Spectrum Mill Quick Start Guide

(Updated December 2021)

1. What is Spectrum Mill?

Spectrum Mill (SM) is a suite of tools for high-throughput processing of proteomics LC-MS/MS datasets to provide identification and relative quantitation at the protein, peptide, and modification sites (phospho, acetyl, ubiquityl, etc) levels.

SM directly processes LC-MS/MS data from Thermo Fisher Scientific and Agilent Technologies Inc. instruments. SM executes on a web server running a Windows operating system, and you access the programs via a web browser.

Spectrum Mill Core Capabilities include:

Spectral pre-processing
Directly from instrument vendor format files: Thermo Fisher, Agilent

Identification of peptide spectrum matches - PSMs by database search
PSM scoring capitalizes on low ppm product ion mass accuracies with fragment-ion type models optimized for orbitrap HCD, qToF-CID, ion trap CID, and ETD / ETHCD dissociation methods

PTM-site localization
For phosphorylation, acetylation, and ubiquitylation with multi-sample site-level consolidation that handles localization ambiguity

TMT-10, 11, 16, 18 / iTRAQ 4 reporter ion based quantitation
With PSM-level isotope correction and precursor ion purity filtering, and sample-level normalization

Protein-level assembly of PSMs
With attention to isoform-specific and species-specific peptides

Quality metric calculation/reporting
For MS acquisition, chromatography, spectral interpretation and FDR, digestion efficiency, PTM enrichment, and label incorporation

Automation on an SM server through a service request manager
Maximizes use of all CPUs, queues and processes tasks for multiple users via workflows with stored parameters

2. What's New in this version?

For brief description of new features introduced in each Spectrum Mill release see:
https://proteomics.broadinstitute.org/millhtml/SM_instruct/whats_new.htm

3. Where to Find More Information?

4. For Scientists: slide sets, online help, example data, saved default parameter sets

Slides section on SM home page lower left

- Spectrum Mill Overview
  https://proteomics.broadinstitute.org/millhtml/SM_slides/SpectrumMillOverview.pdf

Online Help links are consistently placed throughout SM pages:

- Far right side of menu bar at top of page
- Section headings in colored trim bars, link directly to help for that section

Example Datasets:

Located in SpectrumMill_releases shared folder on Cloud Platform along with SM download

- Thermo Fisher Scientific
  - ExampleData\Thermo\Exploris480_JurkatQC
- Agilent
  - ExampleData\Agilent\QTOF6550_Hela120

Saved Default Parameter Sets

Near the top of each SM form, press the Load button, then select the Defaults directory. See **Figure 1**.

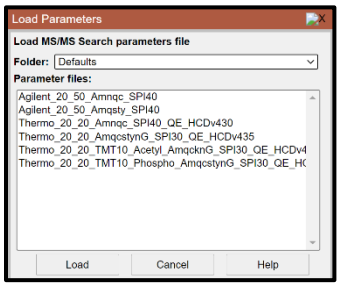

Figure 1 Load parameters for saved Default Parameter sets.

5. For System Administrators: Read these guides and online help

Links to the following are included on your local Spectrum Mill home page:

Installation Guide

https://proteomics.broadinstitute.org/millhtml/SM_instruct/SpectrumMillInstallationGuide.pdf

Configuring Custom Modifications Guide

https://proteomics.broadinstitute.org/millhtml/SM_instruct/SpectrumMillCustomModifications.pdf

Protein Sequence Databases

https://proteomics.broadinstitute.org/millhtml/SM_instruct/faman.htm

Server Administration

https://proteomics.broadinstitute.org/millhtml/SM_instruct/servadmn.htm

6. Set Up Spectrum Mill, Server and Client

7. To set up the server

See the Spectrum Mill Installation Guide. If you wish to update from a previous version of Spectrum Mill, see Step 0. Note that the server name cannot have an underscore.
https://proteomics.broadinstitute.org/millhtml/SM_instruct/SpectrumMillInstallationGuide.pdf

8. To set up your client PC

Spectrum Mill is web based so you interact with your SM web server from a computer with any operating system (in principle) and use a web browser. As of 2021, SM is primarily developed using Windows 10 and a Google Chrome Browser. The Lorikeet spectrum viewer runs by default from links in the scored percent intensity (SPI) column of PSM/peptide reports. SM's aged Java applet based Spectrum Viewer still works great, but annoyingly requires switching to use of Internet Explorer 11, which still comes standard with Windows 10 for legacy applications. After removing all other IE-dependencies from SM in 2017, performance was tested and found suitable with the following browsers: Edge, Firefox, Opera, and IE-11.

Internet Explorer 11 Compatibility View

You must disable Compatibility View to use Spectrum Mill. If pages do not display properly, it is likely that Compatibility View remains enabled (Microsoft default). If you get a warning to disable Compatibility View when you access your Spectrum Mill server, do the following:

1. In Internet Explorer, click **Tools > Compatibility View Settings**. If Spectrum Mill server is listed under **Websites you've added to Compatibility View**, remove it.
2. Clear the check box for Display intranet sites in Compatibility View (checked by default).

Cookies

Enable cookies so you as you move between SM pages, selected parameters shared between pages (like data directory) will be carried forward and you won't need to re-select each time

1. For detailed instructions to enable cookies for your browser, search help for **cookies**.
2. When using IE-11, if the drop-down menus in Spectrum Mill appear empty, you may need to add Spectrum Mill server as a trusted site. Search Internet Explorer help for **trusted site**.

Java (JRE) required on each client workstation to view spectra in Internet Explorer 11

Check that the Java Runtime Environment is installed. If not, Spectrum Mill installation provides a version of Java that you can install.

9. Transfer Raw data to the Spectrum Mill Server

Before you open the Spectrum Mill home page with a web browser, transfer Thermo Fisher (*.raw files) or Agilent (*.d folders) data to the SM server using Windows Explorer. Data files require no special preparation; you simply copy the files into an appropriate directory on the SM server, beneath the base SM data directory: **msdataSM/**

❖ **Note:** Spaces are not allowed in filenames, so replace them with a dash or underscore. The following characters are also not permitted: ()|, ; % < > ? + !. Only 1 dot (.) is allowed, for a suffix (e.g., myfile.raw or **myfile.d**).

Access msdataSM/ from the server

Navigate to the directory, possibly on the D: partition.

SpectrumMill/msdataSM/

If you don't know where to find your Spectrum Mill file system, ask the person who installed SM on your server.

Access msdataSM/ from a PC client

1. Make sure your administrator has shared the **msdataSM** folder and given you permissions. See *Installation Guide*.
2. Map the shared folder, **\\servername\msdataSM**, to a client PC. (Right-click **Computer** and select **Map network drive**. Follow the instructions.)
3. Make sure you have both read and write permissions for each subdirectory that contains your data files. (To check, right-click the folder and select **Properties**.) Make sure all user groups have full permissions. (Check the **Security** tab.)

10. Set up a sensible hierarchy of data directories

To make it easy to compare datasets and get the desired columns in subsequent result tables, it is important that you set up the appropriate directory structure for your data on the Spectrum Mill server. Think of data directories and subdirectories in terms of the hierarchy: projects, experiments, and samples. When later making result tables you have the choice of columns corresponding to either directories or files (LC-MS/MS runs), usually directories. See **Figure 2**.

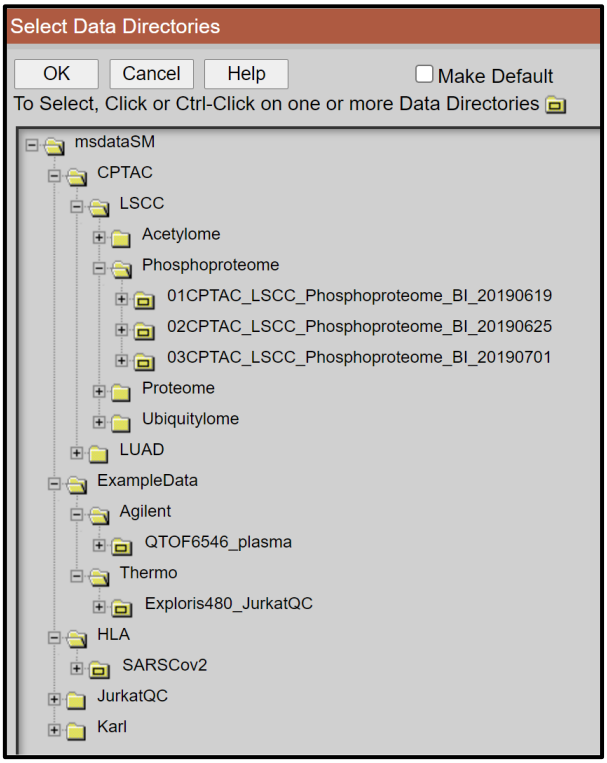

Figure 2 Example directory tree for Spectrum Mill data

Some general suggestions:

- Create top-level directories for each Project or Lab Member.

- If you have an Experiment in which you will compare samples in a set, you should create a directory for the Experiment and a subdirectory for each sample.
- A sample directory may contain multiple data files corresponding to off-line LC fractions or gel slices.
- Beware of "Bundler's remorse". Suppose you have 1 or more cohorts of samples, where each sample is a single LC-MS/MS run and you have so many samples that it will be a nuisance to make directories for each sample. So you decide to bundle all runs for a cohort into a single cohort-level directory and process them together. Later in P/P Summary you will be able to make a report with 1 column for each sample. During autovalidation PSM-level FDR thresholds can be applied separately to each sample or to the cohort as a whole, but you will not be able to use Protein Polishing or VM-site polishing to selectively remove single-sample protein subgroups or single-sample VM-sites (those modes work only across directories not files). Nonetheless, the remorse most-often comes when one or more samples are found to be unsatisfactory and you regret having included them in the cohort in the first place. If a sample is a directory, then you just repeat the autovalidation polishing step and the P/P Summary report generation step without including the undesired sample's directory. If the undesired sample is 1 of many files in a directory, you could choose to delete that sample's column from the final table, but the remaining analysis will still have depended on the presence of that sample during certain steps including autovalidation threshold setting, overall FDR calculation, protein grouping, and VM site collapsing. The same malady often afflicts those who find themselves with a leftover channel when designing a TMT-plex, and say to themselves, "let's tack on sample X, it will be a useful control".

Some specific examples:

- If you run an off-line fractionation method on your sample to generate factions, where each fraction is then analyzed by LC-MS/MS, transfer all the runs of fractions to a single directory on the SM server.
- If you run the same off-line fractionation method and LC-MS/MS analyses on fractions of a second sample, or if you repeat the run on the first sample, transfer all these files to a second directory on the SM server.
- If you conduct a differential expression quantitation experiment where you use SILAC labeling, transfer all the files to a single directory on the SM server. This is because all of the versions of the peptide (light, medium, heavy) are found in the same file or sets of files.
- If you conduct a label-free differential expression study where you will use the precursor ion abundances to do the quantitation, transfer samples from one cell state into one directory and the second cell state into a second directory on the SM server.
- If you conduct an immunopeptidomics experiment of the same sample with multiple replicate immunoprecipitations that each correspond to a subsequent LC-MS/MS run, transfer all the runs to a single directory on the SM server.
- You should have 3 directories of 6 runs each if you conduct a TMT10-plex phosphoproteome experiment with 27 samples, where you construct 3 plexes, each with 9 samples and 1 common control (a mixture of all 27 samples). On each plex you perform basic reversed-phase fractionation. After concatenation you have 6 fractions, which are each enriched by IMAC and the eluate of each enrichment is analyzed by LC-MS/MS.

11. Running Spectrum Mill

SM executes on a web server running a Windows operating system, and you access the programs via a web browser from a client computer.

Access SM from a PC client

Open a web browser window, type the address of your SM server. This should have the general format:

- server/millhome.htm
- If you don't know the URL, ask the person who set up your SM server.

Access SM from the server

Double-click the desktop icon to launch Spectrum Mill.

12. Process an Example Dataset to learn the basic tasks in Spectrum Mill

Please follow the process below with one of the Example datasets and sequence database provided in the Cloud directories

- SpectrumMill_releases/ExampleData copy to your SM server msdataSM/ExampleData
- SpectrumMill_releases/BenchmarkResults copy to your SM server msdataSM/BenchmarkResults
- SpectrumMill_releases/ExampleData/SeqDB/ copy to your SM server SeqDB

After copying the above files into place, use your web browser to navigate from the SM home page.

First click the **Protein Databases** link (in the lower **Utilities** section) to update the list of available sequence databases with the one you downloaded from: SpectrumMill_releases/SeqDB/

- Press **Update Database List**

Now begin to process an example dataset following the steps in the section below for either the Thermo or Agilent example. Start by clicking the **Data Extractor** link to execute task 1. Use the menu bar at the top of each page to access the page for the next task.

When using your own datasets the basic steps should still apply, though you should expect to revise and save parameters according to your own experiments.

❖ When using your own datasets the parameters most likely to be revised are:

- Sequence database
- Cysteine modification-fixed
- Modifications - variable

Pressing the **green button** on the page for an individual task, submits the task to SM's service request manager (SRM) queue. Hence, you do not need to wait for the prior task to finish before submitting the next task to the queue. All selected data directories will process in parallel, while the tasks for each directory will execute in serial order.

ExampleData\Thermo\Exploris480_JurkatQC

A single LC-MS/MS run of a tryptic digest of a lysate from the human Jurkat cell line analyzed with an Exploris 480 instrument at the Broad Institute. Cysteines were reduced/alkylated with DTT/iodoacetamide.

1. **Data Extractor** - extract the dataset
   a. Select the data directory: ExampleData\Thermo\Exploris480_JurkatQC
   b. Load the parameter file: Thermo_RAW_45similar_STL0_z5_600_6000_IAA_HCDv430.params
   c. Press **Extract.**
2. **MS/MS Search** - Search a database in **Variable Modifications** mode.
   a. Load the parameter file: Defaults\Thermo_20_20_Amnqc_SPI40_QE_HCDv430
   b. Update the database menu with example UniProt.human.20210902.RInrNF.602contams.fasta
   c. Press Save As and Save the default parameters with the example database.
   d. Press **Start Search**
3. **Autovalidation –** autovalidate the PSMs at a specified FDR using **Auto thresholds** strategy, with **Peptide** mode.
   a. Load the parameters file: Defaults\peptide_auto_z25_ppm_MSL6_1_0
   b. Press **Validate Files.**
4. **Quality Metrics** – compare your results to those provided as a Benchmark to verify performance of your SM installation.
   a. Load the parameters file: Defaults\FDR_ppm_z_PIP_PAU_id_chrom_digest_mods
   b. Select both data directories
      i. ExampleData\Thermo\Exploris480_JurkatQC
      ii. BenchmarkResults\Thermo\Exploris480_JurkatQC
   c. Press **Report.**
      i. Inspect the results by loading into Microsoft Excel the file:
         - ExampleData\Thermo\Exploris480_JurkatQC\qualityMetricsExportDir.1.ssv
      ii. Or Inspect the results in the lower pane if before pressing **Report** you have unchecked the boxes to the right of the load button for both **Queue request** and **Excel Export.**

ExampleData\Agilent\QTOF6550_Hela120

A single LC-MS/MS run of a tryptic digest of a lysate from the human HeLa cell line analyzed with a QTOF 6550 instrument at Agilent. Cysteines were reduced/alkylated with DTT/iodoacetamide.

1. **Data Extractor** - extract the dataset
   a. Select the data directory: ExampleData\Agilent\QTOF6550_Hela120
   b. Load the parameter file: Defaults\Agilent_D_60similar_STL0_z5_600_6000_IAA
   c. Press **Extract**
2. **MS/MS Search** - Search a database in **Variable Modifications** mode.
   a. Load the parameter file: Defaults\Agilent_20_50_Amnqc_SPI40
   b. Press **Start Search**
3. **Autovalidation –** autovalidate the PSMs at a specified FDR using **Auto thresholds** strategy, with **Peptide** mode.
   a. Load the parameters file: Defaults\peptide_auto_z25_ppm_MSL6_1_0
   b. Press **Validate Files**
4. **Quality Metrics** – compare your results to those provided as a Benchmark to verify performance of your SM installation.
   a. Load the parameters file: Defaults\FDR_ppm_z_PIP_PAU_id_chrom_digest_mods
   b. Select both data directories
      i. ExampleData\Agilent\QTOF6550_Hela120
      ii. BenchmarkResults\Agilent\QTOF6550_Hela120
   c. Press **Report.**
      i. Inspect the results by loading into Microsoft Excel the file:
         - ExampleData\Agilent\QTOF6550_Hela120\qualityMetricsExportDir.1.ssv
      ii. Or Inspect the results in the lower pane if before pressing **Report** you have unchecked the boxes to the right of the load button for both **Queue request** and **Excel Export**.

Further self-guided activities for either example

1. **Protein/Peptide Summary** - Summarize the results to file (Protein/Peptide Summary page).
   a. Load a parameters file (press the category select button to clear Broad Institute - specific values and select gene symbols for the example database)
      i. Defaults\PSM_export (all autovalidated PSMs)
      ii. Defaults\PeptideDistinctCS_Export (PSMs collapsed to peptide level – separate modified peps)
      iii. Defaults\PeptideDistinctCI_Export (PSMs collapsed to peptide level)
   b. Press **Summarize**
   c. Inspect the results by loading into Microsoft Excel the file(s) created in the data directory on the SM server
   d. Before pressing **Summarize,** unchecking the boxes for both **Queue request** and **Excel Export** will direct the summary back to a web page. Then clicking the link in the SPI column will launch an interactive spectrum viewer. Adjust filtering and sorting settings in the middle of the P/P Summary form to limit the number of spectra reported.
2. **Autovalidation -** with Protein Polishing mode remove low scoring single-peptide protein groups
   a. Load the parameters file: Defaults\proteinPolishing_both_minDir1_minScore0_protFDR0
   b. Press **Validate Files**

Online Help links are consistently placed throughout SM pages:

- Far right side of menu bar at top of page
- Section headings in colored trim bars, link directly to help for that section

13. Run Workflow Automation or Individual Task Automation

After becoming familiar with running the individual tasks involved in basic data processing and working with saved parameter sets, you may then prefer to save the set of individual tasks as a workflow and run the whole workflow with the single-click of the **green Execute button** on the workflow page instead of starting each task separately. Once you initiate a workflow, the Spectrum Mill service request manager (SRM) executes the individual tasks for a data directory in serial in the proper order. If you initiate multiple data directories, they execute in parallel.

Look for the Process Automation Tools section at the top of the Spectrum Mill home page. Thorough documentation is available in the Manuals section on the left side of the SM home page. Click Workflow Automation.

Workflow automation uses the Workflows page, an Edit Workflow page, and the Request Queue/Completion Log viewer.

Central to the design of workflow usage in SM is that whether executing individual tasks manually or as an automated workflow, one can use the same user interface for parameter review and the same parameter files. The Workflows and Edit Workflow pages let you view an individual task's page in read-only mode, with the parameters shown for the parameter file you select.

Set up workflow automation with parameter files

In order to run a workflow you must first create a workflow file, which is a simple tab-delimited file listing the tasks and their parameter filenames. See example below.

workflow.phospho_TMT11_v4_Gencode_EMAA.tsv

requestScript	paramsFile
runXtractor.pl	CPTAC3human\xtractor.TMT11fullLysonly_HCD_v4_35_Xcent_800_6000_45sec_z6.params
batchTagPara.pl	CPTAC3human\mstag.TMT11FullLysOnly_Phospho_v4_CU_AmqcstynG_Gencode.params
validateTable.pl	CPTAC3human\autovalidation.CPTAC_peptidevalidation_z24_acrossRun_MSL7_0_8_BCS3.params
validateTable.pl	CPTAC3human\autovalidation.CPTAC_peptidevalidation_z56_acrossDir_MSL7_0_4_BCS3.params

In order to set up a workflow you must first save the parameters for each task in the workflow using the web page for each task. Most Spectrum Mill pages let you save and load parameter files. Data Extractor, MS/MS Search, Autovalidation, and Protein/Peptide Summary include buttons to save settings in parameter files, and to load settings from parameter files. You use the parameter files, whose filenames are called tasks in the Workflows page, to build and execute workflows.

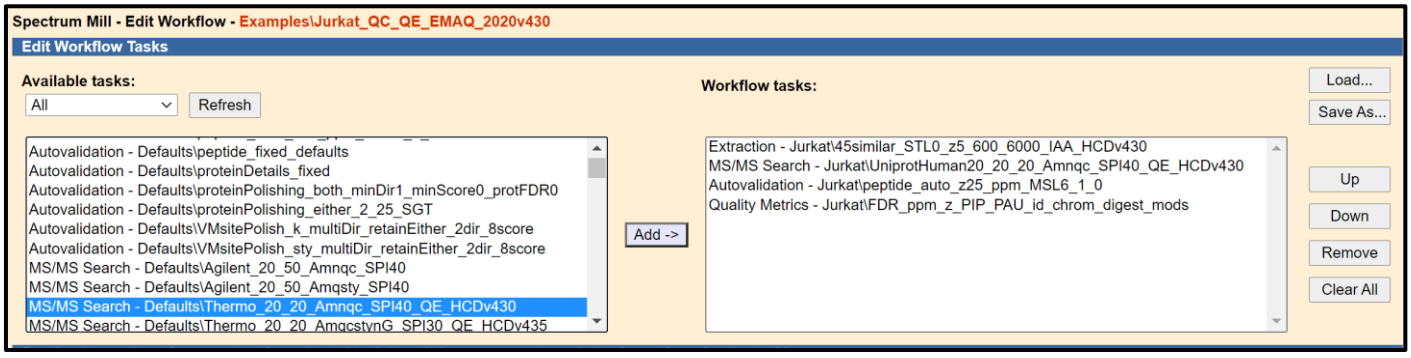

Figure 3 Edit Workflow user interface – upper Edit pane.

To create a workflow file:

- Visit the Workflows page (click the Workflow link near the top of SM home page).
- Click the Edit Workflows button.
- Use the user interface within the Edit Workflow page to select task parameter files from the Available tasks pane on the left. See Figure 3.
- Click the Add button to append to the list in the Workflow tasks pane on the right.
- Use the Up and Down buttons to re-order the Workflow tasks as desired.
- The lower Preview pane should show an individual task's page in read-only mode, with the parameters loaded for the currently selected task's parameter file.
- Click the Save As button to save the Workflow.

For Process Report, Peptide Selector, and MRM Selector, it is convenient to save parameter files, but you do not use them in an automated workflow.

Execute a workflow

To run a workflow:

- Visit the Workflows page (click the Workflow link near the top of SM home page).
- Select a workflow from the Workflow pane on the left.
- Review the parameters. The lower Preview pane should show an individual task's page in read-only mode, with the parameters loaded for the task you select in the Task pane on the right.
- Click the Edit Workflows button if you want to revise the workflow.
- Click the **Execute** button!

❖ Note: Checkboxes for Maximize CPUs: Extraction, Search, and de novo, should always be checked, unless you seek to force those tasks to use only a single CPU. While primarily intended to aid SRM development, single CPU capability might also be useful if one has a very large datafile or a very large sequence database and limited memory on the server.

Interactively place individual tasks in the SRM queue

SM uses the request queue concept to coordinate individual tasks even when not using a workflow. For example, on the Data Extractor page, the **Extract** button submits the extraction task to the Request Queue, as does the **Start Search** button on the MS/MS Search page.

Pages where interactive execution may be helpful – Autovalidation, Protein/Peptide Summary, and Archive Data – have a Queue request check box. If you mark the Queue request check box, the program puts the request into the queue, delivers an acknowledgment of queue submission back to the web browser, and saves all output files to the selected data directories on the SM server, which you can view later. If you do not mark the check box, the SM server bypasses the queue and delivers some form of results directly back to the web browser, while some files will also be written to the selected data directories on the SM server.

Use the Request Queue / Completion Log Viewers

The Request Queue Viewer shows a list of all tasks that are currently executing and those that are queued for execution. It lists the tasks in the order they were queued. Because some tasks depend upon earlier tasks, the tasks that are currently executing do not always appear at the top of the list.

The Completion Log Viewer shows a list of all tasks that have completed, with most recent shown at the top. This log includes all queued requests, whether you queued them individually or via a workflow.

14. Quick Reference

15. Protein/Peptide Summary

During the early days of working with SM, after processing a dataset, you may find yourself wondering which of the various reporting modes you should use in Protein/Peptide Summary. Table 1, below, is intended as a quick reference that outlines an intended example application for each mode.

Also, from the P/P Summary page, click the Load button and select the Defaults folder for a few saved parameter sets.

Table 1 Protein/Peptide Summary Modes

If you want to summarize results by:	And you want to organize results by:	Then use this mode:	Example application
PSMs	Row for each PSM	Peptide-Spectrum Match	List of all peptides identified in the data
Peptides	Rows: Distinct peptides	Peptide-Distinct	Collapse PSM list to best-scoring representative PSM per distinct peptide. Allows filtering options to view only peptides with variable modifications.
Proteins	Rows: Proteins Columns: samples	Protein-Protein Comparison	Compare protein across samples. Provides table of protein groups, subgroups. Collapse all PSMs that belong to the same protein into one row. Multiple options for handling peptides shared between protein subgroups. Combine quantitation to the protein level: median of all PSMs.
Variable modification sites (phospho, acetyl, ubiquityl)	Rows: variable modification sites clustered by protein Columns: samples	Protein- Var Mod Site Comparison	Compare variable modifications across samples. Provides table of phosphorylation sites. Collapse all PSMs that contain the same phosphorylation site into one row. Combine quantitation to the site level: median of all PSMs.
Peptides	Rows: peptides clustered by protein Columns: samples	Protein-Peptide Comparison	Compare distinct peptides across samples, case-insensitive (CI). Collapse all PSMs that contain same peptide into one row. Combine quantitation to the peptide level: median of all PSMs.
Proteogenomic sites (SAAVs, Spliceforms)	Rows: PG sites clustered by protein Columns: samples	Protein-Prot Genom site Comparison	Compare PG sites across samples. Collapse all PSM that contain the same SAAV or Spliceform into one row. Combine quantitation to the PG site-level: median of all PSMs.
Peptides	Rows: peptides clustered by protein	Protein Summary Details	View proteins, with supporting peptide details
Peptides	Rows: peptides clustered by protein Columns: samples	Protein-Peptide Distribution Columns	Method development (evaluation of 2D LC-MS/MS or other fractionation scheme)