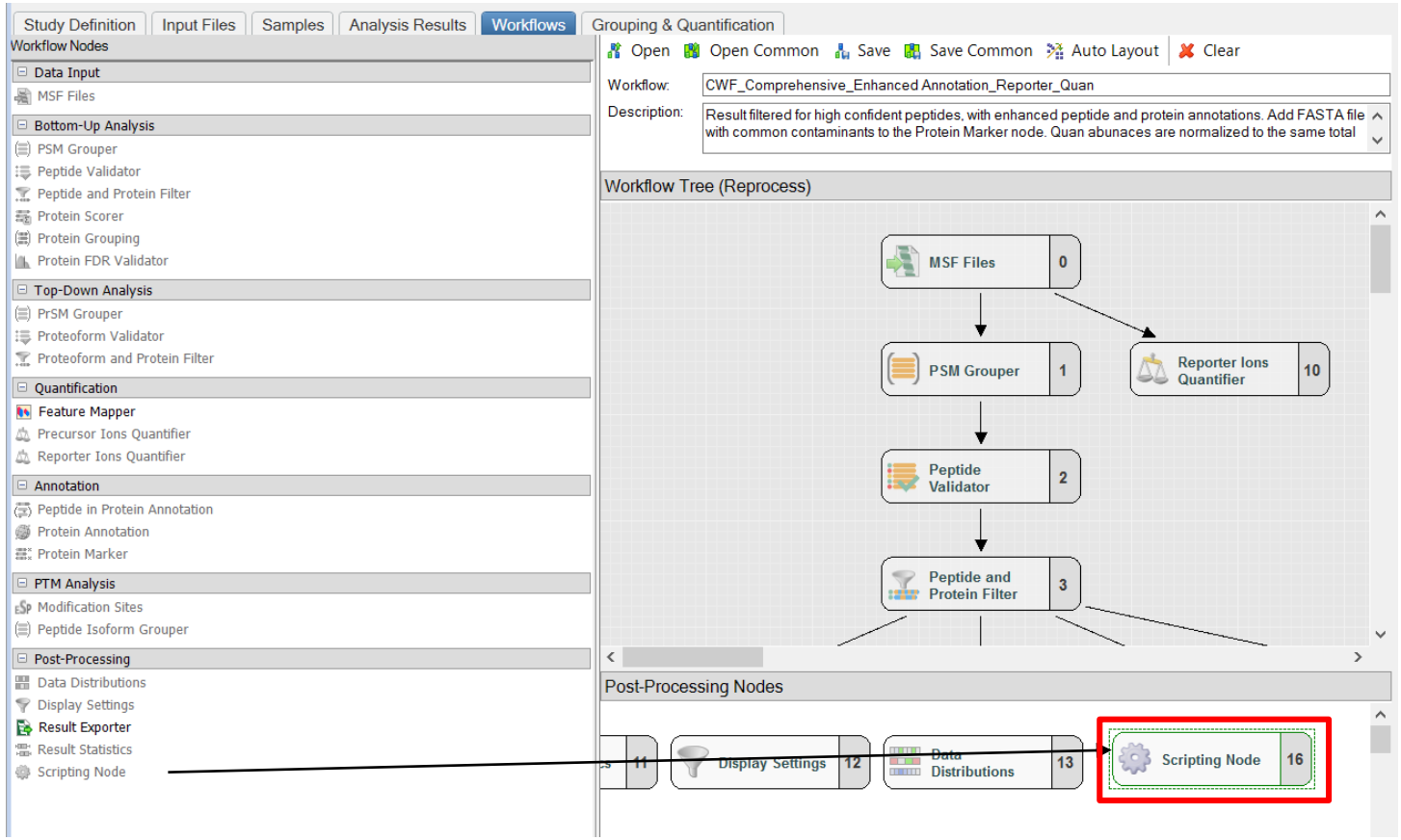
What about customers who know R, Python, Java, etc?

- In past releases, it is possible to create nodes in C# that call R, Python scripts or other executables
- These required learning C# .NET and using Visual Studio in past releases.
- **New Scripting Node** in Proteome Discoverer 2.4 and Thermo Scientific™ Compound Discoverer™ 3.1
- Enable users to incorporate R and Python scripts to perform their own type of data analysis
- Scripts have access to data from any of the result tables

- Caveats:
  - Scripting node for PD 2.4 does not have access to:
    - The raw mass spectra
    - The study information

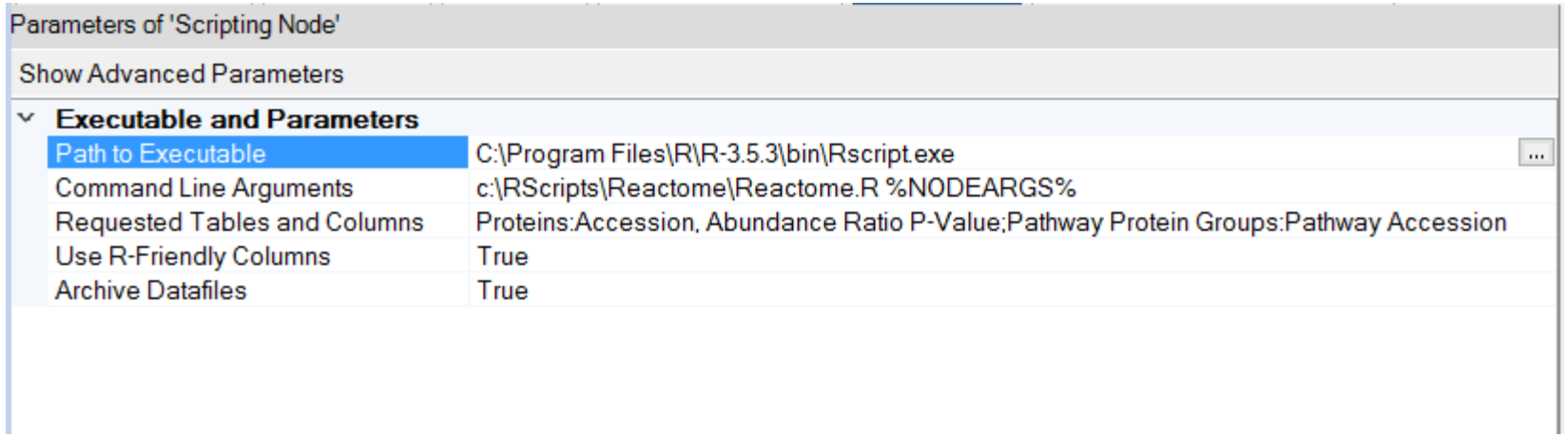
Scripting node – primary use will be as a post-processing node

- Scripting node in Proteome Discoverer 2.4 is shown as a Post-Processing node:



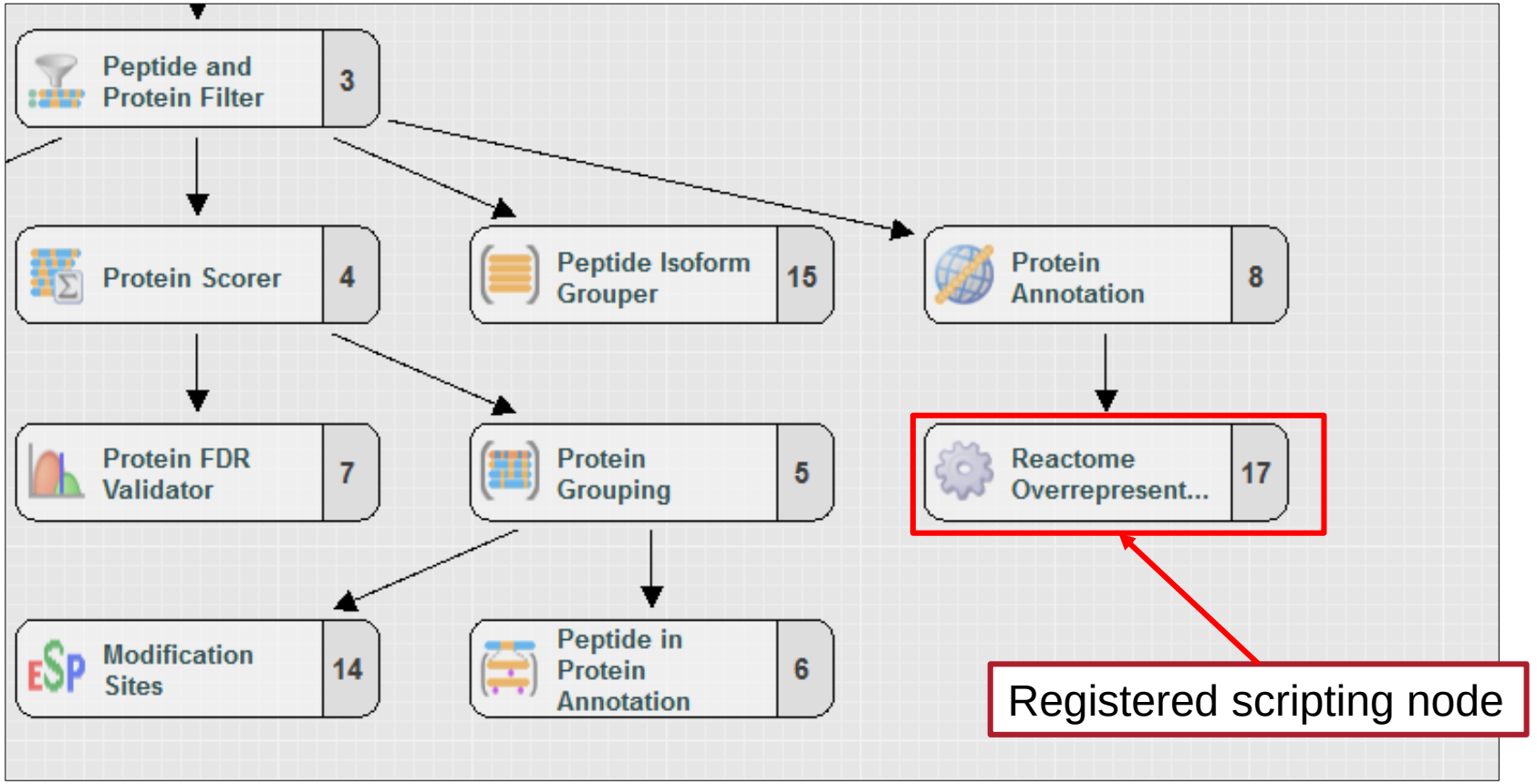
- It is possible to register a scripting node as a Processing or Consensus workflow node

Scripting Node parameters



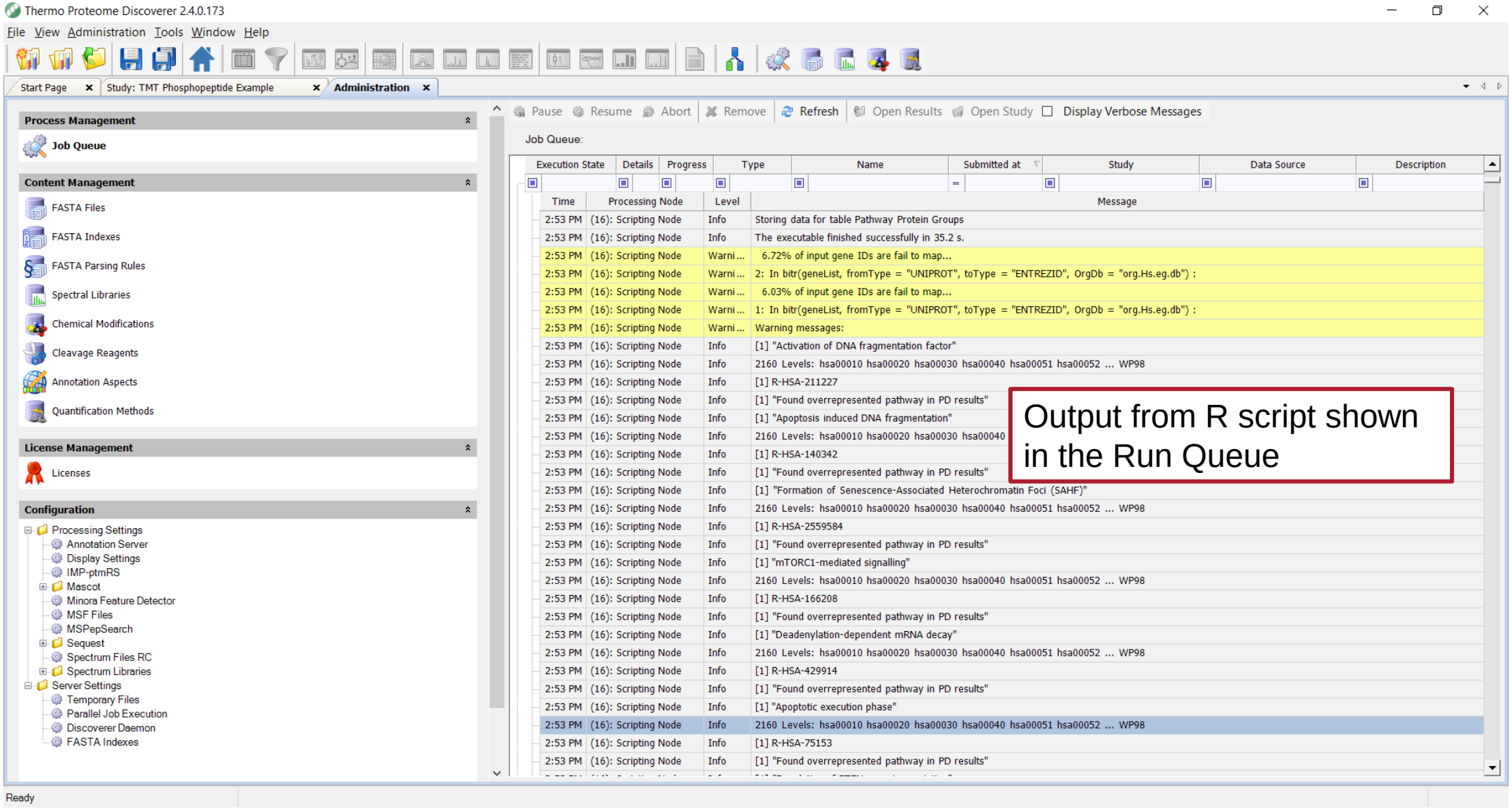
- Path to Executable – usually the executable for the scripting language (e.g. Rscript.exe) but any executable will work
- Command Line Arguments - Name of the script + what will be the link to the input data for the script
- Requested Tables and Columns –PD result tables and columns to be sent to the script
- Use R-Friendly Columns – removes special characters (e.g. #, /, (), []) from column headers
- Archive data files – saves a copy of files used by and produced by scripting node

Registering the scripting node for use in Processing or Consensus Workflows



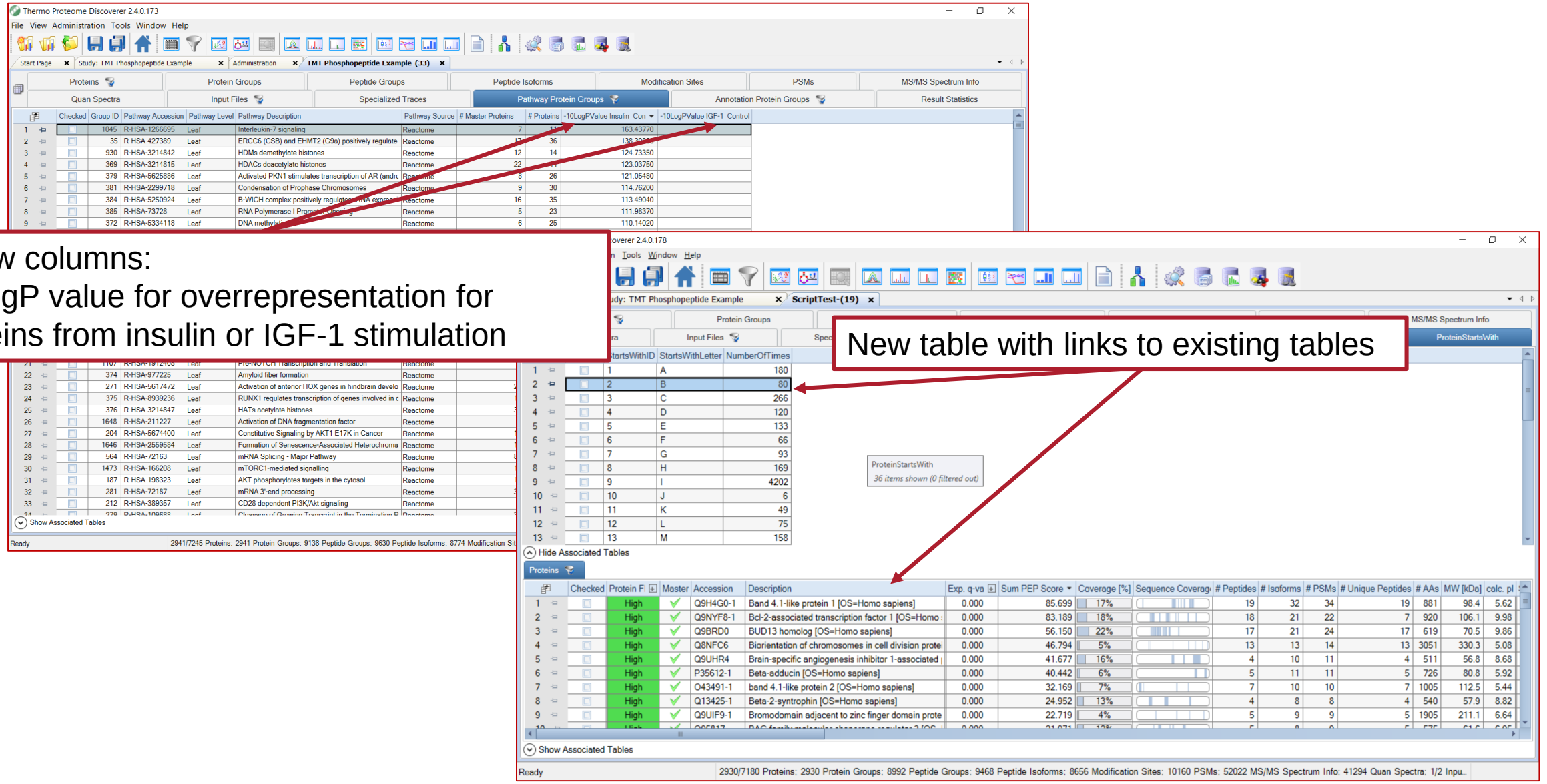
Registered scripting nodes are easier to transfer to other Proteome Discoverer installations

Scripting node in action – messages written by script appear in the Run Queue



Output from R script shown in the Run Queue

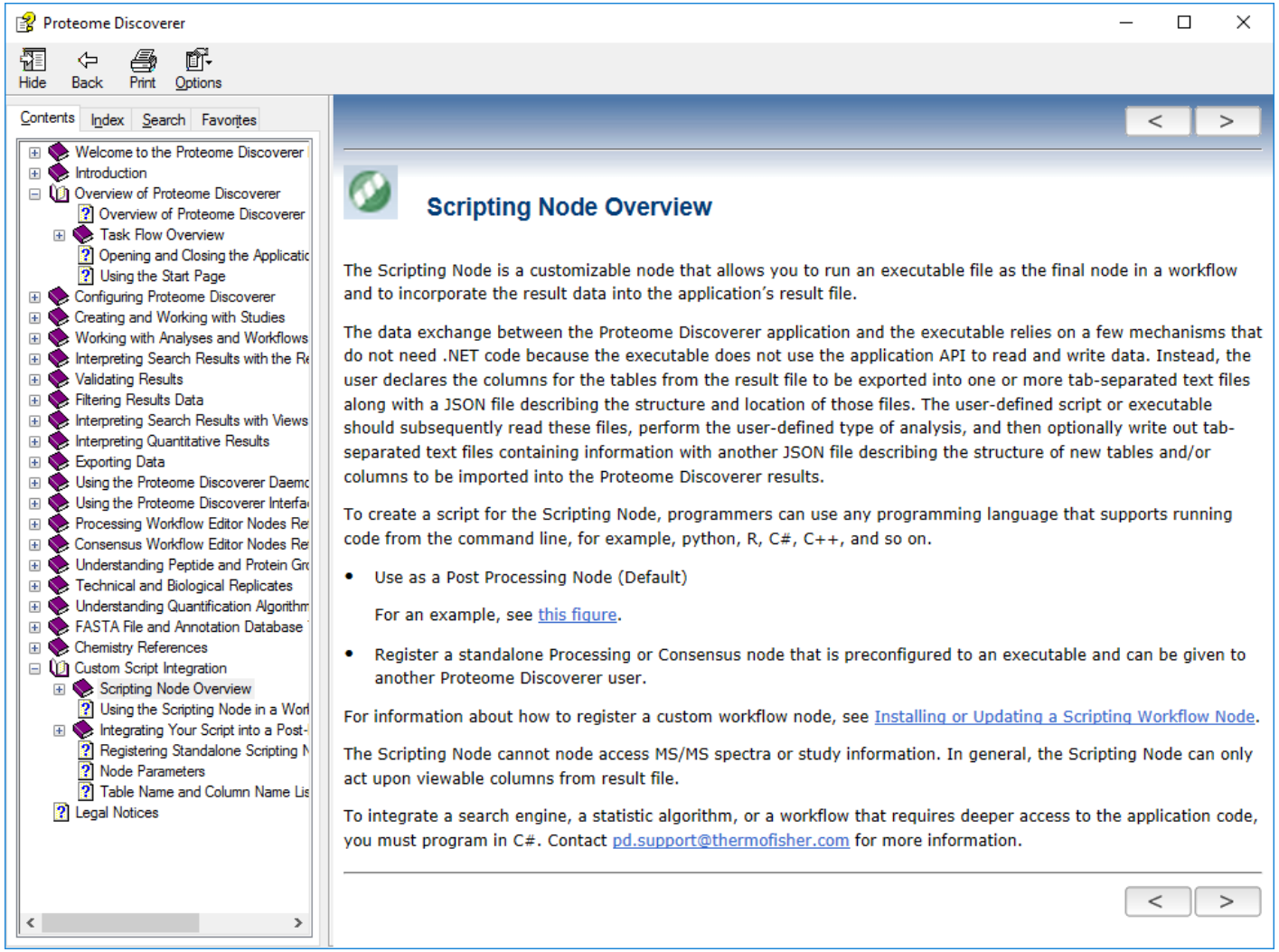
Scripting node can add new columns and tables to the result



New Scripting Node section in the Help

- New scripting node section in the Help explains how to create nodes from scratch
- PD media includes “User Scripts” folder with example scripts in R, Python, and Java with tutorials on how to run

Java	7/29/2019 4:37 AM	File folder	
Python	7/29/2019 4:37 AM	File folder	
R	7/29/2019 4:38 AM	File folder	
Study	7/29/2019 4:48 AM	File folder	
.project	7/16/2019 1:16 AM	PROJECT File	1 KB
README.txt	7/16/2019 1:23 AM	Text Document	1 KB
Scripting Node Custom Script Examples.docx	7/28/2019 10:57 PM	Microsoft Word D...	80 KB
selftest.bat	7/16/2019 1:16 AM	Windows Batch File	1 KB



Scripting Node Poster at ASMS 2019 by Frank Berg *et al* (MP434)

Implementing a Generic Scripting Node to a Standard Proteomics Workflow Processing Software

Frank Berg[1]; Carmen Paschke[1]; Kai Fritzemeier[1]; Pedro Navarro[1], Torsten Ueckert[1]; David Horn[2]; Bernard Delanghe[1], [1]Thermo Fisher Scientific (Bremen) GmbH, Bremen, Germany; [2]Thermo Fisher, San Jose, CA

ABSTRACT

Purpose: Implement an easy-to-use mechanism to enrich workflows with results of non-C# user algorithms in Thermo Scientific™ Proteome Discoverer™ framework.

Methods: Creating a family of preconfigured nodes as well as general mechanisms that integrate the calculation results of arbitrary external executables or scripts into Thermo Scientific™ Proteome Discoverer™ 2.4 software result files.

Results: We show by means of a custom R script that employs the widely used limma package [1] the integration of its results into Proteome Discoverer 2.4 software and use the additional statistical results of quantification data to compare them to the built-in Proteome Discoverer 2.4 software statistics algorithms. For this we use the rich set of plots and table presentations in Proteome Discoverer 2.4 software as well as R Studio.

INTRODUCTION

Proteome Discoverer software offers flexible analysis of proteomics mass spectrometry measurement data. Analyses are done by customizable workflows of configurable nodes that perform workflow subtasks, e.g., peptide identification, statistical validation or consolidation of protein findings. As of now, custom nodes may be implemented by third parties using a .NET programming language (typically C#) against the richly featured Proteome Discoverer API, thus extending the set of factory-provided analysis features. However, for rapid prototyping in context of, e.g., academic teaching or research contexts with compact and fast changing algorithmic ideas written in popular scripting languages like R or python this poses a certain cannon-on-sparrow situation.

Here we present a node family for PD that allows integrating arbitrary executables or scripts into an analysis workflow by using pre-implemented scripting nodes that adhere to a predefined data exchange protocol for external executables, thus providing an easy and fast method to extend workflows with user algorithms.

MATERIALS AND METHODS

The software was implemented within the Proteome Discoverer 2.4 framework using C#

Results

In principle our implementation offers to the user two ways of using a scripting integration in Proteome Discoverer:

- Predefined post processing scripting nodes for both the consensus and the processing workflow that only need a few parameters and an external script to be ready to go.
- Registration and creation of a custom standalone-node that follows the same principles as described above but additionally involves a registration process in PD. With this it appears as a an "ordinary" workflow node that can also be given away to other users in a standalone fashion.

We now describe the principal mechanisms of doing the data exchange between Proteome Discoverer and an external process as defined by our implementation in the post processing node. Further below, the mentioned registration process is outlined.

The post processing nodes involve the following basic parameters (Figure 1):

- **Path to Executable**: Location of the executable or script. If only a filename is given the system PATH environment variable is used to find it.
- **Command Line Arguments**: Any additional argument the executable needs.
- **Requested Tables and Columns**: A string that encodes all information about the data tables from the current PD result file that should be provided to the executable. The data is exported as a CSV text export.

Prior to executing the script the node provides the requested data in CSV tables and additionally stores a json file named "**node_args.json**" (Figure 2) containing meta information about the data.

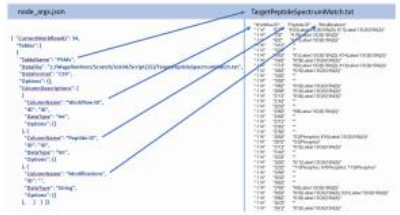
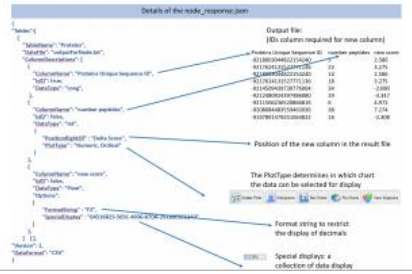


Figure 2. Example of exported Protein data from the PD result file as it is available to the scripting node.

- The following information is stored in file "**node_args.json**":
  - The path and name of every requested data table file
  - Type information about any column that is contained in the data table file
  - ID information, i.e., information about which columns are ID columns that are needed to insert new data into an existing table and to connect tables.

While the path information is vital to find the exported data, the type description is useful to parse the data values when importing them into a custom executable or script context.



After the external executable has calculated its results it may return data to the Proteome Discoverer result file. This is done by writing a very similar json file as "**node_args.json**" named "**node_responde.json**" (Figure 3) . This way the executable may perform the following actions:

- Add columns to existing tables (but not change existing columns/data)
- Add new tables with arbitrary data
- Add connections between tables. These connections may only be made between tables that are not related to each other yet.

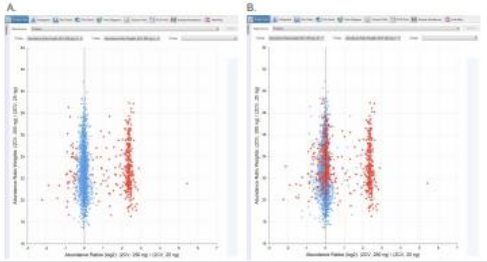
Case study: Integrating an R script for statistical analysis

To demonstrate the functionality of the scripting node, we implemented an R script using the limma package to calculate protein ratios and corresponding statistical values for a known mixture of human (HeLa) (1:1) and yeast (250ng : 25 ng) proteomes. Data reading and writing of json files (by using the RJSONIO package) are shown in the code snippets in (Figure 4 and 5). After running the scripting node, the data is available in Proteome Discoverer 2.4 and can be plotted using the plotting tools.

Figure 4. An example on how to import a table in R by using "node_args.json" properties. Here the table "Proteins" is imported, catching the index and corresponding abundances (by abundancesType) columns from the node_args.json file properties.

Figure 5. The response node json file can be written by using the "node_args.json" file as template. Store the new generated columns in a list, and substitute former list at $ColumnDescriptions of the corresponding Table ("Proteins" in this case). Change also the $DataFile of the Table to the outputFilePath.

To compare the two different calculations we display the max. Abundance between the samples (in log scale) versus log2 of the sample ratios. The proteins with a significant q-value (< 0.01) are highlighted in red. The plot of the Proteome Discoverer 2.4 calculated values is shown on the left in Figure 6, values calculated by the R-script using limma package are plotted on the right.



Custom Deployable Scripting Nodes

As mentioned above the user can deploy a custom version of her scripting node by registering a node in a standalone fashion. Necessary steps are outlined in Figure 8. A special definition file named "**node.json**" (Figure 7) needs to be provided that contains the following information:

- Name, Icon and target workflow (consensus or processing)
- Connection points that define where in the workflow the node can be placed
- Parameters to the node. Here standard parameters can be used that are known from ordinary PD nodes.
- All parameters needed for the scripting mentioned above.

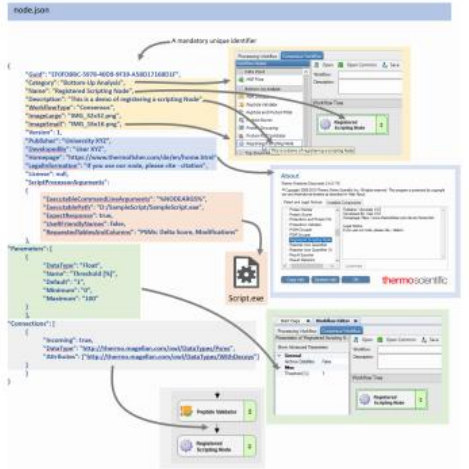


Figure 7. Example "node.json" file that defines the registration of a deployable scripting node

In the parameters section any types of parameters are available that can be used through the .NET API of Proteome Discoverer. The connections section describes a data contract that specifies the allowed connections of the scripting node to other nodes in the workflow.

Future Work

The current implementation of the scripting node mechanism involves some limitations that will be addressed in future versions of Proteome Discoverer.

1. An access to the study information that corresponds to the current analysis is not yet available.
2. Spectrum filters can not be implemented by scripting nodes.

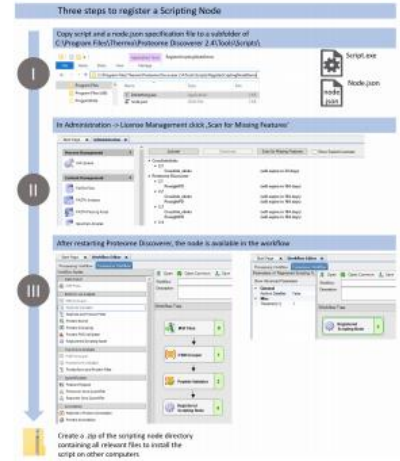


Figure 8. Three-step registration and deployment process for standalone scripting nodes in Proteome Discoverer.

CONCLUSIONS

We present a family of nodes that allow for rapid prototyping of proteomics algorithms in Proteome Discoverer™ 2.4. With these nodes the user can pass data to external executables or scripts and then import calculation results back into Proteome Discoverer. Moreover, the user can choose to define and register a deployable version of his scripting node for further distribution and sharing with collaborators. We demonstrate the usability by connecting the results of an R script that uses limma [1] to do statistics on a quantification workflow to Proteome discoverer™. For this we performed an analysis of proteomics data inspired by [2] and [3].

REFERENCES

[1] Ritchie ME, Phipson B, Wu D, et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

[2] Kai Kammers, R guide: Analysis of Cardiovascular Proteomics Data, http://www.biostat.jhsph.edu/~kkammers/software/CVproteomics/R_guide.html

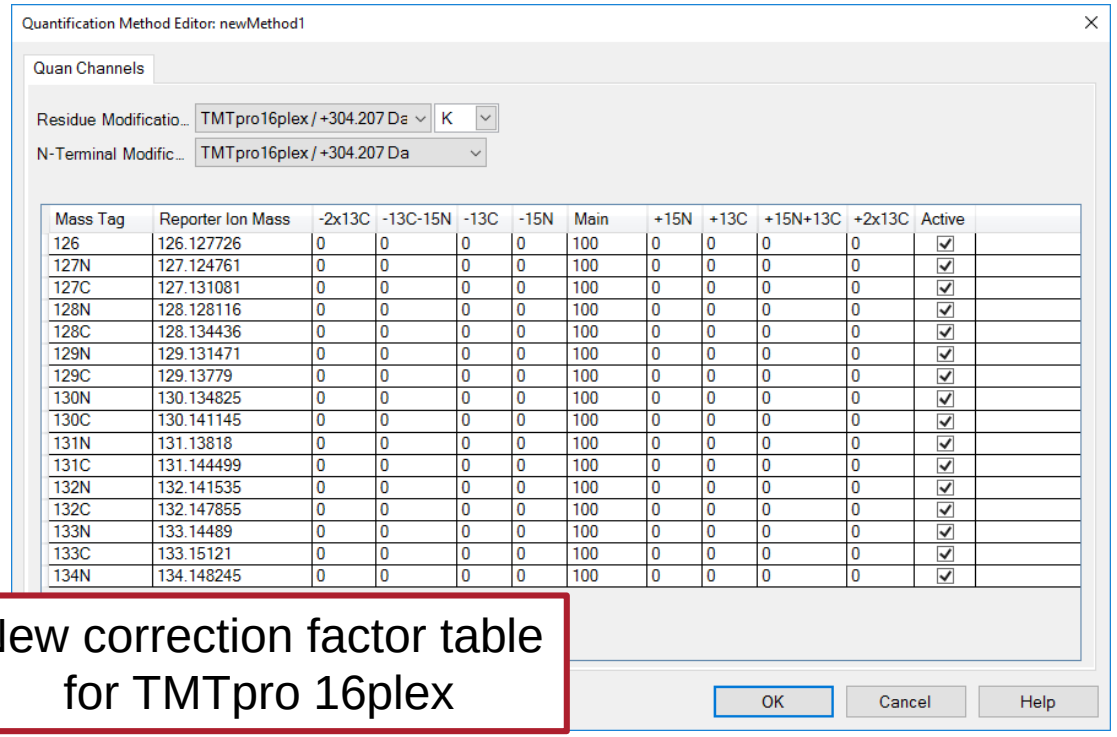
[3] Kai Kammers, D. Brian Foster, Ingo Ruczinski, Analysis of Proteomic Data, In: Manual of Cardiovascular Proteomics Pages 275-292, Springer International Publishing Switzerland 2016

TRADEMARKS/LICENSING

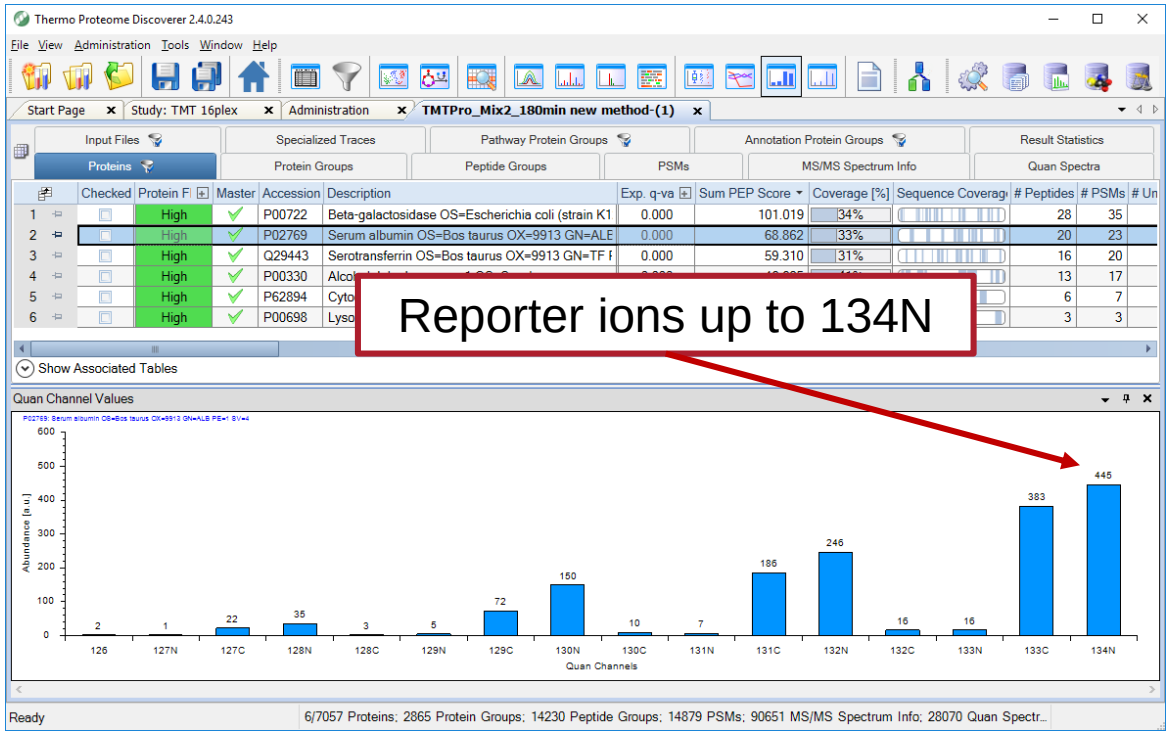
ThermoFisher SCIENTIFIC

TMTpro 16 plex support

- New TMTpro 16plex modification
- New table for TMTpro 16 plex correction factors
- New default workflows for TMTpro 16plex



New correction factor table for TMTpro 16plex



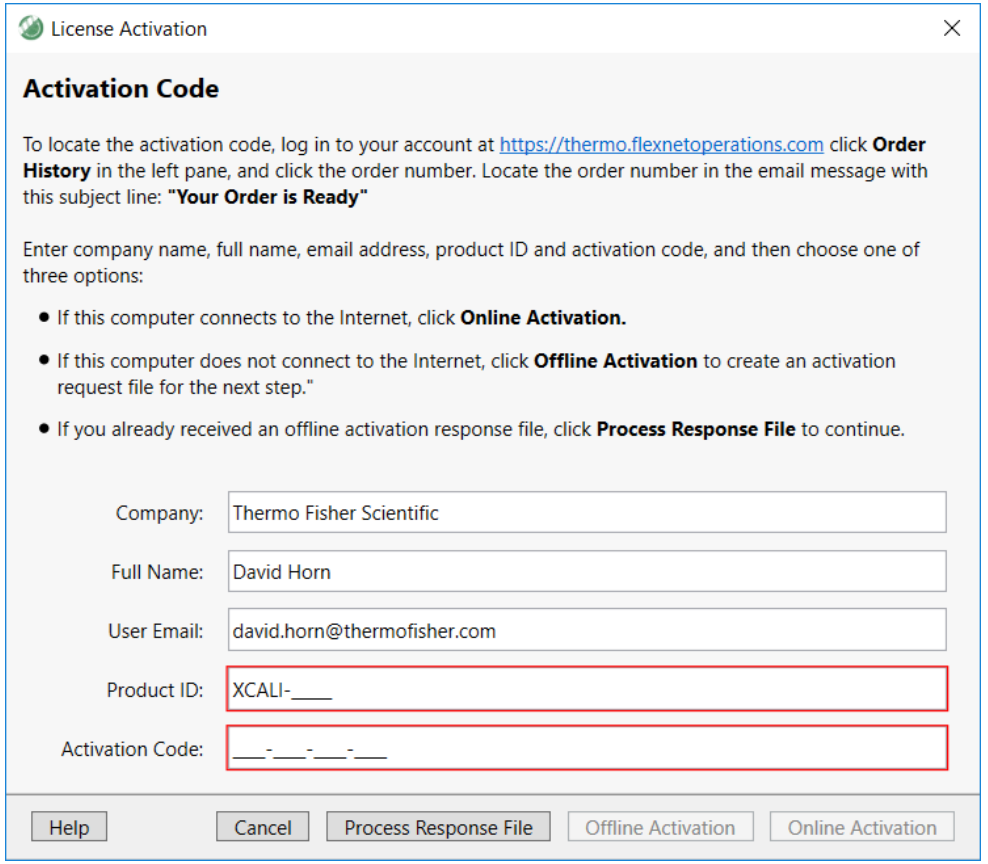
Reporter ions up to 134N

ThermoFisher SCIENTIFIC

Flexera licensing

- Same licensing scheme as BioPharma Finder
- Activation now requires a part number and license key
- Upgrades from previous versions with active maintenance are still free (see next slide for instructions on how to upgrade)
- License keys are sent by e-mail (no more lost licenses!)

- **No more maintenance license for PD 2.4**
- **Protein Annotation service is now tied to the Base license with no expiration**



ThermoFisher SCIENTIFIC

Procedure for customers with active maintenance to upgrade to 2.4

1. Install PD 2.4 on the same system as the previous version of Proteome Discoverer. PD 2.4 now automatically installs with a 60-day demo license without the need for a license key.
2. Send an e-mail to ThermoMSLicensing@thermofisher.com with the following information:
    1. Name
    2. E-mail address
    3. Institution
    4. Current active maintenance license key from previous PD installation (accessible from Administration->Manage Licenses, selecting the Discoverer Annotation license)
3. ThermoMSLicensing will send an e-mail to the e-mail address above with the license key
4. Open the Administration->Manage Licenses dialog and click Activate.
    1. For the Product ID, type XCALI-98057.
    2. For the Activation code, use the code provided by e-mail via ThermoMSLicensing.
5. If connected to the network, click Online Activation. If offline, click Offline Activation and follow the instructions to sending the license file to ThermoMSLicensing.

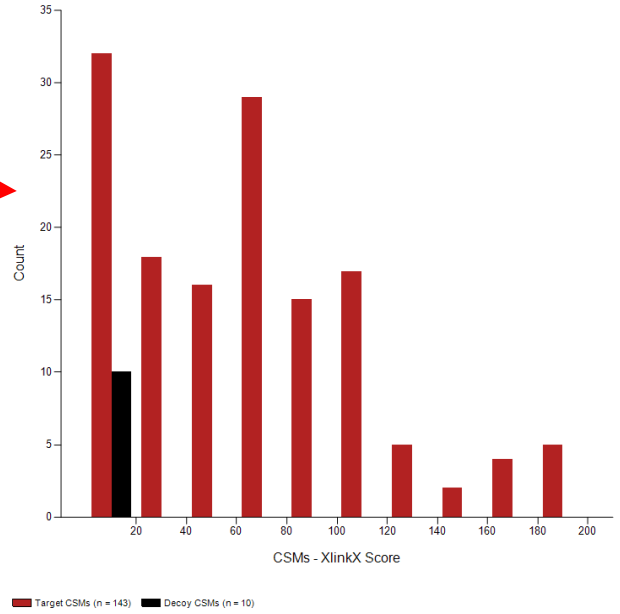
Proteome Discoverer 2.4 Third Party Node Installer

- MS Amanda 2.0 (IMP Vienna)
- Byonic and Preview nodes (Protein Metrics, Inc.)
- ProSightPD 3.0 (Proteinaceous, Inc.) – includes 60-day demo license
- XlinkX for PD 2.4 – includes 60-day demo license

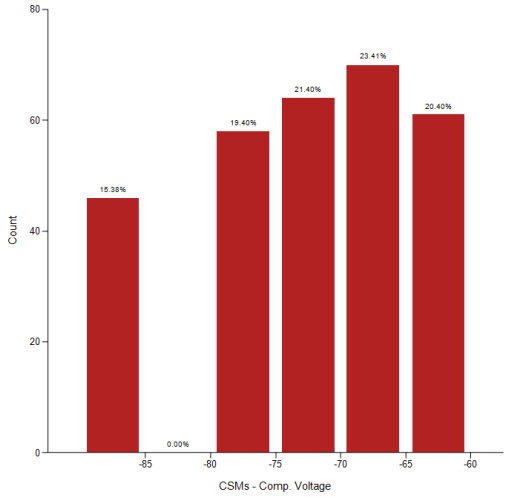
- The Byonic and Preview nodes require the associated standalone software to be installed and licensed on the same PC. The Proteome Discoverer installation media includes the latest installers for both and a 30-day demo license key.
- The ProSightPD 3.0 nodes requires the standalone Thermo Scientific™ ProSightPC™ 4.1 software to be installed. A demo version can be downloaded from www.proteinaceous.net.

XlinkX 2.0 for PD 2.4

- Now have a “target-decoy” view for XlinkX score to help confirm FDR calculation
- Added compensation voltage as a column in CSM table for FAIMS data.
- Upgrades from previous versions of XlinkX are free. The license is automatically transferred from previous versions of PD.



	Exclusive	Total	Label
A	9	22	Comp. Voltage: -60
B	3	16	Comp. Voltage: -80
C	7	28	Comp. Voltage: -70
B | C	8	13	Comp. Voltage: -80 | Comp. Voltage: -70
A | C	8	13	Comp. Voltage: -60 | Comp. Voltage: -70
A | B | C	5	5	Comp. Voltage: -60 | Comp. Voltage: -80 | Comp. Voltage: -70
Sum	40		



ThermoFisher SCIENTIFIC