Scaffold Quant

Version 5.0

User's Guide

Release Information

The following release information applies to this version of the Scaffold Quant User's Guide. This document is applicable for Scaffold Quant, Release 5.0 or greater, and is current until replaced.

Copyright

Limit of Liability

Proteome Software, Inc. has made its best effort in preparing this guide. Proteome Software, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this guide and specifically disclaims any implied warranties of merchantability or fitness for a particular purpose. Information in this document is subject to change without notice and does not represent a commitment on the part of Proteome Software, Inc. or any of its affiliates.The accuracy and completeness of the information contained herein and the opinions stated herein are not guaranteed or warranted to produce any particular results, and the advice and strategies contained herein may not be suitable for every user.

The software described herein is furnished under a license agreement or a non-disclosure agreement. The software may be copied or used only in accordance with the terms of the agreement. It is against the law to copy the software on any medium except as specifically allowed in the license or the non-disclosure agreement.

Trademarks

The name *Proteome Software*, the Proteome Software logo, *Scaffold, Scaffold Q+*, *Scaffold Q+S*, *Scaffold Quant, Scaffold perSPECtives, Scaffold DIA* and *Scaffold Elements* logos are trademarks or registered trademarks of Proteome Software, Inc. All other products and company names mentioned herein may be trademarks or registered trademarks of their respective owners.

Customer Support

Customer support is available to organizations that purchase *Scaffold, Scaffold Q+, Scaffold Q+S, Scaffold Quant,Scaffold PTM, Scaffold DIA or Scaffold Elements* and that have an annual support agreement. Contact Proteome Software at:

Proteome Software, Inc.
1340 SW Bertha Blvd
Suite 10
Portland, OR 97219
1-800-944-6027 (Toll Free)
1-928-244-6024 (Fax)
www.proteomesoftware.com

Table of Contents

Chapter 1
Getting Started with Scaffold Quant

System Requirements

For information about the system requirements for Scaffold Quant, see:

https://www.proteomesoftware.com/system-requirements

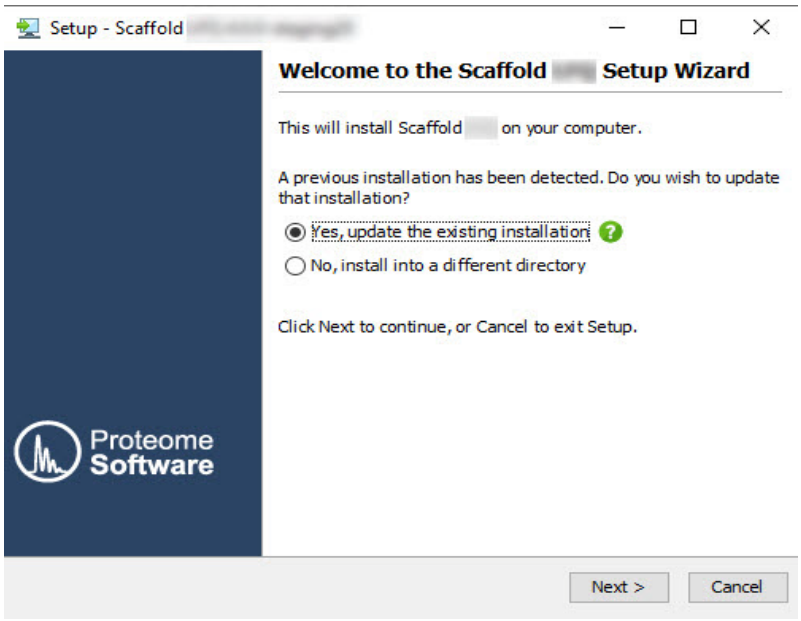
Installing Scaffold Quant

Scaffold Quant runs on Windows, MAC or Linux systems.Follow these instructions to install the application on your system:

Request an evaluation by filling in the form found at ***http://www.proteomesoftware.com/products/scaffold/evaluate/.*** You will receive download instructions and a license key to activate the software via email.

1. Download and launch the installation executable.
2. Carefully follow the instructions provided in the installation wizard, accepting the user agreement when prompted and moving through the screens by clicking Next.

Figure 1-1: Scaffold Quant installation Setup Wizard



3. The installer will then provide you an opportunity to allocate memory to Scaffold Quant. We recommend that you set the Maximum Memory to approximately 80% of the amount of physical RAM on your system. Click “Next”.

4. You may then select a Start Menu Folder for the application and choose whether or not to create shortcuts for all users of the system. The next screen allows you to set a file association between SFDB files and Scaffold Quant, and the following screen allows creation of desktop icons. Clicking “Next” begins the installation.

5. Finally, Scaffold Quant allows you to select the option to have the program open at the closing of the wizard. Click “Finish”..

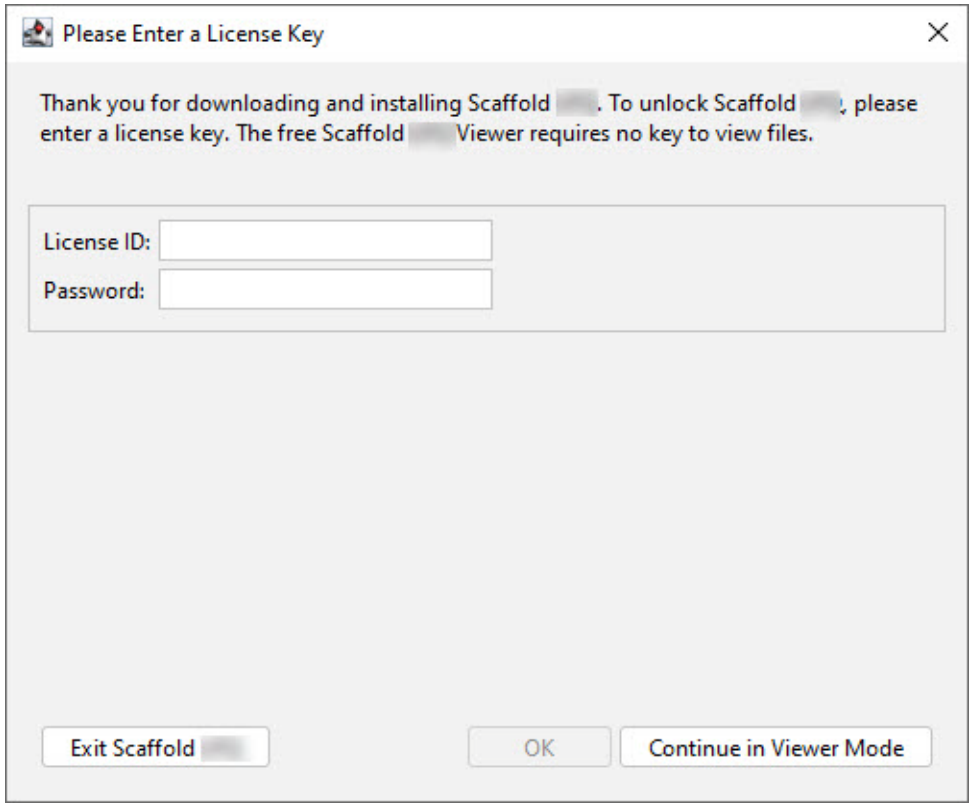
For better performance you should allocate more RAM to Scaffold Quant. The memory setting can be adjusted after installation by selecting the menu option ***Edit > Preferences - Memory tab****. You must close Scaffold Quant and restart the program in order for the new memory setting to take effect.*

After Scaffold Quant has been installed on a computer, a shortcut icon for the application is placed on the desktop. An option is also available from the Start menu. Double-clicking the desktop icon launches Scaffold Quant, as does, for Windows computers, selecting the option from the Start menu (**Start > All Programs > Scaffold Quant > Scaffold Quant**)

Licensing

The first time Scaffold Quant opens after installation, the Enter License Key dialog box opens.

Keys and passwords may be typed, pasted or dragged into the appropriate fields. Both items may be pasted or dragged together.



Two kinds of purchased keys are available to activate the software:

- **Standard (label-free Quant) Key** - this type of key allows the user to perform spectral counting, TIC and precursor intensity quantification.
- **Labeled Quant Key** - this premium level key allows the user to perform all of the standard quantification methods as well as to analyze multiplexed isobaric labeling experiments.

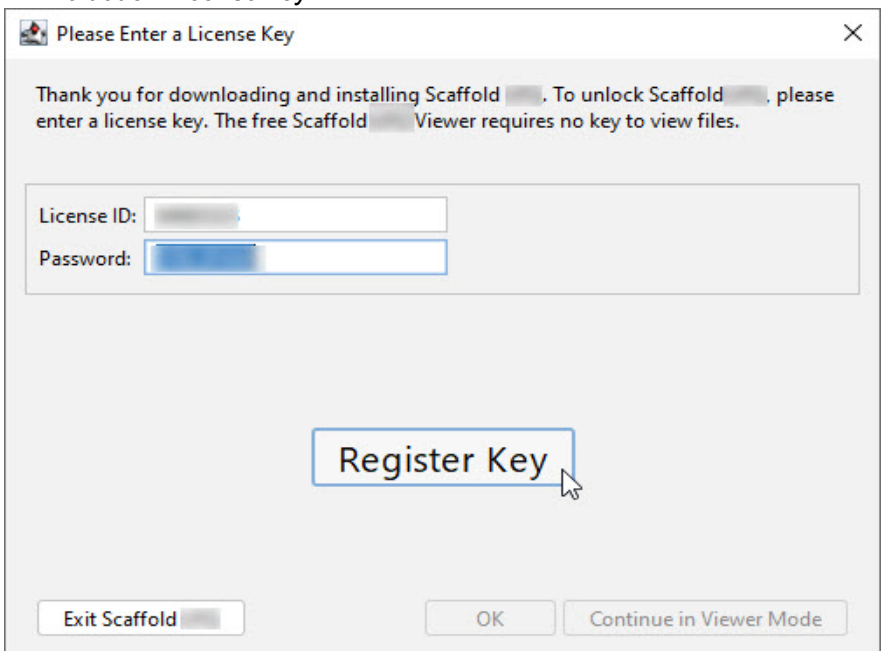
Purchased keys are also known as Time-based license keys, because while they allow full operation of the program indefinitely, the user is only allowed to upgrade the software while a valid software support contract is in effect. The time-based key tracks the term of the support contract and blocks upgrades when the contract expires. The time-based key may be reset by renewing the support contract.

Evaluation keys are also available to allow prospective users to evaluate the software for a limited time prior to purchase. Evaluation keys allow access to all functions available with the Labeled Quant key.

Evaluation key - An Evaluation key is valid for a limited period. A free evaluation key for Scaffold Quant

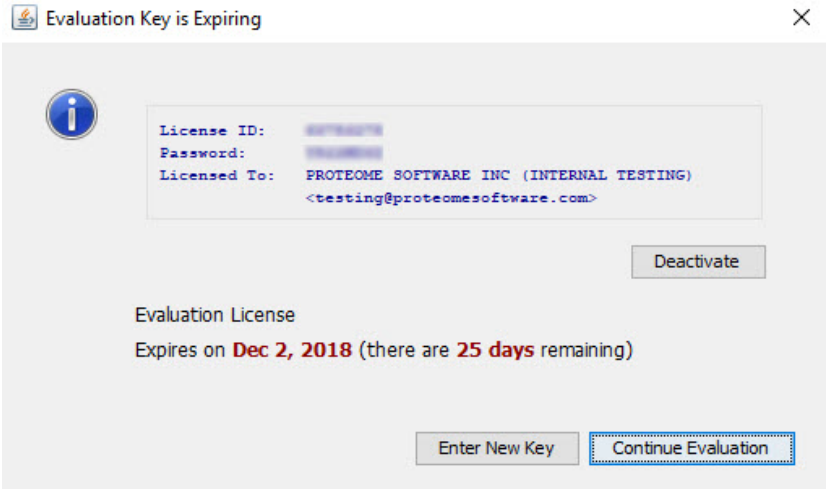
may be obtained through *www.proteomesoftware.com*. An evaluation key may be used on two computers. Once the key and password have been copied and pasted into the license key dialog box, a message will appear below it, displaying confirmation of the key registration. Pressing OK starts the application.

Figure 1-2: Evaluation License key



Every time Scaffold Quant is launched in evaluation mode, a message appears showing the remaining time available for evaluation and offering the option to enter a new key.

Figure 1-3: Message appearing when launching an evaluation copy of Scaffold Quant



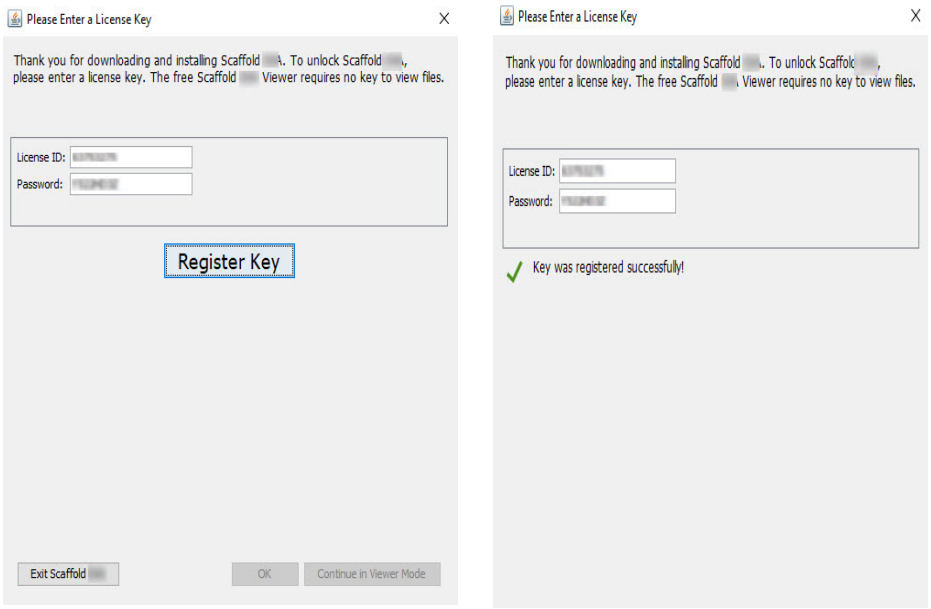
Time-Based License key—a Time-Based License key allows the user to access all features of the software permanently. It only allows upgrades within a certain time limit, however. The time tracks the length of the support contract. Once expired, Scaffold Quant will continue to work beyond the expiration date, but no

upgrades are allowed unless the support contract is renewed.

Contact *sales@proteomesoftware.com* to purchase the appropriate key.

A Time-Based License key is valid only for a single computer. If it is necessary to move the Scaffold Quant installation to a different computer, see for instructions to transfer the key at no charge.

Figure 1-4: Time-Based License key



When the Time-Based License key and password are entered, pressing **Register Key** verifies their validity and a message appears describing the status of the key.

Once the key is successfully registered, pressing OK closes the dialog box and a Scaffold Quant Welcome message opens.

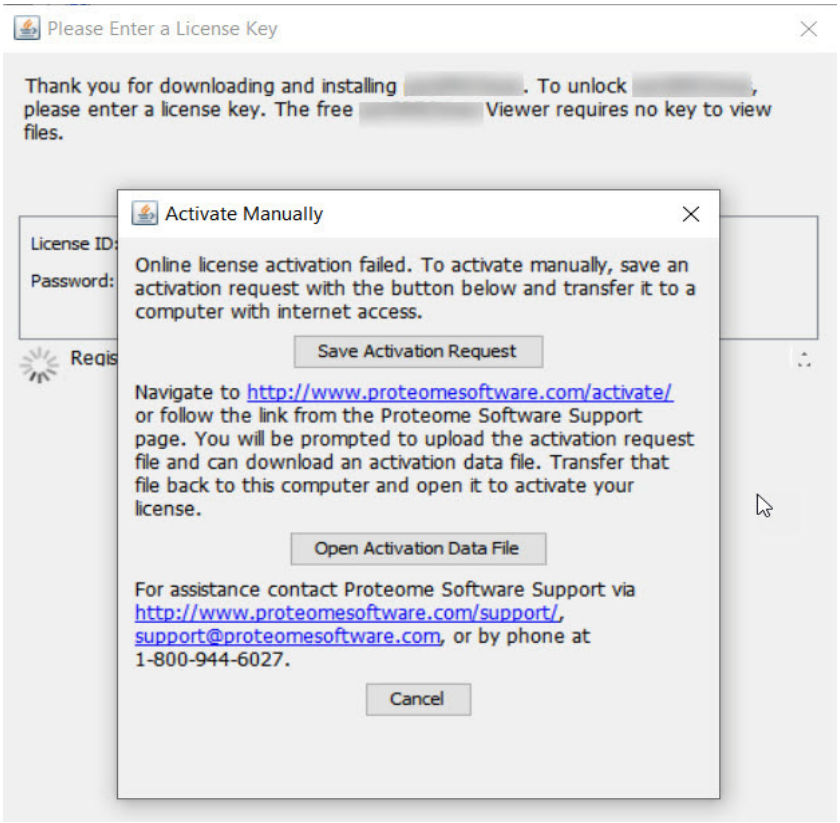
If the user is using an evaluation copy of Scaffold Quant, then an Evaluation message opens, indicating the number of days left in the evaluation period. The user must click OK to close this message and then the Scaffold Quant Welcome message opens.

From this window, the user may create a new experiment, open an existing experiment (*.SFDB file), or work with the demonstration data that is provided in the Scaffold Quant installation.

Registering a Time-Based License key with no INTERNET connection

When a Time-Based License key is entered and the Register Key button is pressed, but no INTERNET connection is available, a dialog appears, providing instructions for manual activation.

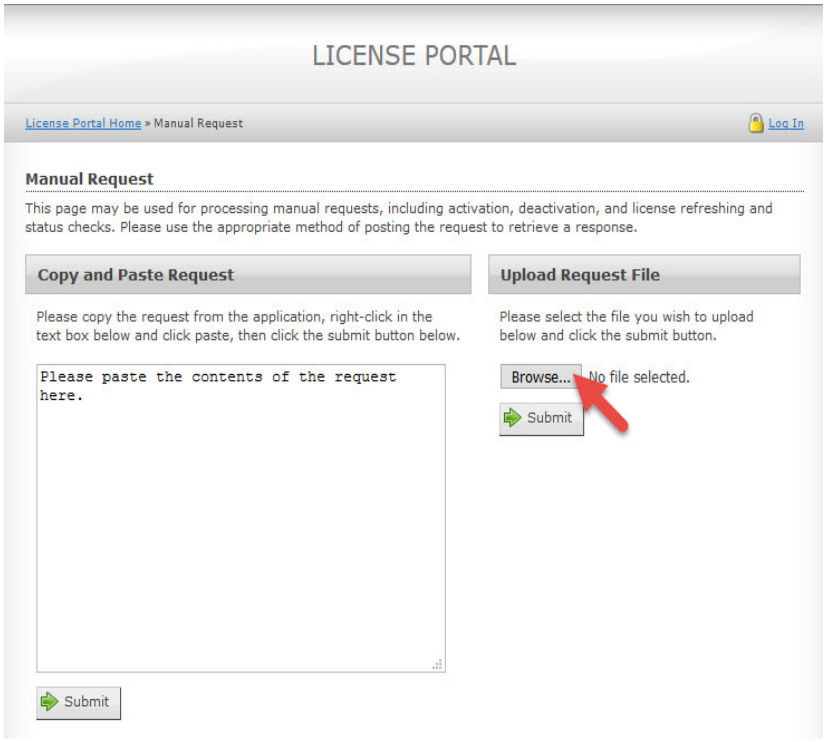
Figure 1-5: Manual or offline activation dialog



To activate Scaffold Quant without an internet connection:

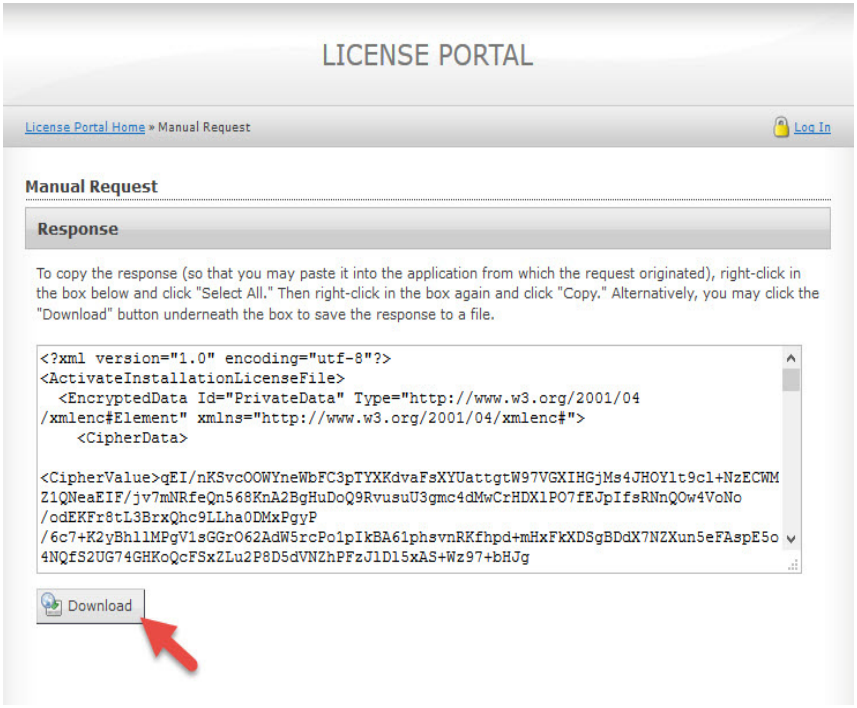
1. First, use the Save Activation Request button to create an activation request file.
2. Transfer this file to a computer with internet access (e.g. using a USB drive).
3. On the connected computer, navigate to *http://www.proteomesoftware.com/activate/*This link is also accessible from the Proteome Software Support page *(http://www.proteomesoftware.com/support/)* to make it easier to access from the internet-connected computer.
4. The License Portal will open. The Portal provides two different options for activating your software. Use the Browse button in the Upload Request File section on the right, and select the activation request file that was transferred from the offline computer(See Figure 1-6 below).

Figure 1-6: The Proteome Software License Portal



5. Click the Submit button just below the Browse button to upload the activation request file. The license portal will respond with a long text sequence(See Figure 1-7 below).

Figure 1-7: .License Portal Response to Activation Request

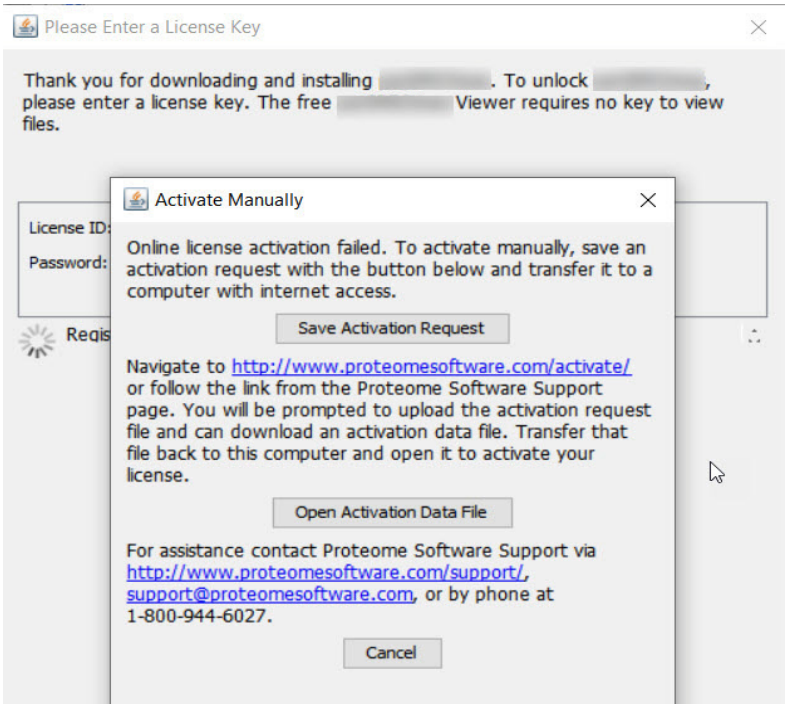


6. Click the Download button to save the response to a file named response.xml, which will be downloaded to the default download location.

7. Transfer the response.xml file to the computer on which Scaffold Quant has been installed.

8. Return to Scaffold Quant on the disconnected computer. Select Open Activation Data File.

Figure 1-8: Select Activation File returned by the License Portal



9. Browse to locate the response.xml file and click Open.

10. Scaffold Quant should report that the key was registered successfully. If not, please contact Proteome Software Support for assistance.

Time based license key renewal

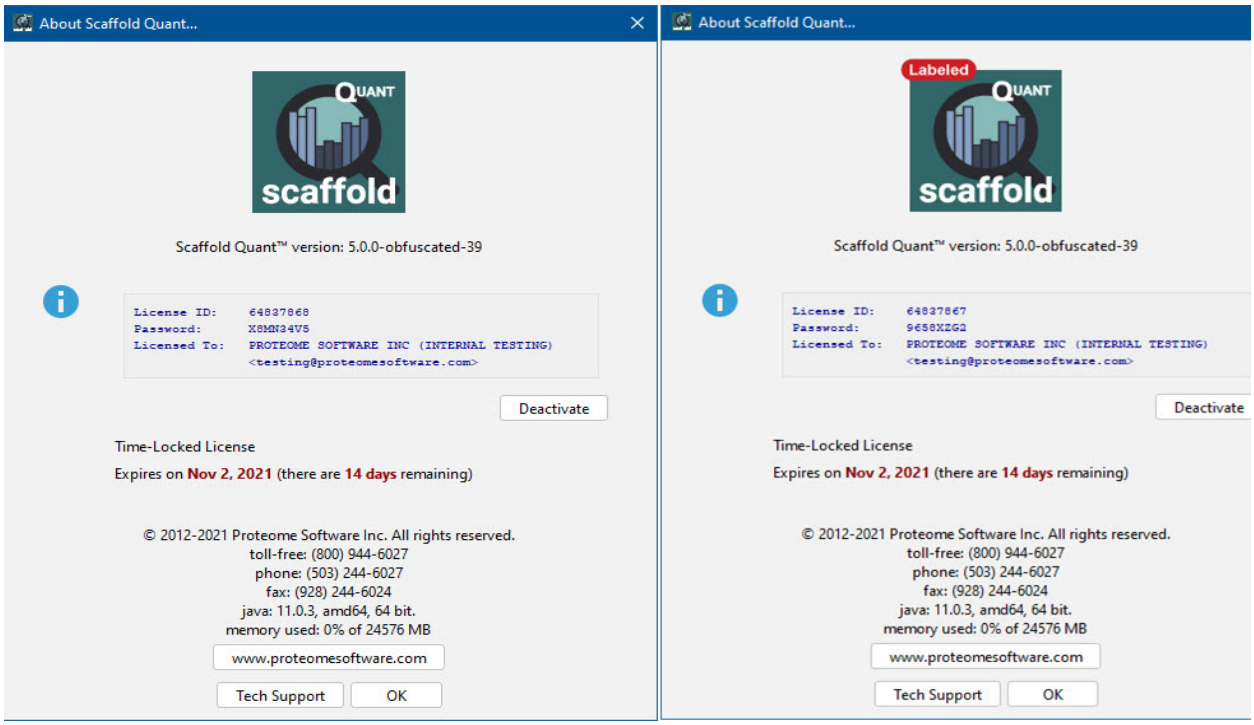
Time based license keys have time limits connected to the term of the user's support contract. When the support contract expires, Scaffold Quant continues to work but upgrades are not allowed until the contract is renewed. The status of the Scaffold Quant license key may be checked by selecting **Help > About Scaffold Quant** from the main menu.

If the contract has expired and the user wishes to upgrade Scaffold Quant, clicking the **Renew** button in the dialog opens the **Key reset Request** page on the Proteome Software website. The user should complete the request. A sales representative will promptly contact him/her providing further information.

Checking the status and type of License

Selecting **About Scaffold Quant** from the **Help** menu displays a screen that supplies information about the licensing status of the software. If the premium **Labeled Quant** license has been applied, the word "Labeled" will appear in the product logo. Similarly, the **Welcome** screen which appears when the program opens will display a logo which indicates the license type in effect.

Figure 1-9: About Scaffold Quant dialog with Standard or Labeled Quant license



Moving Scaffold Quant to a different computer

Each permanent Scaffold Quant key allows activation of the program on a single computer. If it becomes necessary to reinstall the program either on a different computer or on the same computer following an operating system upgrade or hardware replacement, the user may deactivate the key and then reactivate it on the new system. This may be done once per support contract period. If additional reinstallations are required within the same period, please contact Proteome Software Support.

To deactivate a key:

1. Be sure you have a record of your key and password. These were sent via email at the time of purchase, or may be copied from the Help>>About Scaffold Quant dialog.
2. Select Help>>Update License Key and click the Deactivate button.

To reinstall Scaffold Quant:

1. Download the program from the Proteome Software website to the new system and run the

installation program.

2. Paste in the key and password and register as described in Installing Scaffold Quant.

Scaffold Quant Viewer

A free Scaffold Quant Viewer may be downloaded from*www.proteomesoftware.com* The Viewer can open and display any *.SFDB file created by Scaffold Quant, and allows users to distribute Scaffold Quant results to colleagues, collaborators or reviewers.

The Viewer may be installed on any number of computers, and multiple instances of the Viewer may be run on a single computer simultaneously. It performs most of the functions of the full Scaffold Quant program, but it cannot load search results files and analyze data.

With the Viewer, the user (or his/her collaborators) can view the data, just as in Scaffold Quant, by samples, proteins, peptides or spectra. The user may apply thresholds, change summarization, adjust legend colors, move columns and hide rows. The Viewer user may also validate the peptide/spectrum matches.

Only a single fully-licensed instance of Scaffold Quant may be run on a computer at one time. Additional instances will function as Viewers.

Scaffold Quant highlights

Scaffold Quant is a software tool designed to help researchers organize, summarize, refine and visualize proteins across a large number of biological samples. This program supports analysis of complex experiments because it allows users to categorize samples based on various attributes to help extract biologically meaningful patterns in Mass Spectrometry results.

Organize

The power begins with the Organize View. Big data sets are often derived from complex experiments which have more categorical data types than Scaffold's basic category, biosample and MS/MS sample can accommodate. In Scaffold Quant, the user can create categorical variables or "Categories" and annotate samples using values or sample attributes associated with these categories. Attributes may be added through a graphical interface or loaded from a file. This provides the foundation for evaluation of a proteomics experiments from many viewpoints.

Summarize

Once attributes are applied, Scaffold Quant offers a great deal of flexibility to allow easy viewing of similarities and differences across sample groups. Flexible summarization allows the user to select Categories and use them to create a hierarchical categorization of the data. This makes it easy to:

- Compare protein and peptide similarities and differences among samples at any level of summarization.
- Compare the impact of tissue types, treatment types, demographic differences, measurement conditions and more.
- Properly account for technical and biological replicates.

Refine

Scaffold Quant imports data from any search engine or program (including Scaffold) that exports mzIdentML and it respects scoring from these sources. In Scaffold Quant, the user may:

- Filter by FDR, modification, peptide sequence and taxonomy.
- Validate spectra
- Perform label-free quantitation with spectrum counting or precursor intensity.
- Evaluate significance with a variety of statistical tests, such as t-test, ANOVA and Fisher Exact test as well as (non-parametric) Permutation testing.

Visualize

All proteins are easily visible, even those with exactly the same peptide evidence, providing greater transparency to view the samples' contents. Many specialized visualization tools are provided, along with the ability to:

- Cluster proteins with various degrees of similarity.
- Use customizable color to easily visualize counts, fold change, and quantitative differences between samples at various summarization levels.
- Annotate with Gene Ontology.

Scaffold Quant helps display patterns of expression across many samples with various attributes to provide new insight into an experiment.

Scaffold Suite of Products

Scaffold Quant is one of the Scaffold Suite of applications developed by Proteome Software, Inc. to facilitate the inspection and analysis of proteomics mass spectrometry data.

While Scaffold Q+S is an add-on to the core Scaffold product, Scaffold Quant is a standalone application and requires an independent license key provided by Proteome Software.

Application	Description
Scaffold	Visualize and validate MS/MS proteomics experiments.
Scaffold Q+S	Calculate and display relative protein expression levels in a sample determined by tandem mass spectrometry of stable isotopically-labeled (for example, SILAC) proteins.
Scaffold Quant	Catalog, summarize and analyze complex large-scale experiments. Compare protein and peptide similarities and differences at any summarization level. Easily reorganize samples to compare the impact of tissue types, treatment types, demographic differences, experiment conditions and more.

Referencing Scaffold Quant Results

Users are free to copy, modify, and distribute the following examples for citing Scaffold Quant in their publications and reports.

Scaffold Quant (Proteome Software, Portland, Oregon, USA) was used to validate and statistically compare protein identifications derived from MS/MS search results.

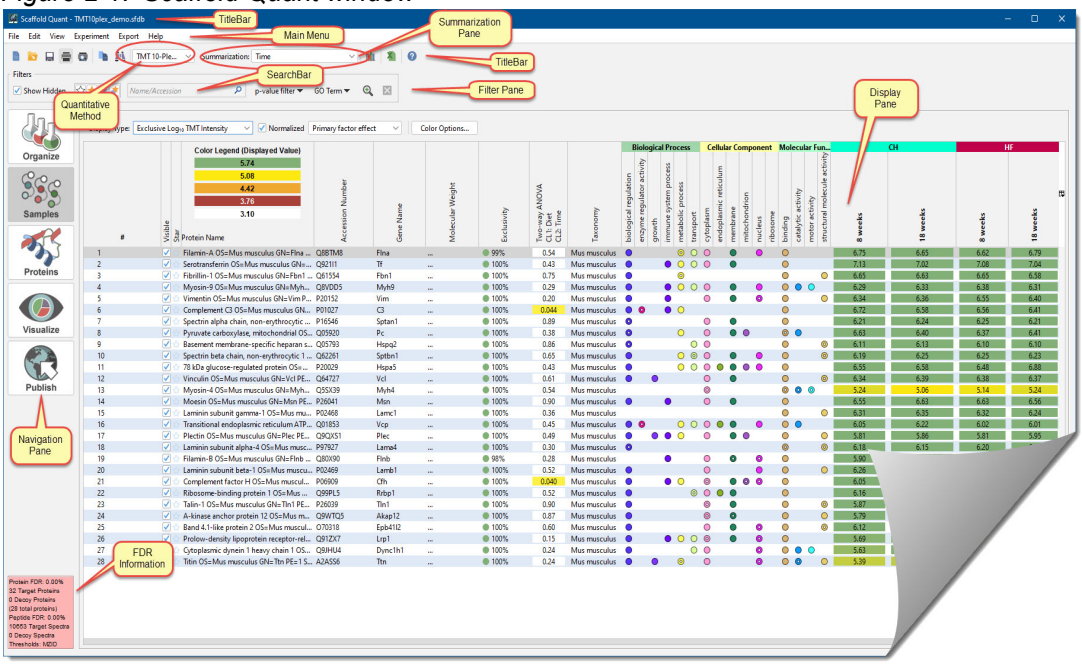
Chapter 2
Scaffold Quant Main Window

Like most applications in the Scaffold Suite, Scaffold Quant consists of a main window which provides access to a number of specific views. In each view, the data loaded into a Scaffold Quant experiment are organized so that a user can easily view the results from various points of view.

The Scaffold Quant Main Window provides quick access to all of the Scaffold Quant features and functions through the following features:

The Main Window

Figure 2-1: Scaffold Quant window



Title bar

Figure 2-2: Title bar

The title bar at the top of the Scaffold Quant window always displays “Scaffold Quant”. Additional text is displayed in the title bar depending on the actions that the user has performed. For example, when a new experiment is created, the default experiment name “-_Experiment” is appended. When a file is saved with a different name, the default name is replaced by the new file name. When an .SFDB file is opened, the title bar displays the file name.

---

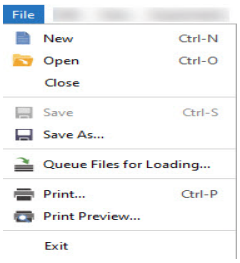
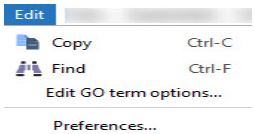
*The version of Scaffold Quant in use is not displayed in the Title bar. The version may be accessed through the **Help > About Scaffold Quant** option in the main menu. See “Main menu commands” below.*

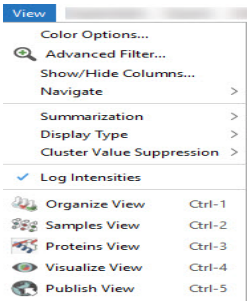
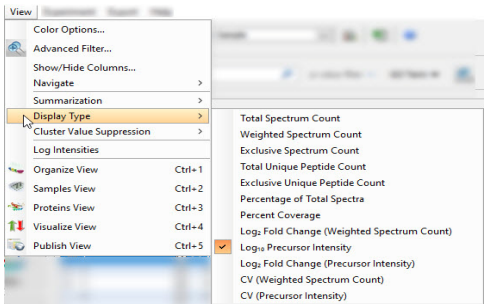
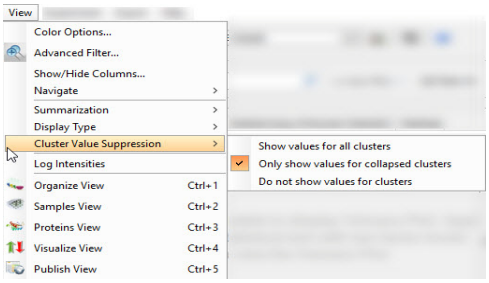
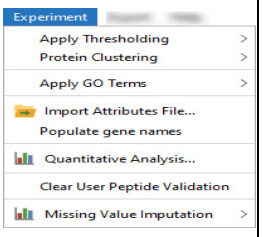
Main menu commands

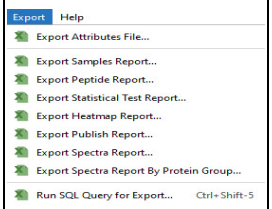
Figure 2-3: Main Menu

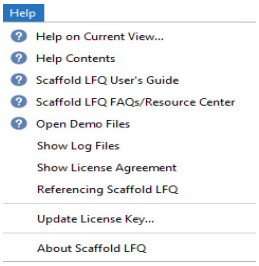
The Scaffold Quant main menu is organized in a standard Windows menu format with commands grouped into menus (File, Edit, View, Experiment, Export and Help) across the menu bar.

Some of these menu commands are also available in other areas of the application.

Menu	Menu Commands
File 	• **New**—Starts a new experiment and opens a file browser to allow the selection of files to load into Scaffold Quant. See Loading data in Scaffold Quant. • **Open**—Opens a saved Scaffold Quant experiment file, *.ftdb, through a file browser. • **Close**—Closes the current experiment, standard Windows behavior. • **Save**—Saves the current experiment, standard Windows behavior. • **Save As**—Saves the current experiment offering the option to use a different name, standard Windows behavior. • **Queue Files for Loading**—Opens a browser to select the files to be added to the current experiment. See Loading data in Scaffold Quant. • **Print**—Prints the current view. • **Print Preview**—Previews the current view with the option of printing the document. • **Exit**—Closes the Scaffold Quant window.
Edit 	• **Copy**—For each View, copies the first table appearing at the top of the View to the clipboard so it can be pasted into a third-party program such as Excel or Microsoft Word. • **Find**—Opens a find dialog box that searches the first table present in the Current View • **Edit Go Terms Options**—See Edit GO Term Options • **Preferences**—see Preferences

Menu	Menu Commands
View 	• **Color Options**— Opens the Adjust Display Options dialog, see Color Options button • **Advanced Filter**— Opens the Advanced filter dialog, see Advanced Filter • **Show/Hide Columns...**—Opens the Table Column Control menu • **Navigate** — Allows selection of tabs in a View. • **Summarization**—Equivalent to the Summarization Bar pull down menu • **Display Type**—Equivalent to the Display Type option in the Samples View  • **Cluster Value Suppression**—Helps the user select the clustering values that need to be visible, see Cluster Value Suppression 
Experiment 	• **Apply Thresholding**—See Applying Confidence Thresholds to the Protein List • **Protein Clustering**—See Grouping and clustering options. • **Apply Go Terms**—Applies imported Gene Ontology Annotations to the Samples Table.To import GO databases see Edit GO Term Options. • **Import Attributes File**— See Import Attributes File... • **Populate Gene Names**— See Gene Names. • **Quantitative Analysis**— See Configure Sample Organization and Statistical Analysis dialog • **Configure Labeled Quantification...**— See Configure Labeled Quantification • **Clear User Peptide Validation...**—This option reestablishes the initial selection of peptides at load. For more info see The Validation Pane. • **Missing Value Imputation** — Allows the user to direct how missing values should be treated. There are two options: QRILC - replace missing values with values imputed by QRILC or None - do not impute values at all.

Menu	Menu Commands
Export 	• **Export** :: • **Attributes file...**—Generates a comma-delimited-delimited text file of the meta-data attributes assigned to each ms sample in the current experiment, see Sample Organization tree table. • **Samples Report to Excel...**—Generates a comma-delimited text file of the Samples table appearing in the Samples View, that can be opened and viewed in Excel. • **Peptide Report to Excel...**—Generates a comma-delimited text file of the Peptide table for all proteins appearing in the Samples View, that can be opened and viewed in Excel. • **Statistical Test Report to Excel...**--Generates a comma-delimited text file providing the details of the calculation of any applied statistical test. • **Heatmap Report to Excel...**—Generates a comma-delimited text file of the information depicted in the Heatmap appearing in the Visualize View, that can be opened and viewed in Excel. • **Publish Report to Excel...**—Generates a comma-delimited text file of the information about the current experiment provided in the Publish View, that can be opened and viewed in Excel. • **Spectra Report...** —Generates a comma-delimited text file displaying the characteristics of the peptide-spectrum matches that can be opened and viewed in Excel. • **Spectra Report by Protein Group...** —Generates a comma-delimited text file showing the peptide-spectrum matches associated with each protein in the experiment. It can be opened and viewed in Excel. • **Run SQL query for Export...**—Opens the SQL dialog box see SQL Export tab

Menu	Menu Commands
Help 	• **Help on Current View**—Opens the Online Help that is specific for the currently displayed topic. • **Help Contents**—Opens the Contents page for the Online Help. • **Scaffold Quant User's Guide**—Opens the current Scaffold Quant User's Guide. • **Scaffold Quant FAQs/Resource Center**—Opens the user's default web browser to the Home page of the Proteome Software's resource center. • **Open Demo Files**—Opens the folder where Scaffold Quant demo files are stored. The user can choose any of the pre-loaded files to test Scaffold Quant capabilities. • **Show Log Files**—Opens the folder containing Scaffold Quant error log and output_log files • **Referencing Scaffold Quant**—Opens the Online Help that contains a sample of how to reference Scaffold Quant when publishing data analyzed with this application. • **How to Purchase** — opens the Purchase Information page of the Proteome Software website in the default web browser. • **Update License Key...**—Opens a dialog where the user can update the license key to activate Scaffold Quant, see Licensing. • **About Scaffold Quant**—Provides the release information for the current version of Scaffold Quant, license information, contact information for Proteome Software, Inc. It also reports information about the system where Scaffold Quant is installed, the type of license installed, the amount of memory available to the software and the percentage of memory used by the application.

Edit

Some options appearing under the Menu Edit:

Edit GO Term Options

Preferences

Edit GO Term Options

Selecting Edit>Edit GO term options opens the GO Term Configuration Dialog

See Edit GO Term Options

Preferences

In the main menu, selecting **Edit > Preferences** opens the **Preferences** dialog. It contains the following tabs:

- General
- Analysis

- System
- User Interface
- Colors
- Heatmap

General

This tab contains general options:

- **Preferred Protein identifier pull down list** -- Here the user can select the type of identifier to appear in the Protein menu. Two choices are available: protein name or protein accession number.

Analysis

This tab allows the user to define default analysis settings each time a new experiment is loaded:

- **Protein Grouping options** - Choices available are:
  - Externally specified grouping - as specified in the originally loaded files (.e.g MZID or Scaffold-generated SFDB)
  - Perfect Shared evidence - as defined in .
- **Thresholding -** Options are provided for setting FDR Thresholds for the experiment.
  - FDR Thresholding - which allows the user to set target Protein FDR and Peptide FDR percentages and to specify a minimum required number of peptides. FDR is calculated as described in the Appendix "Computation of protein and peptide FDR in Scaffold Quant."
  - Externally-specified Thresholding - which respects the "passThreshold" designations in the original (e.g. mzIdentML) files that are loaded. In SFDB files exported directly from Scaffold, this type of thresholding respects the thresholding performed in Scaffold.

Note: In externally-specified thresholding, proteins with zero total spectrum counts in every MS sample sometimes appear, especially when MZID files exported directly from search engines are loaded.

System

This tab provides information to Scaffold Quant about interactions with the computer's Operating System.

- **Memory Usage** - Allows the user to set the maximum amount of memory Scaffold Quant can use

Note: On a Windows machine, to establish the maximum amount of memory that can be assigned to Scaffold Quant, the user can open the Task Manager, go to the Performance tab and check for the total amount of memory in use. Then assign at max a little bit less than the available memory for use to Scaffold Quant.

Note: MAC OS, once Scaffold Quant is installed, does not allow memory resetting unless the User is an administrator. A non-administrator user can update the memory only by reinstalling the software.

- *The new memory setting will take effect only after the application has been closed and restarted.*

- **Number of Processors** - the maximum number of processors available for threading computations. The default value is the maximum number of processors available on the computer where the application is installed.
- **Internet Settings** - determine how Scaffold Quant interacts with the internet. A user may:
  - **Allow Scaffold Quant to connect to the Internet** If this box is unchecked, then Scaffold Quant cannot access the Internet. Users may wish to uncheck this box if their organizations prohibit connection to the Internet.

    Use an HTTP Proxy Server - proxy servers may be used by an organization's IT department to filter communications to and from the Internet. In this case, the user must set the Proxy Server Name and Port Number. To check if this is necessary, a user may look at how the his/her web browser is connected to the web.
    - **Proxy Server name (or IP address)**
    - **Proxy port number**

User Interface

- **Search Fields - Regular expression check box** -- Once selected this option allows the use of regular expressions in any search box present in the program.
- **Prompts**- **Always run as viewer check box** - turns this option on or off.

  Reset "don't show" messages button -- Messages or warnings sometimes appear in which include the option not to show the warning anymore. This button reactivates the warnings to the original settings.
- **Views** -- Select which View should be opened by default when data has been loaded into Scaffold Quant or when opening an existing file.

Colors

This tab allows the user to change the colors assigned to Post Translational Modifications. Double clicking on the colored square assigned to a particular modification opens a dialog which includes standard color picking options.

Heatmap

This tab contains a checkbox which allows the user to select whether the Heatmap will cluster the columns according to their similarity or will display the columns (samples or categories) in the order in which they appear in the Samples View. In either case, rows (proteins) will be clustered. The selection remains until it is changed, even if the program is closed.

View

Some items appearing in the View menu:

Color Options - see Color Options button

Advanced Filter - see Advanced Filter

Show/Hide Columns - see Table Column Control

Navigate - selects the tab to be displayed. The options are to Select previous tab or Select next tab. These options are disabled when the currently displayed View contains only a single tab.

Summarization - see Summarization Bar

Display Type - see Display Type

Cluster Value Suppression - see Cluster Value Suppression

Cluster Value Suppression

This option allows the user to show or hide the rolled up values at the cluster level. The option is meant to help the user navigate through the samples list when clusters are applied.

- **Show Values for all clusters** - The rolled up values are shown in any of the rows that represent the top level of a cluster
- **Only Show values for collapsed clusters** - The rolled up values are shown only for the rows that represent the top level of a collapsed cluster
- **Do not show values for clusters** - The rolled up values are not shown for any of the rows that represent the top level of a cluster

Gene Names

In the Uniprot Protein Fasta format, the protein description often contains the gene name that has been associated with the protein. Selecting the **Populate gene names** option in the **Experiment Menu** parses the gene name from the Protein Name field. This only succeeds if the Protein Name is in Uniprot format and contains the string “GN=...”. If the Protein Name does not contain the gene name in this format, no gene names will be extracted and the column will remain empty.

Gene Ontology Annotations

Many genes and proteins are annotated in public data repository by Gene Ontology or GO terms or annotations. Scaffold Quant can annotate protein using the GO terms found in the GOA database downloaded from the EMBL-EBI websiteor in any custom GOA database imported in the program. The GO system annotates proteins with a hierarchy of terms. For example, one biological process of the protein albumin may be described at a high level as a "physiological process". At a more detailed level it can be described as "regulation of body fluids". At an even more detailed level the description says that albumin is involved with "water homeostasis".

Proteins are described by GO terms in three different categories:

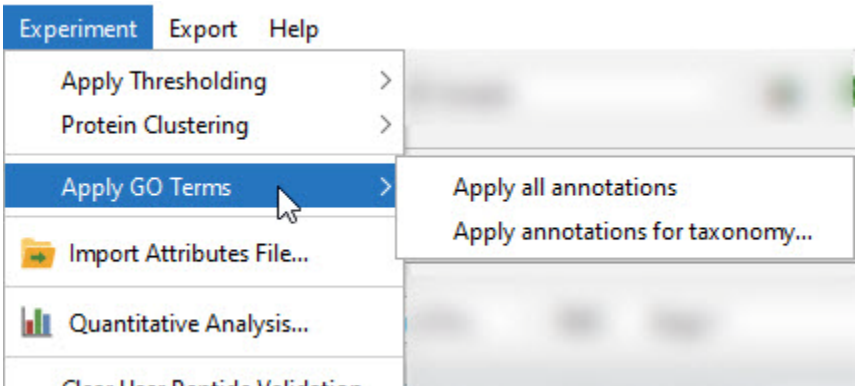
- Biological process
- Cellular component
- Molecular function.

Scaffold Quant shows the high-level GO annotations as colored dots in the appropriate column added to the Samples table. Depending on the type of evidence used to define a GO term the colored dots will appear either as open or closed circles, see GO Terms - Open vs Closed Circles.

Apply GO terms

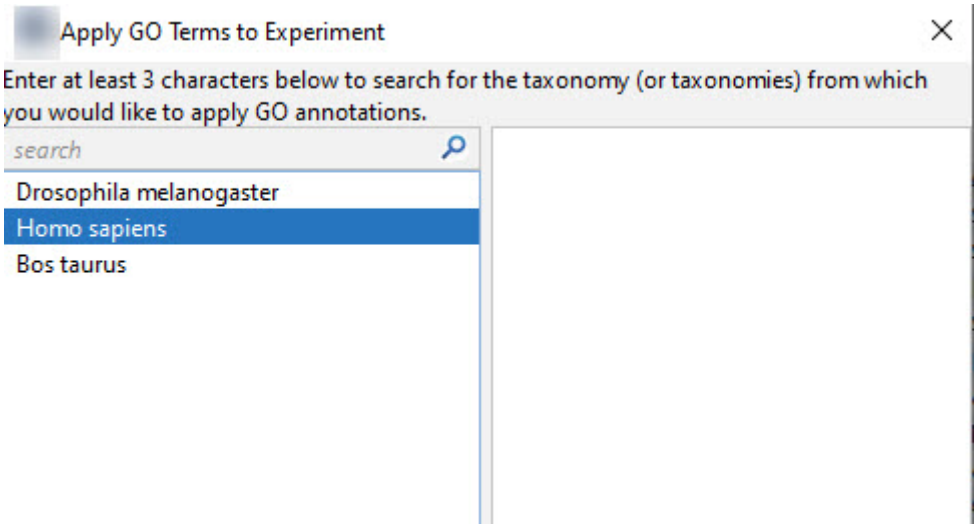
Scaffold Quant adds GO terms to the Samples Table when the command **Experiment > Apply GO terms** is chosen. The user may also select to filter the list of GO terms which will be displayed by taxonomy.

Figure 2-4: Applying GO terms



Clicking on Apply annotations for taxonomy... brings up a dialog from which the user may select the taxonomies for which GO terms should be accepted. All taxonomies currently in the GOA database are displayed. Select one or more taxonomies and click Add.When all desired taxonomies have been selected, click Apply.

Figure 2-5: .Selecting a taxonomy for GO terms



The terms are searched against the set of GOA databases appearing in the list of imported databases shown in the GO Annotations Tab. Once the search is finished, Scaffold Quant displays the GO terms found in separate columns added to the Samples Table. The number of added columns depends on the number of terms added to the GO terms Display List.

The GO Annotations columns may be hidden by going the Table Column Control and unselecting the **GO Terms** entry.

GO Terms - Open vs Closed Circles

Scaffold Quant utilizes the Gene Ontology (GO) “evidence code” to determine how to display GO phenotype/function data. It displays a closed circle if the annotation was derived by a human experimenter/curator; an open circle if the annotation was assigned by a computer algorithm. If any of the GO associations for a given sequence have a non computational evidence code, then a closed circle is displayed. Table 2-1 indicates which evidence codes are human curated (closed circle), or computationally annotated (open circle). For a detailed explanation of each of the evidence codes, see http://www.geneontology.org/GO.evidence.shtml .

Table 2-1: GO Evidence Code

Type of evidence	Code	Type of Circle
Unrecognized Evidence Code	UNKNOWN	closed circle
Inferred from Mutant Phenotype	IMP	closed circle
Inferred by Curator	IC	closed circle

Table 2-1: GO Evidence Code

Inferred from Genetic Interaction	IGI	closed circle
Inferred from Physical Interaction	IPI	closed circle
Inferred from Sequence or Structural Similarity	ISS	closed circle
Inferred from Direct Assay	IDA	closed circle
Inferred from Expression Pattern	IEP	closed circle
Inferred from Electronic Annotation	IEA	open circle
Traceable Author Statement	TAS	closed circle
Non-traceable Author Statement	NAS	closed circle
Not Recorded	NR	closed circle
No biological Data available	ND	closed circle
Inferred from Reviewed Computational Analysis	RCA	open circle
Inferred from Sequence Orthology	ISO	open circle
Inferred from Sequence Alignment	ISA	open circle
Inferred from Sequence Model	ISM	open circle
Inferred from Genomic Context	IGC	open circle
Inferred from Experiment	EXP	closed circle

Note that GO terms downloaded from NCBI will always have closed circles since the information provided does not allow establishing the difference between experimentally verified and computationally derived GO terms.

Edit GO Term Options

Selecting **Edit > Edit GO Term Options** from the main menu**,** opens the **GO Term Configuration** dialog. It contains the following tabs:

- The Displayed GO Terms Tab
- GO Annotations Tab

The Displayed GO Terms Tab

This tab contains tools that allow the user to create and modify a custom list of GO terms. The list is then displayed by Scaffold Quant as extra columns in the Samples Table.

The Tab window is divided into the following sections:

- **Search Field** - Searches terms available in the GO terms database loaded in Scaffold Quant.

- **GO Tree list** - Hierarchical list of all the terms present in the loaded GO database
- **Add and Remove GO terms** - Provides tools for creating the custom Display list
- **Display List** - List of GO terms, selected by the User, that will be visible in Samples Table.
- **Save and Apply**- Allows the User to save the current Display List if changed

To create a new custom GO terms Display List the User needs to follow these instructions:

1. If the **Display List** is not empty select all the rows and press delete.

2. Search and select any GO term of interest present in the loaded GO database either by typing a name in the **Search Field** or by selecting a row in the **GO Tree List**.

3. Click **Add**; the selected term or group of terms is added to the **Display List**. Terms may be selected individually or by domain or group. If a group or domain is selected, all terms in that group will be added to the **Display List**.

4. To remove terms from the **Display List**, select a term or group of terms to be discarded then click **Remove**.

5. To save the current selections as User Defaults check the box **Save displayed GO terms as user default**.

When a Scaffold Quant experiment is saved, the displayed GO terms are saved within the *.SFDB file.

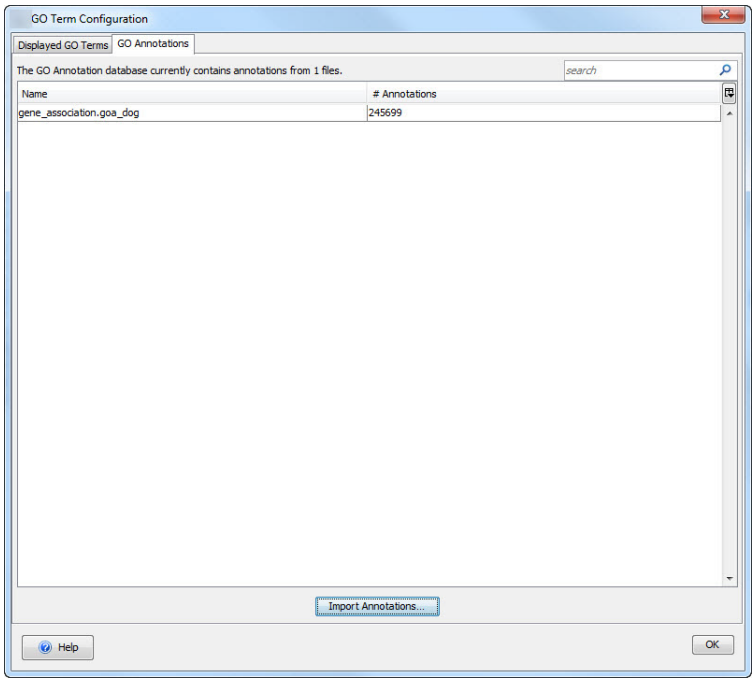
When a new file is created, or when Scaffold Quant is closed, the list of displayed GO terms is unchanged. To reset the list to the defaults, the user may click the **Reset to User Default** or the **Reset to Scaffold Quant Default** button.

GO Annotations Tab

Scaffold Quant includes a GO terms table that the User can populate with information from existing GO terms databases through the Import Annotation function.

The GO Annotations Tab shows the list of already imported GO databases, a search box and the Import annotations button to import a GO database.

Figure 2-6: Go annotations tab

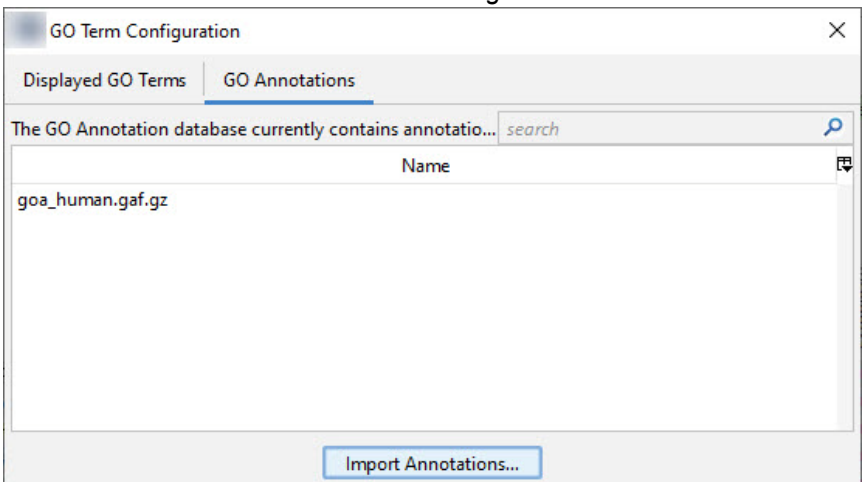


Import annotations

This button opens a dialog from where the user can import GO databases in Scaffold Quant.

Through a pull-down menu it is possible to direct Scaffold Quant to a location where the GO database can be downloaded .

Figure 2-7: Add GO Annotations Database dialog



- **Human Only** - provides a download of the human subset. It takes about 10 minutes to download.

- **Other Website** - the user may type or paste in a website address from which a GO Database can be downloaded.
- **Other File** - the User can direct Scaffold Quant to a location in his/her computer where the GO database is stored.
- **Specific GO databases for various taxonomies** from ftp://ftp.geneontology.org/pub/go/gene-associations/.

Note: **Other species-specific GO databases** are available from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/. To access these, locate the correct taxonomy beneath this site in a web browser, hover over the *.gaf.gz file, right-click and choose “Copy Link Location”. Select “Other Website” in this dialog and paste in the copied link.

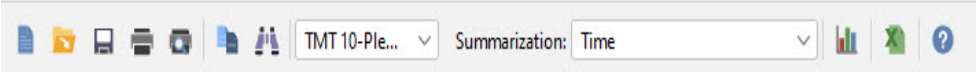
After one of the options has been selected, clicking **Add** starts the operation of importing the GO annotation database into Scaffold Quant. A new row appears in the list of already loaded databases providing the name of the newly added database and the number of annotations imported through it.

Clicking **OK** closes the dialog and now Scaffold Quant is ready to annotate the proteins in the proteins list with GO terms by selecting, the now available option, **Experiment > Apply GO Terms.**

The command Experiment > Apply GO Terms is available for use only when one or more GO Annotations databases are loaded into Scaffold Quant.

Tool-bar

Figure 2-8: Scaffold Quant tool ba



The Scaffold Quant tool-bar contains icons that represent equivalent commands for frequently used main menu options.

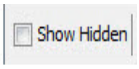
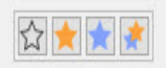
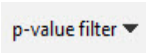
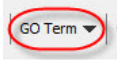
Icon	Function
	New—Starts a new experiment and opens a file browser to allow the selection of files to load in Scaffold Quant. See Loading data in Scaffold Quant.
	Open—Opens a saved Scaffold Quant experiment file, *.SFDB, through a file browser.
	Save—Standard Windows behavior.
	Print—Prints the current view.
	Print Preview—Previews current view with the option to print the document.
	Copy—For each view copies to the clipboard the first table appearing at the top of the view. From there, the user can paste it into a third-party program such as Excel or Microsoft Word.
	Find—Opens a find dialog box that searches the first table present in the current view
	Quantitative Analysis—See Configure Sample Organization and Statistical Analysis dialog.
	Excel—Exports the information that is contained in the current view to a comma-delimited text file that can be opened and viewed in Excel.
	Help—Opens the Scaffold Quant Online Help.

Filter control bar

The Scaffold Quant Filter control bar, located under the Tool-bar, contains an ample selection of filtering options. When the GO terms are applied, a GO term filter appears in the control bar. When a Quantitative test is applied a P-value filter becomes available.

Figure 2-9: .I

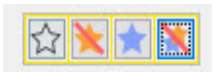
The Filter control bar includes the following functions:

Icon	Function
	**Show Hidden**—Toggles the view of hidden proteins in the Samples View's Samples table. see Hidden Proteins.
	**Star Filter**—Filters the proteins that were tagged with a specific star in the Samples View Proteins Table, see Star Filter.
	**P-value Filter**— Filters the proteins appearing in the Samples View Proteins table according to whether or not their p-values meet the significance criteria set in the Signifcance Level tab. If Family-wise error correction has not been applied, the "Significant after FWER correction option is grayed out. This filter appears in the Filter control bar only after a Quantitative test has been applied, see Quantitative Tests.
	**Advanced Filter** icon— Clicking this icon opens a dialog box where the advanced searches can be set up, see Advanced Filter.
	**GO Term filter**—Filters the proteins appearing in the Samples View Proteins table for a specific group of Gene Ontology Annotations. This filter appears in the Filter control bar only after GO terms have been searched, see GO Annotations Tab.

Star Filter

The Star Filter pane contains an icon for each type of star. Clicking on a star icon filters out

proteins tagged with that type of star. When a star filter is applied, the icon displays a bar across the star.



Selecting one or more of the stars will filter all proteins that are tagged with that particular combination of stars out of the Protein List in the Samples Table. See how to assign stars to proteins in the Samples View:

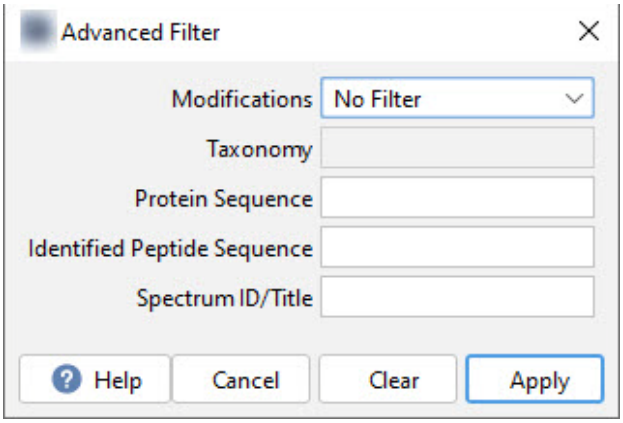
P-Value Filter

Selecting one of the filter options will filter the Protein List in the Samples Table to show only proteins for which the p-value calculated by the currently applied statistical test meets the filter criterion. The statistical test is selected in the Statistical Analysis Tab of the Configure Sample Organization and Statistical Analysis dialog. In that same tab, the user also specifies whether a Multiple Test Corrections should be applied to the results, and the acceptance level for the resulting statistic. The p-value filter allows the user to display only those proteins with statistically significant test results after applying the multiple test correction, or those for which the test result would have been significant had the multiple test correction not been applied. If no multiple test correction has been applied, this option is grayed out.

Advanced Filter

Selecting the **View > Advanced Filter...** option from the main menu, opens a dialog containing a list of the advanced filters included in Scaffold Quant. All filters except the Modification filter, include a dedicated text box search field where the user can type in terms that define how the list of protein is filtered. By selecting the menu option **Edit > Preferences** and selecting the **General** tab, it is possible to allow the use of regular expressions in any of the text search fields present in this dialog.

Figure 2-10:



The modification filter is a pull down list which includes all of the variable modifications identified in the experiment.

If more than one filter is selected at the same time, Scaffold Quant "AND"s the chosen filters. In other words, a peptide or protein must satisfy ALL selected filters in order to be displayed. The APPLY button starts the application of the filters and a Wait dialog box appears to monitor its progress.

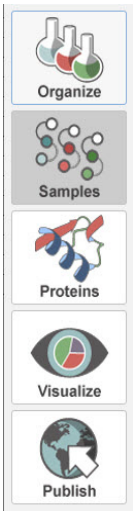
The application of the Advanced Filter affects the Protein list in both tabs of the Samples view tabs and the list of spectra and peptides appearing in the Proteins view.

While the advanced Filter is applied, the imposed filtering conditions are summarized in the Filter control bar. A red cross button also appears, clicking the button cancels the applied filter resetting the protein list to the original one.

Navigation pane

The Scaffold Quant Navigation pane is a vertical bar displayed on the left side of the Scaffold Quant window.

Figure 2-11: Scaffold Quant Navigation pane View selection



The bar contains large buttons that toggle the five different views available in the Scaffold Quant main window:

- The Organize View, see “The Organize View” on page 55
- The Samples View, see “The Samples View” on page 79
- The Proteins View, see “The Proteins View” on page 101
- The Visualize View, see “The Visualize View” on page 117
- The Publish View, see “The Publish View” on page 131

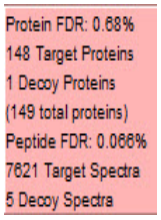
FDR Information Box

The **FDR Info** box (or dashboard) is located under the navigation pane on the left lower corner of the Scaffold Quant main window. The box contains information about the peptide and protein FDR at the current FDR Thresholds together with the number of target and decoy spectra and proteins. It also lists the thresholds currently applied.

When loaded data was searched without decoys the background of the info box is blue.

Note: The information reported in the Info Box when highlighted can be copied by simply using the standard CTRL C key strokes.

Figure 2-12: FDR Info Box



Quantitative Method

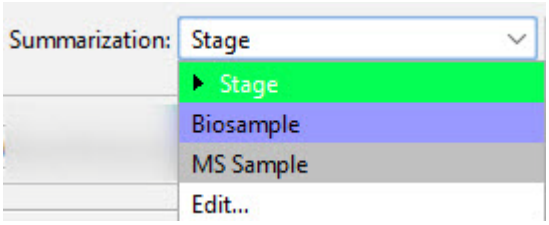
This drop-down allows the user to select the quantitative method that will be used in analyzing the experiment. The methods included in the dropdown depend on the types of quantitative data potentially available in the experiment. For instance, in an unlabeled experiment, the only option is “Label-free.”

If isobaric labeling has been used, there may be several potential options. Choose the correct quantitative type for the experiment and click OK. For information about loading labeled experiments and populating the SFDB file with quantitative values, see Loading Isobaric Labeling Experiments.

Summarization Bar

The Summarization bar allows the user to view the data collapsed or expanded at different hierarchical levels. The Summarization bar operates through a drop down menu containing a list of Categories hierarchically ordered.

Figure 2-13: Scaffold Quant Summarization bar



The last item in the list is the command **Edit...**, which, when selected, opens the Experimental Design dialog which allows addition of Categories to the drop down list and definition of a different hierarchical order.

Display pane

The information included in the different views appears in the Scaffold Quant Display pane. Depending on the view, the type of information reported might appear framed in one or more tables or graphs included in one or more sub-panes. All panes and tables in Scaffold Quant share the following characteristics:

- Tool-tips
- Resizing of columns and panes
- Table Column Control
- Moving columns
- Column sorting feature
- Multi selection of rows

Tool-tips

The user can view information about fields or columns in a View by just hovering the mouse pointer over the location of interest.This operation opens a collapsed tool-tip. Pressing F2 opens an expanded tool-tip. Pressing the Escape (ESC) key on the keyboard closes the expanded tool-tip.

Figure 2-14: Viewing information in a collapsed tool-tip

Figure 2-15: Viewing information in an expanded tool-tip

Resizing of columns and panes

The user can resize columns and different panes in each of the views to better suit his/her working needs. For example, in the Samples Table, the user can change the width of a column by resting the mouse pointer on the right side of a column heading until the pointer changes to a double-headed arrow, and then dragging the boundary until the column is the width that he or she wants.

Figure 2-16:

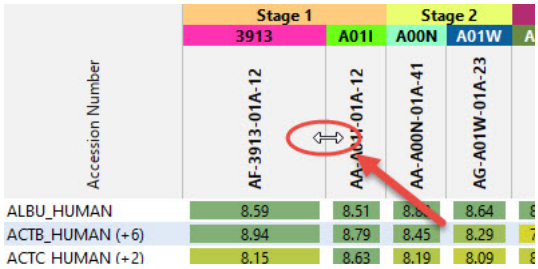
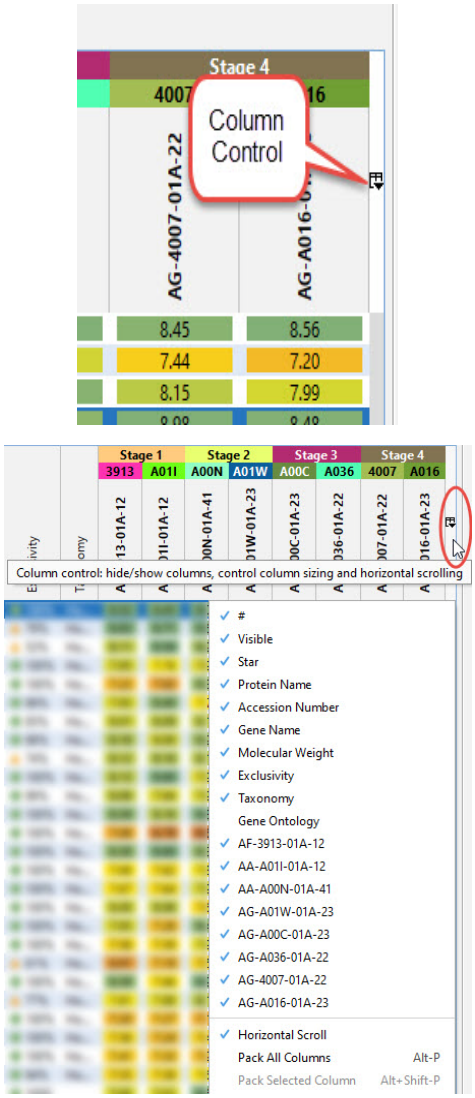


Table Column Control

All tables throughout Scaffold Quant have a feature called Column Control. It is a vertical

button placed on the right side of every table lined up with the column headers. When the user clicks the button or selects the option **View > Show/Hide Columns** from the main menu, a drop down list of all the columns opens. Each column is associated with a check box and at the bottom of the list there are included three group commands.

Figure 2-17: Column Control pull down list

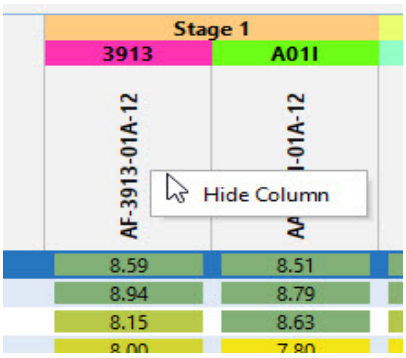


-
- Un-checking columns from the list will hide them from view in the Samples table.
- The Horizontal Scroll command if checked will add a scroll bar at the bottom of the Samples table.
- Pack all columns when selected resizes all samples columns to a common width.

- Pack selected column resizes the column that contains the current selected cell. If no cell has been selected the command is grayed out.

Columns can be hidden also by using the right-click function which brings up the context menu Hide Column when hovering over the heading of a column. To have the column reappear again use Table Column Control

Figure 2-18:

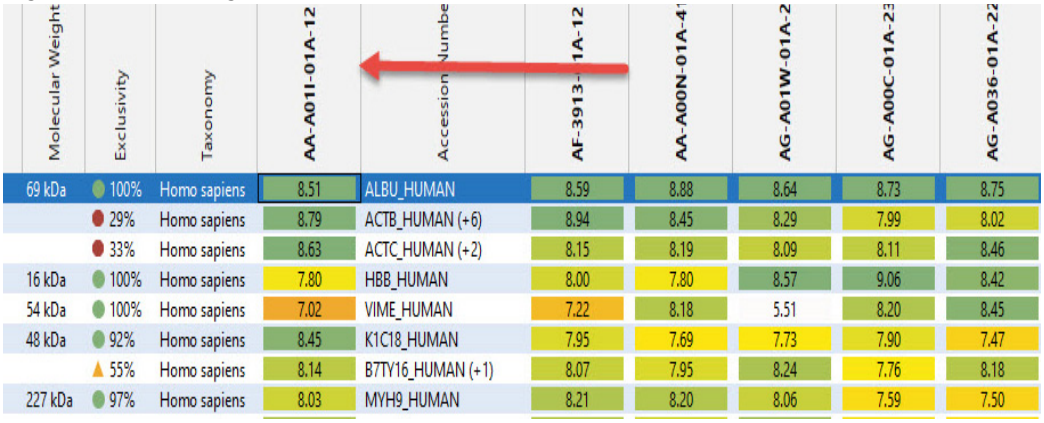


Moving columns

In all tables throughout Scaffold Quant, every column can be moved from one position to another for more comfortable access to the data that is summarized in them.

The user simply clicks on the header of the column that he/she desires to move and drags it to the location where he/she wants to place it. The new position will be retained when the user switches to another view and then returns.

Figure 2-19: Moving columns in tables



Column sorting feature

In all tables throughout Scaffold Quant, the user can use the tri-state column sorting feature

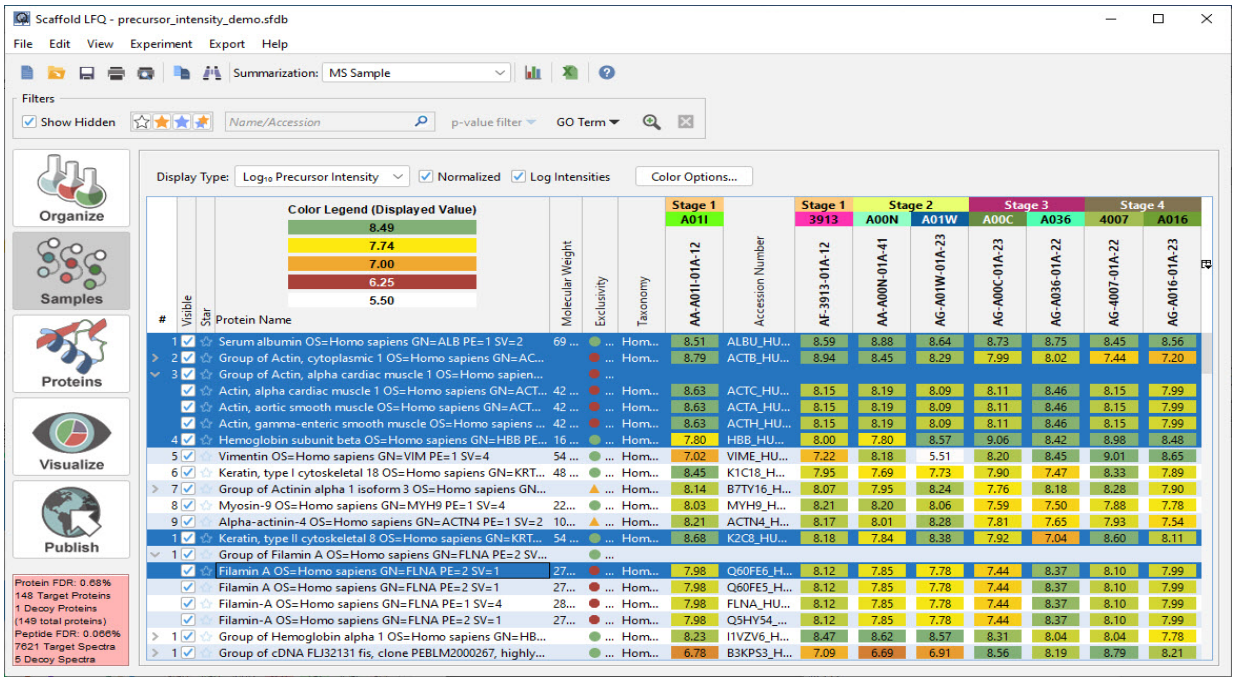
and sort the display by clicking on any column header. For example, to sort the proteins based on their accession number, the user can click the Accession Number column header to initially select the column. Then to sort the proteins based on increasing alphabetical order, the user can click the Accession Number column header a first time. A second click will order the column on decreasing alphabetical order. To return to the default display, the user can click the Accession Number column header a third time.

Increasing and decreasing orders will be indicated by an up and down arrow respectively, shown in the header of the column that is being sorted, while the default order will have no arrow.

Multi selection of rows

In all tables throughout Scaffold Quant the user can select multiple rows by using either the SHIFT or the CTRL key, depending whether the desired selection has contiguous rows or not, and the click of the mouse in a pretty standard fashion. Other functions can then be applied, like assigning a star to the selected group of proteins in the Samples table, for example.

Figure 2-20: Row multi-selection in the Samples Table



Mouse Right-click Context Menus

When the user hits the right-click button of the mouse while hovering over the Display Pane of a View, a menu with various options appears close to the working arrow. Depending on the selected view the list of options available in the menu varies. A description of the mouse right-click command is provided in “Description of Mouse Right Click Context Menu

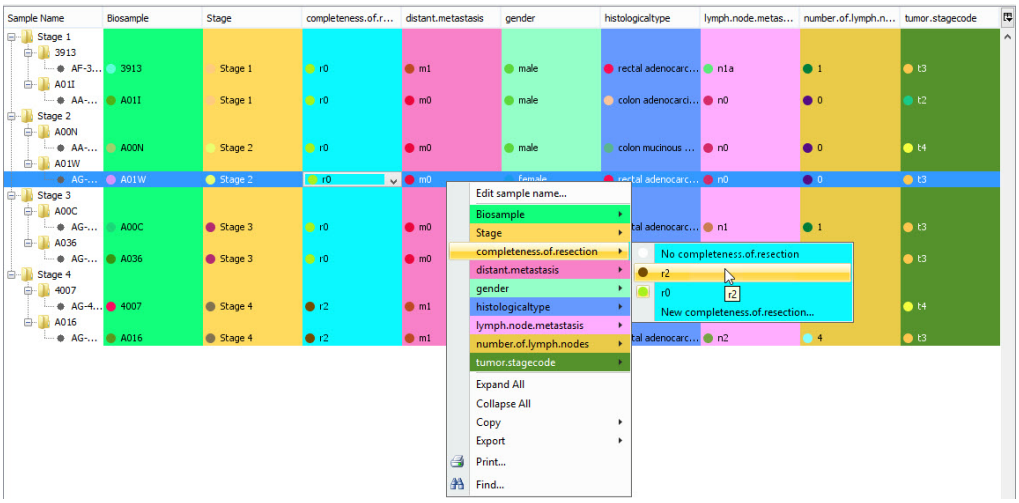
Commands” on page 204.

Organize View

There are two different context menus appearing in this view, one connected to the Sample Organization Table and the other to the Sample Organization tree table.

- When selecting a row in the Organize View Table, Right Click Menu A appears. The menu contains number of commands and a list of the Categories defined in the table. Each attribute group appears in its assigned color and provides in a sub-menu the list of attributes included in it.

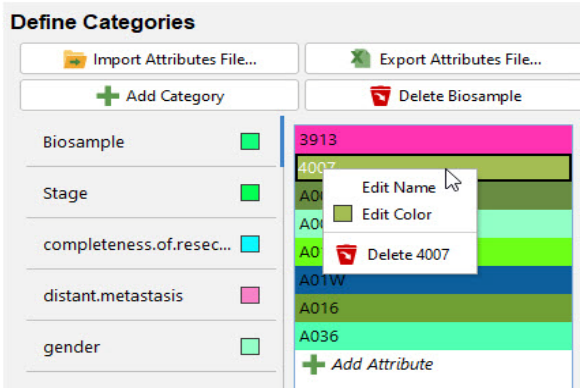
Right Click Menu A



-

- When a row in the Attributes List is selected, a right-click of the mouse brings up the Right Click Menu B . The user may edit the attribute’s name or color or delete the selected attribute.

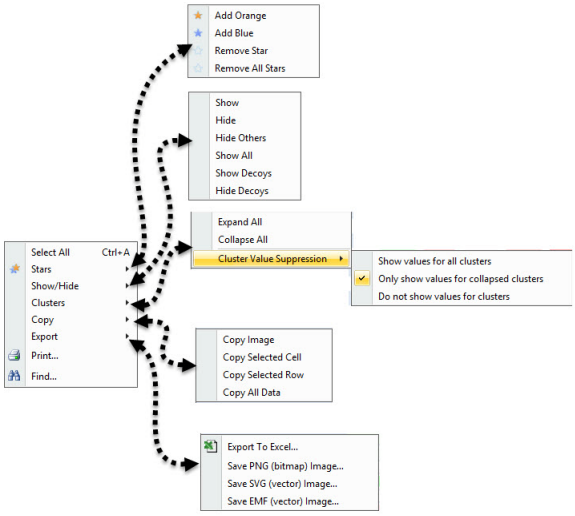
Right Click Menu B



Samples View

When hovering over the Samples table the Right Click Menu C appears. This menu has a number of sub-menus as shown in the picture.

Right Click Menu C



Chapter 3
Loading data in Scaffold Quant

Files supported by Scaffold Quant

Scaffold Quant allows inspection and analysis of peptide and protein identification data from many sources, including Scaffold, Mascot, and other sequence database and spectral library search applications. Scaffold Quant is not designed to combine results from different search engines. It assumes that all identified peptides have been evaluated with a common scoring system. To combine results from different search engines, the user should first load the data into Scaffold and then export it as an SFDB file or in MZID format.

It is possible to view data loaded directly from different search engines, but only in externally-specified grouping and thresholding modes. Attempting to regroup or threshold proteins that do not have a common scoring system will result in errors.

Some Quantitative Methods are available only in SFDB files exported from Scaffold, while others are available when mzIdentML files are loaded directly into Scaffold Quant. Note that labeled Quantitative Methods are only available when a Labeled Quant License (see Licensing) is applied.

mzIdentML files

A common output file format for many search applications is the mzIdentML standard format for proteomics data, developed by the HUPO Proteomics Standards Initiatives. Typically mzIdentML exports create *.MZID files.

A description of the standard specifications is available at the following website http://www.psidev.info/mzidentml.

The Scaffold Quant application creates experiments by loading *.MZID or *.MZID.GZ files version 1.1.0 and higher. It does not need the corresponding peak list files, *.MGF or *.mzML files, but if they are included, spectra will be available for inspection. Loading only *.MZID files dramatically reduces the size of the Scaffold Quant experiment file

Peaklist files are also necessary for processing isobarical labeling data loaded through mzIdentML files with a Labeled Quant License..

*Note that the option to visually inspect a spectrum will be available in Scaffold Quant only if the related *.MGF or *.mzML files were loaded along with the *.MZID files.*

Scaffold Quant support of mzIdentML files

Not all mzIdentMLexports provide sufficient information for analysis with Scaffold Quant.

In addition to identifying peptides through protein sequence or spectral library searches, most search engine applications associate the peptides with specific proteins which contain the identified peptide sequences. They then calculate protein scores that represent the likelihood that each protein is found in the sample based on its associated peptide evidence.

Unlike software such as Scaffold, IDPicker, Mascot or PEAKS, which produce scored lists of proteins from peptide-spectrum-match (PSM) data (from sequence database or spectral library search software), Scaffold Quant does not perform protein assembly or protein scoring. As a result, Scaffold Quant requires that protein scoring information be supplied in the input files.

This means that MZID files from many applications will load into Scaffold Quant, but some will not. Some examples of incompatible tools are SpectraST (through pepXML file conversion), MyriMatch, and Pepitome. Loading a *.mzid file exported from one of these programs into Scaffold Quant will result in an empty protein list, because the file does not indicate that any of the peptide identifications should be considered as evidence for any protein.

How to obtain mzIdentML files

From Scaffold

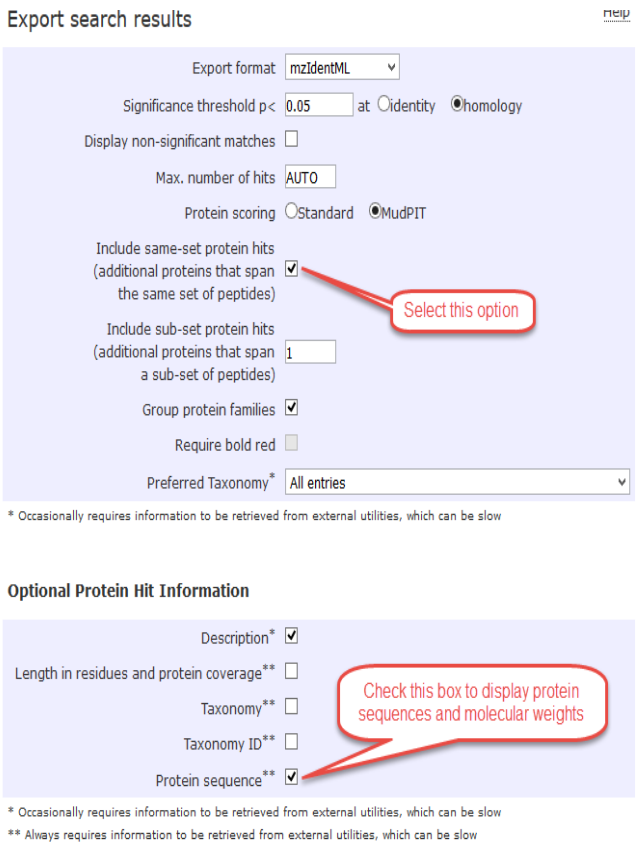
For data analyzed in **Scaffold and Q+S** an mzIdentML export is readily available. *.MZID files are created by going to ***Export > mzIdentML ....*** Two basic options are offered in a small dialog; the suggested option for use with Scaffold Quant is “Scaffold Quant analysis”. By clicking Advanced, it is also possible to further customize the export parameters. Once the desired options are selected, clicking OK brings up a file browser to select a destination for the *.MZID file export. Scaffold creates a new sub-directory that contains *.MZID files and the related *.MGF files. Scaffold Batch also includes commands to create mzIdentML exports from existing Scaffold files or from new files created in Scaffold Batch.

From MASCOT

Mascot creates exports of mzIdentMLfiles from the MASCOT Search Results page. When selecting the option “Protein sequence” the created *.MZID files include information that will allow Scaffold Quant to display sequences and coverage.

When setting up the export, ensure that proper homology information is reported by selecting “Include Same-set protein hits” in the Export Search Results page in the Mascot Server. With this option selected, Mascot properly reports homologous proteins. In Scaffold Quant, the user should apply Experiment >Protein Clustering > Shared Evidence Clustering.... to allow Scaffold Quant to properly group the proteins.

Figure 3-1: Mascot Server Export Search Results page



In order to see decoy hits and calculate FDR, it is necessary to export the decoy reports as separate mzIdentML files. To do so, expand the "Sensitivity and FDR" section of the Protein Family Summary, choose "Decoy Report" and export another mzIdentML report.

Figure 3-2: Creating a decoy report in Mascot

Load the files using one of the methods described in Loading decoy MZID files.

From Proteome Discoverer

Scaffold Quant can load mzIdentML and mzML files exported from Proteome Discoverer 2.4 or higher.

Other compatible exports

MZID files created by PEAKS and IDPICKER can be loaded into Scaffold Quant.

Search engines supported through Scaffold

Currently, MyriMatch *.MZID files are not compatible with Scaffold Quant since they only provide peptide identification, see Scaffold LFQ support of mzIdentML files. Users who want to load MyriMatch data into Scaffold Quant may first load the mzIdentML files into Scaffold and then use either the ***Export>SFDB...*** option to create a Scaffold Quant file or the ***Export > mzIdentML...*** option to create compatible *.mzid files.

In general, sequence database or spectral library search tools that do not provide peptide evidence for protein scores are not supported in Scaffold Quant.

Loading decoy MZID files

There are two possible methods of including decoy *.MZID files in Scaffold Quant, both producing the same results:

1. In the first method, users should place decoy files into the same directory from which they will load an mzIdentML file, eg, *.mzid or *.mzid.gz. The decoy file must have the same name with the word “decoy” appended, e.g. *.mzid.decoy or *.mzid.gz.decoy.

2. Using the second method entails the creation of a directory called decoy within the directory that contains the target MZID files. In the decoy directory, the user should

include the decoy *.mzid file with the same name as the target file. For example, .*.mzid and .\decoy*.mzid or .*.mzid.gz and .\decoy*.mzid.gz.

Loading precursor intensity data

Scaffold Quant processes *.mzid files that include precursor intensity quantitative data. Precursor intensity values must be computed by search engine programs that have the ability to integrate and quantitate precursor intensity peaks, such as Thermo Proteome Discoverer, Mascot distiller and others. Some of these programs may not have the ability to export mzIdentML files, but their output files can be loaded into Scaffold and exported as an SFDB file or the proper MZID files can be exported and loaded into Scaffold Quant. For more information see section Preparing Data for Precursor Intensity Quantitation in Scaffold Quant.

Loading Isobaric Labeling Experiments

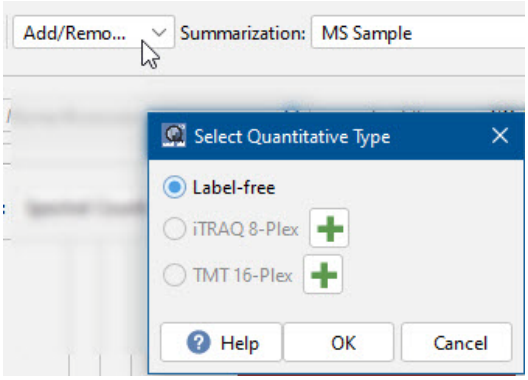
Note: This feature is available only with a Labeled Quantitative License

Data from experiments using Isobaric Labeling, either iTRAQ ® or TMT®, may be loaded in two ways. Which option(s) will work for a given experiment depend on whether the reporter ion intensities were captured by the MS2 or MS3 method.

- **MS2 Data -** In MS2 experiments the reporter ion intensities are specified in the MS2 peaklists. As a result, Scaffold Quant can populate the quantitative data tables directly from the spectra, so MS2 data may be loaded from mzIdentML files accompanied by MGF or mzML peaklist files obtained either from Scaffold or from a search engine which exports mzIdentML files.

  To load mzIdentML files containing MS2 isobaric labeling data, load the mzIdentML files as usual then use the Quantitative Method drop-down to populate the SFDB with the reporter ion intensities. Choose the appropriate quantitative method and click on the green plus icon.

Figure 3-3: .Quantitative Method dropdown

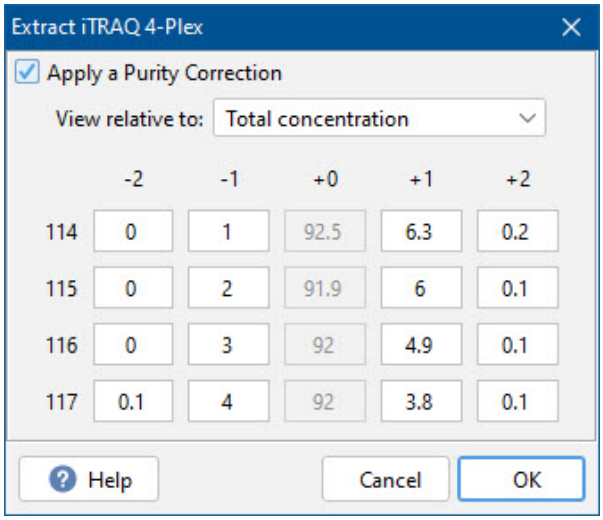


Purity Correction

Where appropriate, Scaffold Quant offers the user the opportunity to apply a purity correction. For iTRAQ, the form is pre-completed with the default purity correction. For

TMT, the form is blank and must be completed by the user. For information about entering a purity correction for TMTpro (16-plex) see "To enter TMT 16-plex purity corrections:" on page 152). The purity correction will be applied as the reporter ions are stored. To populate the quantitative data without applying a purity correction, unclick the checkbox at the top of the dialog.

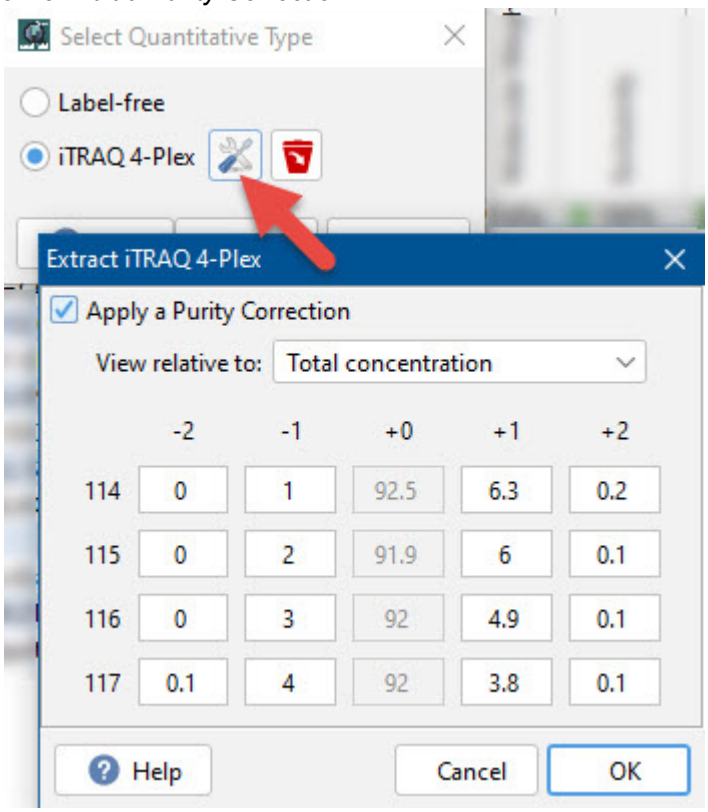
Figure 3-4: Applying a Purity Correction



After the user clicks "OK" the reporter ions are extracted from the spectra, adjusted as necessary to apply a purity correction and written to the quantitative value tables in the SFDB file. At that point, the user may organize and analyze the quantitative samples.

To view or change the purity correction that has been applied, the user may select **Add/Remove Quant Type...** from the Quantitative Method drop-down and select the edit icon.

Figure 3-5: View or Edit a Purity Correction



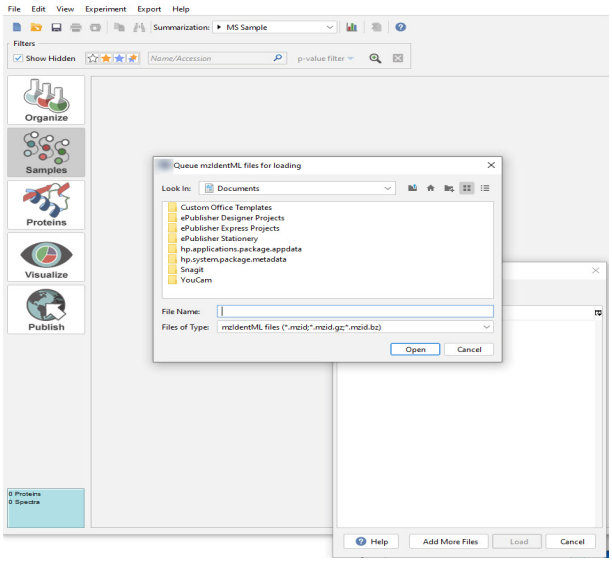
MS2 Data may also be exported from Scaffold in SFDB format and opened directly in Scaffold Quant. In Scaffold-created SFDB files, the reporter ion intensities are already populated, but the user must select the appropriate Quanatitative Method from the drop-down in order to view the quantitative samples.

- **MS3 Data** - Experiments in which the reporter ions are detected in MS3 spectra must be loaded into Scaffold and exported as SFDB files for analysis in Scaffold Quant. Any necessary purity correction must be applied in Scaffold prior to export. It is not possible to alter the purity correction or repopulate the quantitative values within Scaffold Quant. The user may select the appropriate Quantitative Method to view the quantitative samples.

Creating a new experiment

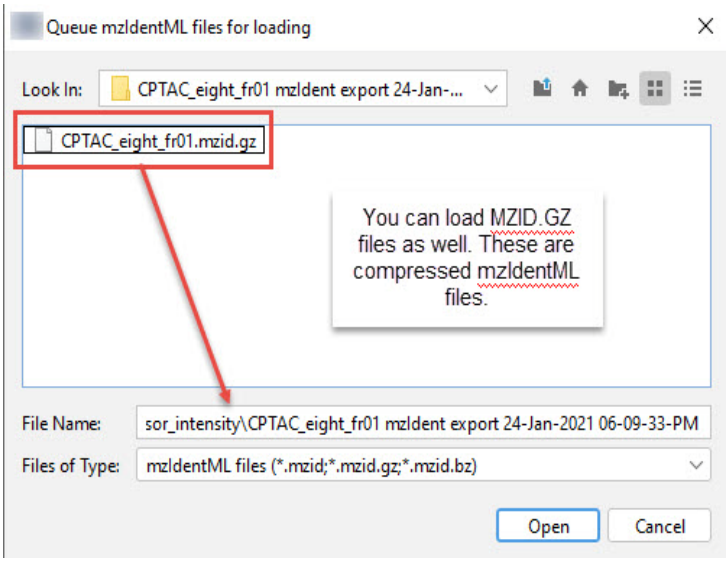
- To create a new experiment in Scaffold Quant, the user can either go to ***File > New*** or click on the “New” icon located in the tool bar below the main menu in the Scaffold Quant window. A dialog box appears asking the user to navigate to the directory where the *.MZID file(s) is (are) located.

Figure 3-6: elect data files dialog box



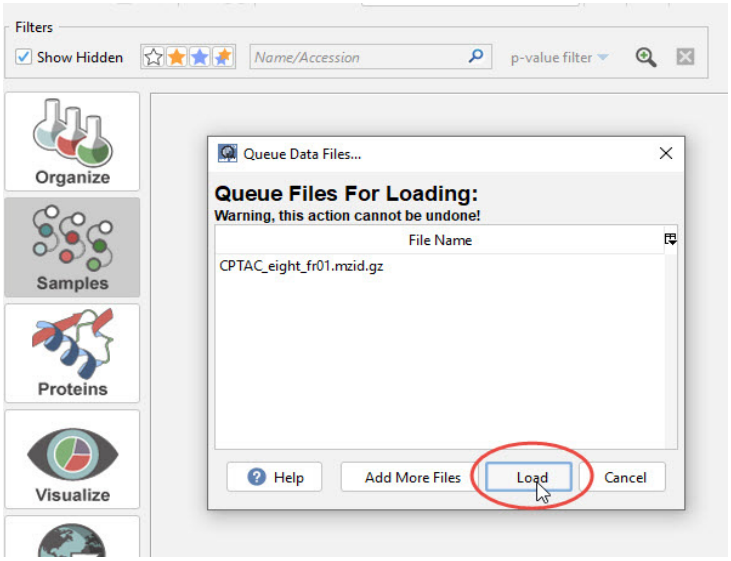
- It is possible to select either a directory that contains the *.MZID files or a compressed directory *.MZID.GZ or a single *.MZID file.

Figure 3-7: electing the file to load



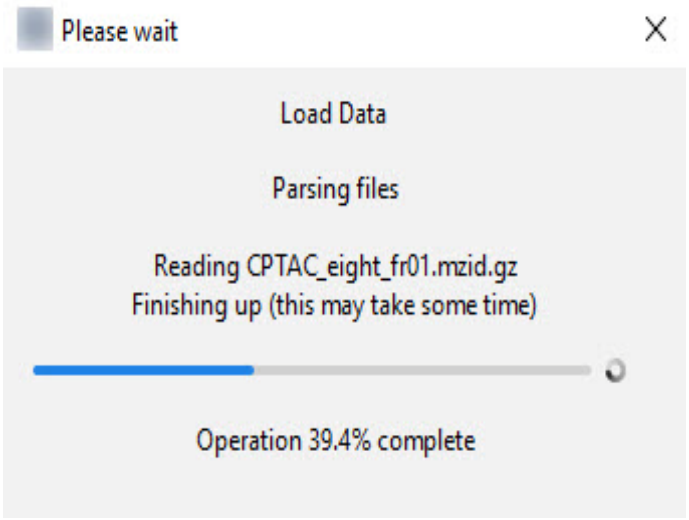
- When Open is clicked, a new dialog box, **Queue Files For Loading,** appears. From this dialog it is possible to choose other files for loading by clicking the button **Add More Files**. All of the selected file names are listed in the dialog box ready to be loaded in Scaffold Quant. Clicking Load starts the operation of loading the listed files into Scaffold Quant.

Figure 3-8: Queue files for loading dialog box



- Depending on the size and number of files in the loading list, the loading operation may take some time to complete. A wait dialog box provides a description of the ongoing operations during the loading phase and a general estimate of the time left to completion.

Figure 3-9: Data loading operation



Scaffold Quant files

Scaffold Quant creates its own file type called SFDB, which stands for Scaffold Database. This file is an SQLite file, a lightweight, high performance SQL database. Indeed, in Scaffold Quant, you can query the experiment using Structured Query Language (SQL) and can also save these queries for future use. This direct access to the data gives Scaffold Quant users a unique capability to manipulate and analyze their data.

Chapter 4
The Organize View

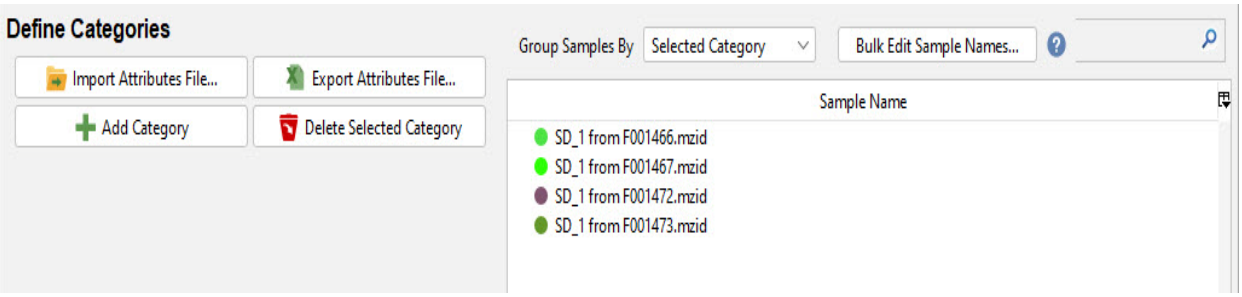
The Organize view displays the loaded samples in a tree structure that can be customized to reproduce the organizational structure of the experiment to be analyzed in Scaffold Quant. It works in conjunction with the “Configure Sample Organization and Statistical Analysis” dialog to support data analysis to expose meaningful biological trends in the experiment.

Organizing data in Scaffold Quant

The Organize View in Scaffold Quant provides an easy method to organize even complex experiments involving multiple variables. A variable or factor in Scaffold Quant is called a Category, and each different level or value assumed by the variable or factor is called an Attribute. Through the Organize View, Categories, along with their Attributes are defined, and the appropriate Attributes are associated with the specific samples to which they apply.

For example, “Treatment Group” might be a Category, with Attributes “Control” and “Treated”. A time-course study measuring response to administration of a drug might have a Category called “Time” with Attributes “0 min”, “20 min”, “40 min” and “60 min”. Many Categories may be applied to the same data. Often many different attributes comprising many categories are recorded for clinical samples, such as age, sex, disease history, etc. All of these may be applied in Scaffold Quant, and the researcher may experiment with analyzing the data on the basis of any one or a combination of these categories.

Figure 4-1: Organize View when files are initially loaded

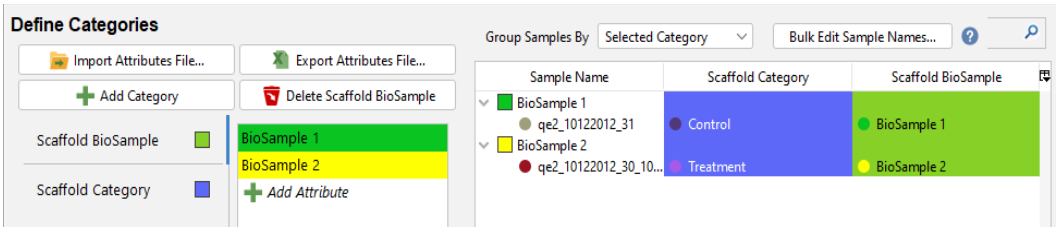


Attributes may be defined and assigned either through the graphical user interface of the Organize View or by reading an Attributes File, which may be created in Excel and saved as a CSV, TSV or .TXT file or exported from a LIMS system.

When files are first loaded into Scaffold Quant, the samples are displayed as originally reported in the input files. Typically, this consists of a simple list of MS samples, but if the files were exported from Scaffold, the Scaffold Biosamples and Scaffold Categories assigned there will be retained. Figure 4-2 shows an example of the Organize View when a series of

*.mzid files from Scaffold have just been loaded:

Figure 4-2: Data from Scaffold as initially loaded into Scaffold Quant



The Tools in the Organize View provide helpful ways to restructure the loaded samples to reflect the proper experimental design. This is done by defining Categories and their associated Attributes and assigning them to the samples. As Categories are added, additional columns are displayed, each corresponding to a Category. As samples are assigned Attributes, they are tagged with customizable labels and colors. Once the Attributes have been associated with the corresponding samples, the user can structure the experiment by creating a hierarchy of categories to view the data at different levels of summarization and can also apply a variety of statistical tests.

Tools in the Organize View

The Organize View consists of:

The Define Categories Pane -- which contains a number of tools to enable the user to provide the meta-data needed to organize the samples in accordance with the experimental design. It consists of:

- **The Import Attributes File... button**--which loads metadata that is already assigned to the samples and stored in a text file structured as a spreadsheet. It creates Categories and assigns Attributes to the currently loaded samples as specified in the file.
- **The Export Attributes File... button**--which saves the current Categories and Attribute assignments as an Attributes File. This is useful for saving Attributes created through the GUI or for exporting a skeletal Attributes File which can be completed in Excel and re-imported to create and assign Attributes.
- **The Add Category button** --which creates a new Category. When this button is clicked, a new Category is added at the top of the Category list below the button. By default, it is named “New Category”

Figure 4-3: On the left, a new Category has been created. On the right, the new Category has been named, which causes it to assume its proper place in the alphabetically

sorted list of categories. The next step is to click Add Attribute to the right and create the appropriate Attributes for this Category

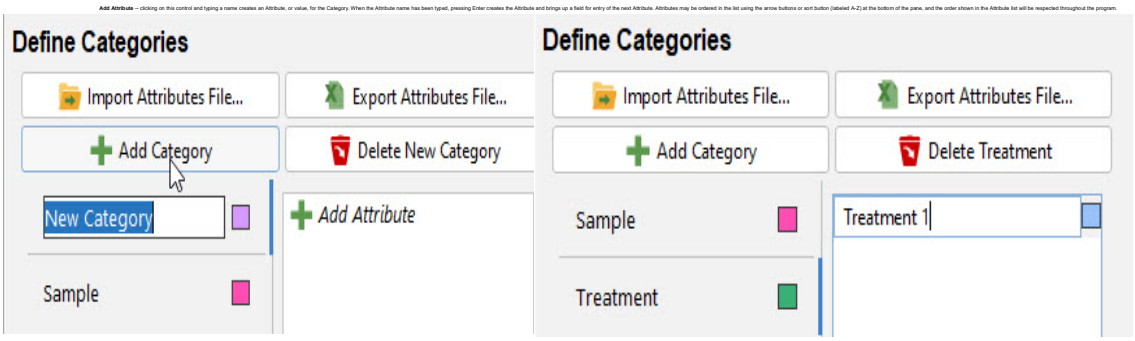
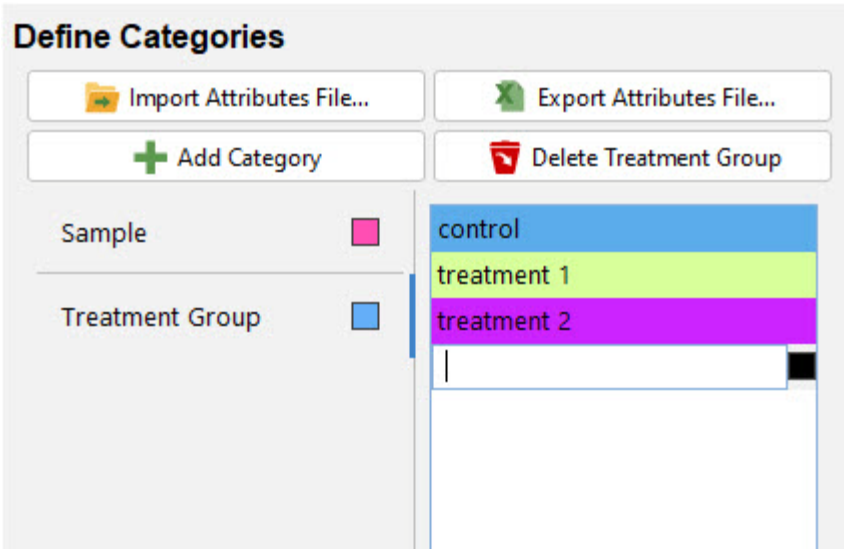


Figure 4-4: The completed Attribute List for the new Category



- **The Delete Category button** -- clicking this button deletes the selected Category and all Attributes associated with it. To delete a single Attribute, select the Category in which it appears, then select it in the Attribute list and use the delete key or right-click and delete.
- **Configure Experimental Design and Statistical Analysis** -- when all samples have been organized according to their attributes, the user should click this button to open a dialog which allows for specification of the design of the experiment. This will establish a Summarization Hierarchy and allow the user to configure statistical analysis (see Specifying the Design of an Experiment).

The Sample Organization Pane - allows the user to assign attributes to the samples. It consists of:

- **Group Samples By** -- a dropdown list which allows the user to determine how the samples will be displayed in the Sample Organization Table. Options are 1) to display the samples sorted by Sample Name, 2) to group the samples based on the Selected Category (the currently selected in the Category List under Define Categories), or 3) to display them

hierarchically, organized according to the Experimental Design (see). Each of these options can be useful at different points in the process of organizing the experiment.

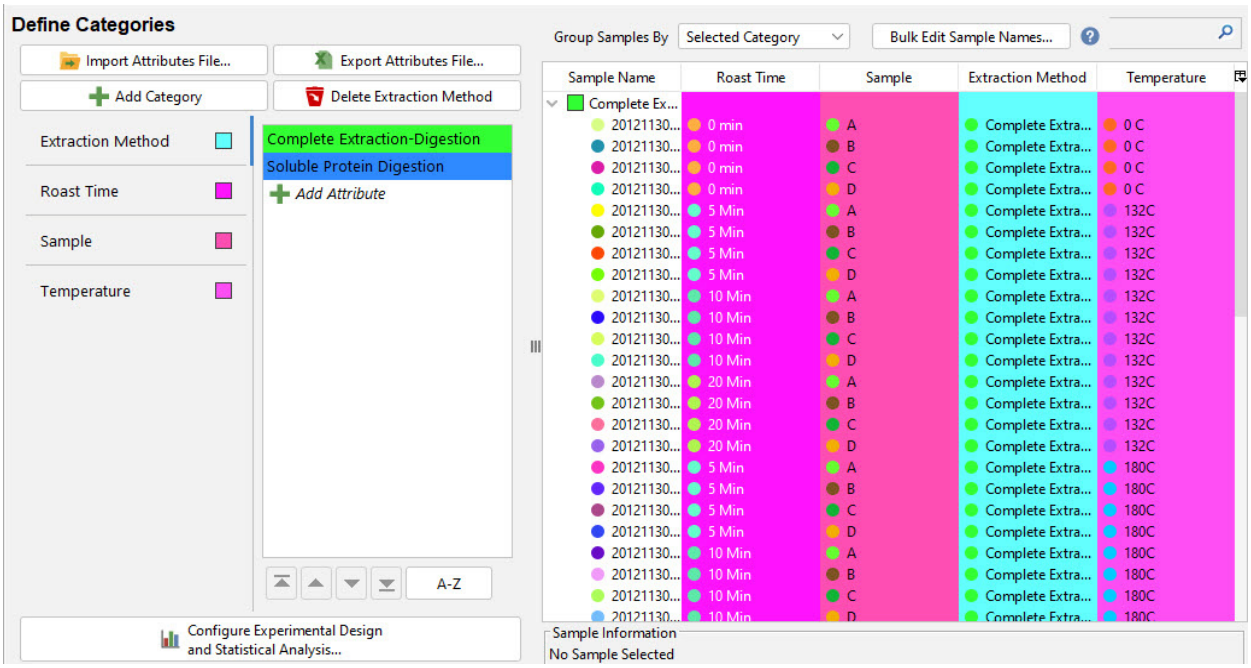
- **Bulk Edit Sample Names**... -- a button which brings up a dialog to assist the user in editing sample names to make them more useful and legible throughout the program. It provides a number of options for trimming the names or allows individual editing if the custom option is selected.
- The **Sample Organization Table** -- which lists the MS samples loaded into the experiment, organized as a tree structure and displays the Attributes associated with each sample.
- **Sample Information** -- displays sample information for the sample currently selected in the Sample Organization Table.
- The **The Configure Sample Organization and Statistical Analysis Dialog** button which allows creation and editing of the Summarization Hierarchy.

Sample Organization Table

Scaffold Quant allows the user to derive a much deeper understanding of the experiment by creating new Categories and then assigning their Attributes to the appropriate MS samples. The Categories may be hierarchically organized using the Summarization pane, described in Specifying the Design of an Experiment. Figure 4-5 shows the data after a series of attributes has been applied.

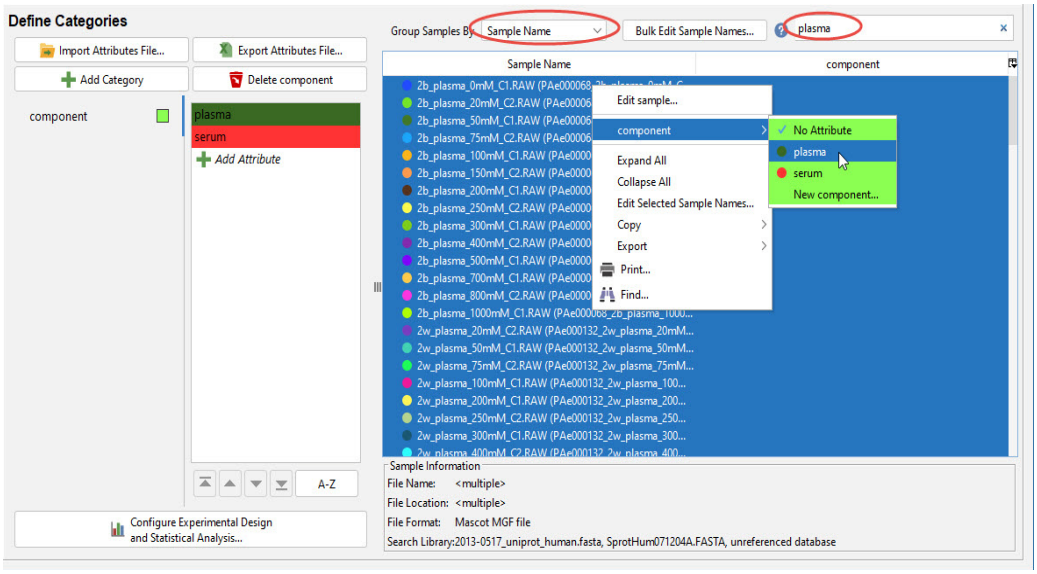
Figure 4-5: Organize View with added attributes

There are several methods by which Attributes may be assigned to samples.



- Drag and drop - Select the Category whose attributes are to be applied by clicking on it in the list under Define Categories. Select Group Samples By Selected Category Drag an individual attribute in the Sample Organization tree table to a sample name. Alternatively, select one or more samples and drag them to one of the Attributes, which appears in the table as the Attribute name and its color block.
- Right-click on Sample(s) - This option is often used in Display Samples by Sample Name mode, but also works in the other modes. Select one or more samples in the Samples column and right-click. Hover over an Attribute Group in the context menu that appears, then select an Attribute to assign to all selected samples.
- Search Box - A helpful method when organizing large experiments is to use the Search Box to display a subset of samples, select them all then either right-click and select an Attribute or drag the Attribute to the set of samples. Often the sample names contain substrings that indicate which attributes belong to which samples. In this case, the Search Box approach allows the user to leverage this information to quickly organize the samples.

Figure 4-6: *Assigning an Attribute to multiple samples using the Search Box*



Sample Organization tree table

When the Group Samples By option is set to Sample Name:

- The table shows the list of samples in alphabetical order in the first column. A colored dot next to the sample name indicates the color associated with that sample throughout the program. The sample name or color may be edited by right-clicking in the cell and selecting Edit sample...

- An additional column is displayed for each Category that has been created, and the cells in these columns show the Attribute associated with the specific sample in that Category. A colored dot indicates the color associated with that Attribute throughout the program.

When the Group Samples By option selected is Selected Category or Experimental Design:

- The table shows the list of Categories as collapsible folders:

  When the folders are collapsed a + sign appears to the left side of the folder. Clicking the + sign expands the folder showing the list of attributes in the group and the +sign becomes a -. Clicking the - sign collapses the folder.

- When the Attribute Group is expanded, the Attributes belonging to the group are listed with a colored dot assigned to each of them.

Right clicking on an Attribute Group folder or on an Attribute in the table:

- Right-clicking on an Attribute Group or Attribute in the tree table:

  This action displays a menu that allows the user to edit or delete the Attribute Group. The Edit Name option makes the attribute name editable and the Edit Color color option opens a dialog that allows the user to select a new color to be assigned to the Attribute.

Organizing Quantitative Samples

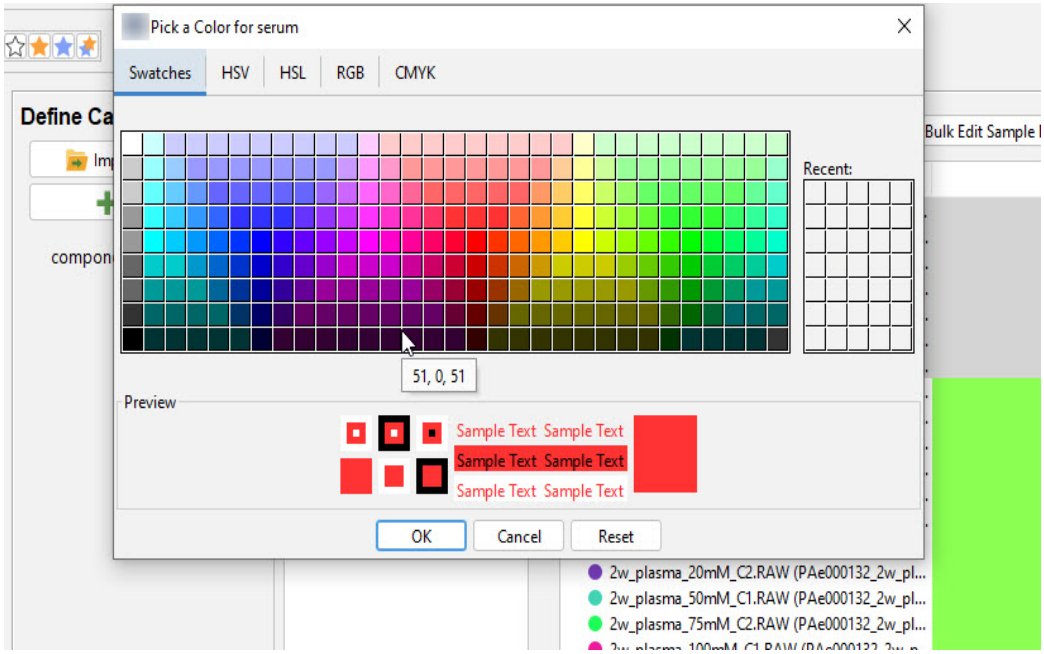
Note: Feature available only with a Labeled Quantitative License

In a multiplexed quantitative experiment, each channel represents a separate quantitative sample. When Scaffold Quant is registered with a Labeled Quant License and multiplexed labeled data has been loaded, the program enables the user to organize and analyze the quantitative samples, rather than the MS Samples. The Quantitative Method drop-down is used to switch between these modes. When a multiplex quantitative method is selected, the Organize View displays Quant Samples and the user may rename them, categorize them, and specify the experimental design based on these categories. This is done in the same way as for MS Samples in label-free experiments (see Organizing data in Scaffold Quant).

Editing Category Colors

Organize View - Edit Category Colors

Figure 4-7: Organize View - Edit Category Colors



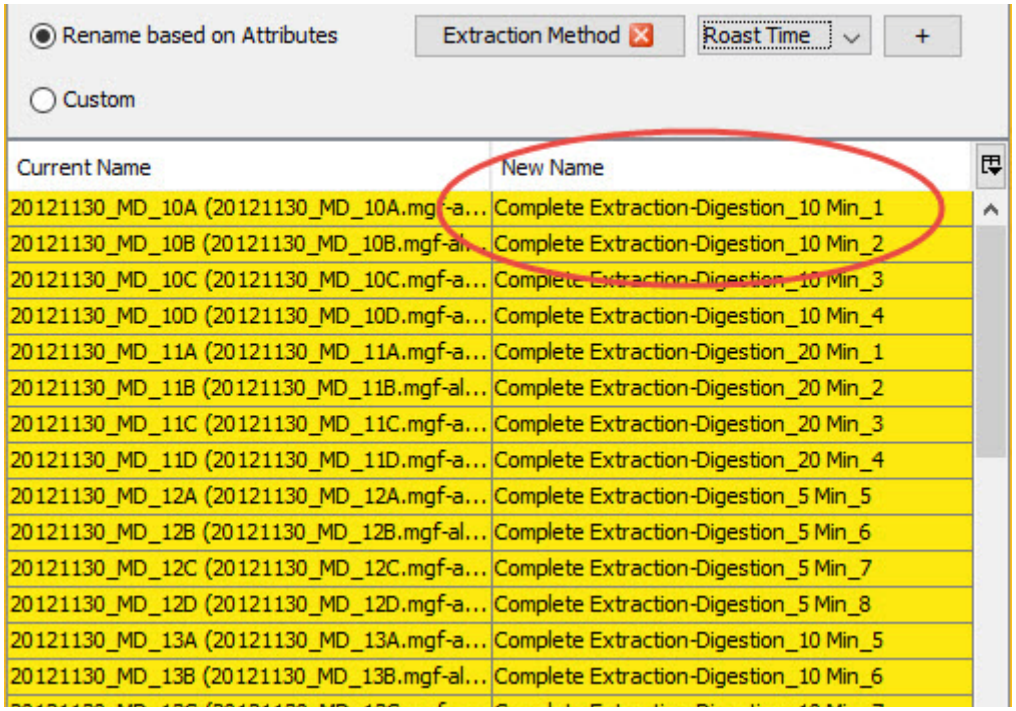
- If Group Samples By Selected Category is selected, all of the samples are shown in folders grouped by the selected category. Each folder contains all samples with a specific Attribute of that Category. If Experimental Design is chosen as the grouping method, samples are organized into a hierarchical set of folders based on the Experimental Design (see Specifying the Design of an Experiment).

Bulk Edit Sample Names

Often sample names are quite long and difficult to distinguish. This can cause problems when viewing the data in the other Views of Scaffold Quant. The user may wish to edit the sample names to make them shorter and/or more meaningful. Several tools are provided to assist in this effort. The names as they appear currently are shown in the left column of the table, and as they would appear if a proposed edit were applied on the right. Buttons at the bottom allow the user to Apply an edit, Cancel an edit or close the dialog (OK). The editing tools provided are:

- Remove prefix - if all sample names share a common prefix, it will appear in the text field. It may be edited if the user wishes to remove just a portion of the common prefix.
- Remove suffix - if all sample names end in a common suffix, it will appear in the text field. It may be edited to allow removal of a portion of the common suffix.
- Remove characters at beginning - the user may select a specific number of characters to be removed from the beginning of all sample names.

- Remove characters at end - the user may select a specific number of characters to be removed from the end of all sample names.
- Rename based on Attributes - the user may select a Category from the dropdown. Samples will be renamed to the Attribute name associated with that sample for that Category appended with a sequential number. Clicking the + button will add a second Attribute from another Category (see Figure •).
- *Samples renamed based on a combination of two Categories*Custom - allows selection



and editing of individual names in the New Name column.

Import Attributes File...

A quick way to apply a number of Attributes to MS samples already loaded into Scaffold Quant is to read them from a formatted list of Attributes saved as a tab delimited text file, see Compiling an Attributes File. The file can be imported through the I**mport Attributes file...** button or by selecting the **Experiment > Load Attributes from File** command from the main menu.

If some Attributes have already been applied, importing an Attributes File will update assignments of existing Attributes listed in the file, but will not affect any Attributes that are not included in the file. Thus an Attributes File may be used to add to existing Attribute assignments, or to update them.

Compiling an Attributes File

The first line, or header line, in a Scaffold Quant attributes file begins with **Sample Name** followed by a list of Attribute Group names. If they do not already exist, these Categories will

be created when the file is imported.

Each successive line must begin with the name of a sample loaded into the program followed by Attributes, each belonging to the Attribute Group listed above it in the header line. Note that the sample names must be precisely the same as the sample names loaded into Scaffold Quant.

One method of creating an Attributes File is to export a skeletal file containing the sample list from the program and then add the Attribute information for each sample using Excel or a similar program. The list of loaded samples can be compiled by clicking the **Export Attributes File...** button or by selecting the **Export > Export Attributes File...** command from the main menu. Once the exported file is opened in Excel, it is easy to add attribute information to each sample in the list. The top row, or header line, will begin with SAMPLE NAME, and the names of the desired Categories, should be added. The list of samples must be the first column in the file, and the Attributes should be added in the subsequent columns (see Figure 4-8below).

Figure 4-8: Example of a Scaffold Quant Attributes text file

	Sample Name	Biosample	Category	Ethnicity	Anticoagulant	HPLC
2	CAK-20-400-900.RAW (F002703)	PAe000817	Lab-1	b1	edta	20-400-900
3	CAS-40-900-1200.RAW (F002698)	PAe000810	Lab-1	b1	serum	40-900-1200
4	CAH-40-900-1200.RAW (F002760)	PAe000862	Lab-1	b1	heparin	40-900-1200
5	CAH-20-900-1200.RAW (F002757)	PAe000862	Lab-1	b1	heparin	20-900-1200
6	CAC-10-400-900.RAW (F002743)	PAe000859	Lab-1	b1	citrate	10-400-900
7	CAC-40-900-1200.RAW (F002751)	PAe000859	Lab-1	b1	citrate	40-900-1200
8	CAH-20-400-900.RAW (F002756)	PAe000862	Lab-1	b1	heparin	20-400-900
9	CAS-20-1200-200.RAW (F002696)	PAe000810	Lab-1	b1	serum	20-1200-200
10	CAK-40-400-900.RAW (F002706)	PAe000817	Lab-1	b1	edta	40-400-900
11	AAS-10-400-900.RAW (F002734)	PAe000797	Lab-1	b3	serum	10-400-900
12	CAS-40-400-900.RAW (F002697)	PAe000810	Lab-1	b1	serum	40-400-900
13	CAC-40-1200-2000.RAW (F002752)	PAe000859	Lab-1	b1	citrate	40-1200-2000
14	CAC-20-400-900.RAW (F002747)	PAe000859	Lab-1	b1	citrate	20-400-900
15	CAS-10-900-1200.RAW (F002692)	PAe000810	Lab-1	b1	serum	10-900-1200
16	CAK-40-1200-2000.RAW (F002708)	PAe000817	Lab-1	b1	edta	40-1200-2000
17	CAH-40-400-900.RAW (F002759)	PAe000862	Lab-1	b1	heparin	40-400-900
18	CAH-10-400-900.RAW (F002753)	PAe000862	Lab-1	b1	heparin	10-400-900
19	CAH-10-1200-2000.RAW (F002755)	PAe000862	Lab-1	b1	heparin	10-1200-2000
20	CAK-20-1200-200.RAW (F002705)	PAe000817	Lab-1	b1	edta	20-1200-200
21	CAC-20-1200-200.RAW (F002749)	PAe000859	Lab-1	b1	citrate	20-1200-200
22	AAS-40-400-900.RAW (F002740)	PAe000797	Lab-1	b3	serum	40-400-900
23	CAK-10-1200-2000.RAW (F002702)	PAe000817	Lab-1	b1	edta	10-1200-2000
24	CAH-40-1200-2000.RAW (F002761)	PAe000862	Lab-1	b1	heparin	40-1200-2000
25	AAS-20-400-900.RAW (F002737)	PAe000797	Lab-1	b3	serum	20-400-900
26	AAS-10-900-1200.RAW (F002735)	PAe000797	Lab-1	b3	serum	10-900-1200
27	CAS-20-400-900.RAW (F002694)	PAe000810	Lab-1	b1	serum	20-400-900
28	AAS-20-1200-200.RAW (F002739)	PAe000797	Lab-1	b3	serum	20-1200-200
29	CAS-40-1200-2000.RAW (F002699)	PAe000810	Lab-1	b1	serum	40-1200-2000
30	CAH-10-900-1200.RAW (F002754)	PAe000862	Lab-1	b1	heparin	10-900-1200
31	CAS-20-900-1200.RAW (F002695)	PAe000810	Lab-1	b1	serum	20-900-1200
32	CAC-10-400-900.RAW (F002744)	PAe000859	Lab-1	b1	citrate	10-400-900
33	AAS-40-900-1200.RAW (F002741)	PAe000797	Lab-1	b3	serum	40-900-1200
34	CAS-10-400-900.RAW (F002691)	PAe000810	Lab-1	b1	serum	10-400-900

If the Scaffold Quant file already contains attribute data, such as Category and Biosample, these Categories do not need to be added again.

Important: Open the file in Excel, add the Attributes and then export from Excel as comma- or tab-delimited text file.

Experimental Design

Supported experimental designs

Scaffold Quant supports quantification and statistical analysis for several types of experimental designs

Basic Design

Experiments of this design consist of two or more biological classes of MS samples, between which variation is considered to be caused by experimental conditions (e.g. control and treatment classes). Each biological class is made up of one or more biological replicates which represent identical experimental conditions, but differ from one another because of biological variation (e.g. multiple organisms raised under identical experimental conditions would be biological replicates of one another). A biological replicate, in turn, consists of one or more technical replicates which originate from the same biological source. Finally, a technical replicate may be fractionated, in which case it is a set of MS samples which correspond to different portions of a sample which contain (ideally, but not exactly) distinct subsets of the total content of the originating sample. A non-fractionated technical replicate is simply a single MS sample.

Figure 4-9: Example experimental design

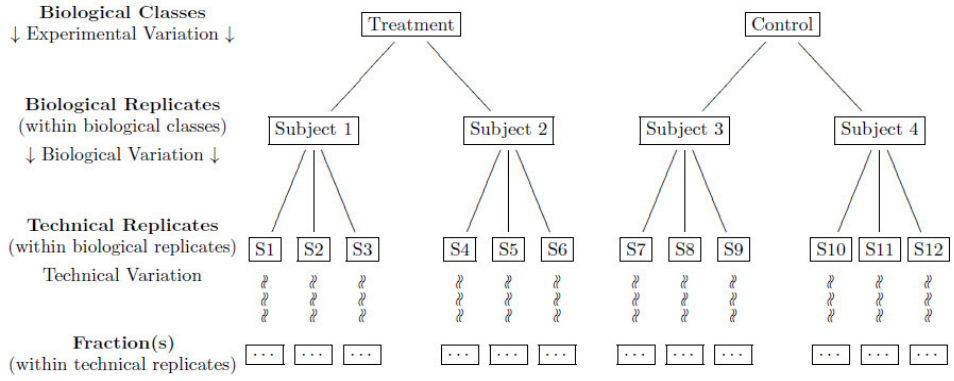


Figure shows an example experimental design which can be easily represented in Scaffold Quant with two categories in a simple hierarchy, for example:

Group -> {Treatment, Control}

Subject-> {Subject 1, Subject 2,Subject 3,Subject 4}

can be summarized with the following hierarchy:

Group

Subject

(MS sample)

Repeated Measures

In a repeated measures experiment, samples are obtained at different times or under different conditions from the same biological entities. The goal is to analyze how each individual's levels change in response to the varying conditions. Each individual may have its own baseline value, but the goal is to analyze whether there are patterns in the changes from these baseline levels in response to the changing conditions.

Some examples of repeated measures studies are time-course studies and crossover studies. Time-course studies are used, for example, for measuring the response of individuals to a drug treatment. Initial baseline levels are measured, then measurements are taken at a series of time points to ascertain the pattern of response to the drug. In crossover studies, a set of individuals are exposed to a series of different treatments, with each individual receiving each treatment, although not necessarily in the same order.

The Samples Hierarchy in experiments of this type may contain biological replicates, technical replicates and/or fractions, as described in Basic Design above.

Scaffold Quant requires that the experiment is complete, meaning that a sample is provided for each individual at each time point or in each condition.

Two-way Design

A two-way analysis compares the effects of two independent variables. It can help in determining whether there is an interaction between these two variables. For example, a study might compare the reaction of males and females to the administration of a drug. A two-way design would allow the researcher to test whether males and females respond differently.

When a two-way ANOVA test is applied to an experiment, three different measures are produced, each of which tests a different hypothesis. One measure assesses the degree to which the two factors interact, the second measures the effect of the category designated as the primary factor, and the third measures the effect of the category designated as the secondary factor. In Scaffold Quant, the user may select which of these measures should be displayed in the test result column.

The Samples Hierarchy in experiments of this type may contain biological replicates, technical replicates and/or fractions, as described in Basic Design above.

Scaffold Quant currently requires that a two-way study be balanced, meaning that each combination of factors is represented by the same number of samples.

Randomized Block Design

Randomized Block is a specific type of Two-way analysis in which samples are divided into groups called blocks. Blocking compensates for situations in which known factors (e.g. age, sex) other than treatment group status are likely to affect what is being observed in the study[1]. The Randomized Block ANOVA measures the treatment effect while minimizing the effect of the blocking category; unlike the two-way ANOVA, it does not provide any assessment of the effect of the blocking category. Scaffold Quant only supports complete randomized block designs, i.e. those which contain one value per cell in the Design Matrix.

Specifying the Design of an Experiment

After samples have been organized into Categories (see Organizing data in Scaffold Quant), it is important to organize the Categories to reflect the design of the experiment. This is accomplished through the Configure Sample Organization and Statistical Analysis dialog.

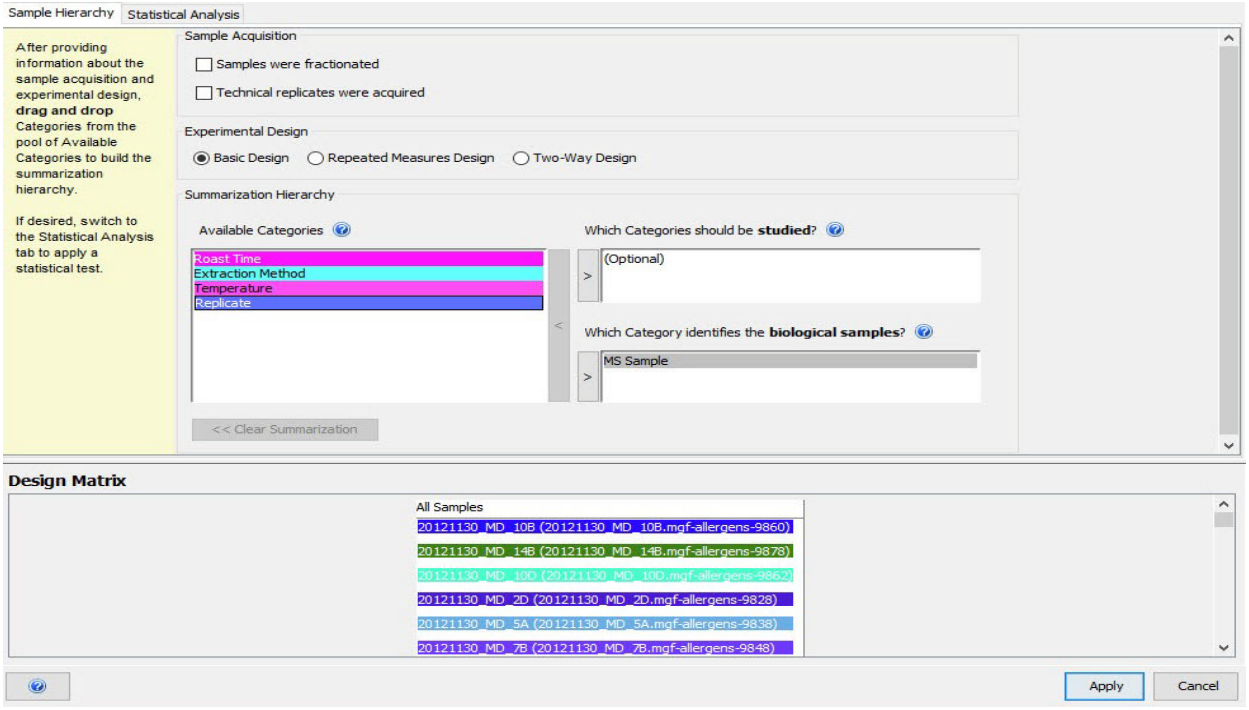
The Configure Sample Organization and Statistical Analysis Dialog

This dialog may be opened by:

- Clicking the Configure Experimental Design and Statistical Analysis... button at the bottom of the Organize View
- Selecting the menu item Experiment>Quantitative Analysis...
- Clicking on the Quantitative Analysis icon in the toolbar
- Selecting Edit from the list in the Summarization dropdown

1. https://www.statsdirect.com/help/analysis_of_variance/randomized_blocks.htm

Figure 4-10: Configure Sample Organization and Statistical Analysis dialog, *initial state*



The Configure Sample Organization and Statistical Analysis dialog includes two tabs:

- Sample Hierarchy Tab -- allows the user to specify the type of experiment to be analyzed and the roles of the various Categories in the analysis.
- Statistical Analysis Tab -- presents the various statistical tests available for analyzing the experiment as it has been specified in the Sample Hierarchy tab, and allows the user to select the test and specify its parameters.

Sample Hierarchy Tab

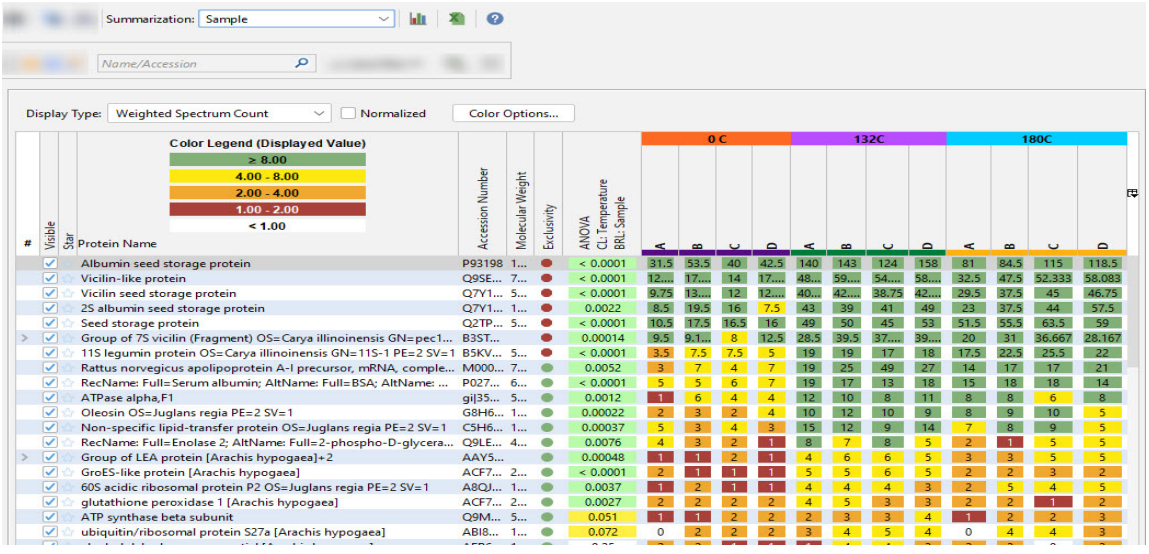
The upper portion of the Sample Hierarchy Tab consists of three sections:

- Sample Acquisition - this portion consists of two checkboxes:
  - Samples were fractionated - should be checked if the samples were fractionated, e.g. if they were separated on a 2-D gel. This will allow quantitative values for peptides detected in different samples to be combined as they would be in a MuDPIT experiment.
  - Technical replicates were acquired - should be checked if technical replicate samples were gathered, e.g. if biological samples were aliquoted and the aliquots were analyzed as separate MS Samples. If samples are designated as technical replicates, their proteinprotein-level quantitative values are first normalized and then summed to give a total value for the biological sample.

- Experimental Design - this section provides options for defining the basic structure of the experiment.
  - Basic Design - this option should be selected if the user simply wishes to view the MS results without performing any statistical analysis, or if a simple comparison based on a single Category is to be carried out (see Basic Design). This allows performance of, e.g., a T-Test or ANOVA.
  - Repeated Measures - this option should be selected if samples from each biological subject have been analyzed under different conditions or at different time points (see Repeated Measures).
  - Two-way Design - this option should be selected if the data is to be analyzed on the basis of two Categories (see Two-way Design). For example, a study might assess the differential effect of a treatment on males and females.
- Summarization Hierarchy - this section allows the user to specify how quantitative values should be summarized based on the Categories. The Available Categories are shown on the left. On the right are a series of boxes used to assign the Categories to the different analysis levels required by the experimental design. Different boxes are displayed depending on the experimental design type and whether or not there are technical replicates and fractionation.

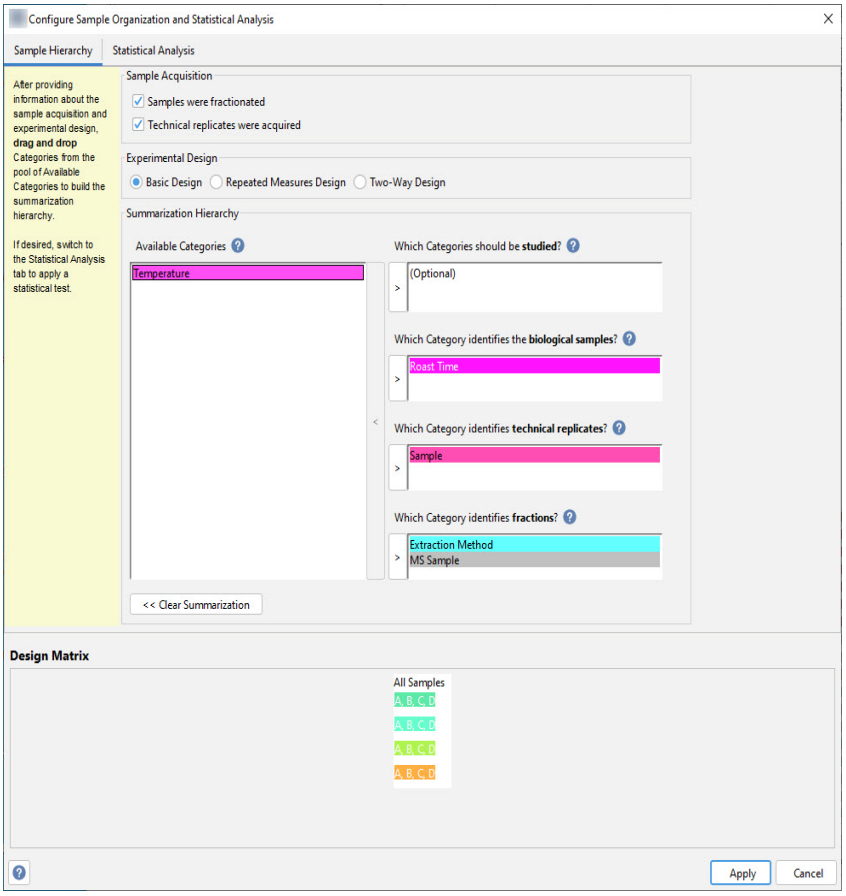
  The Summarization Hierarchy determines how the data may be viewed in the Samples View as well as how it will be analyzed in statistical tests. Once Categories have been assigned to different levels in the Summarization Hierarchy, quantitative values may be "rolled up" or summarized to any of the levels for display and analysis.

Figure 4-11: The Samples View with a Summarization Hierarchy defined



The Sample Hierarchy dialog when Basic Design is selected, showing fractions and technical

replicates



Each Category may be moved from the Available Categories list on the left to an appropriate box on the right either by dragging and dropping or by using the left and right arrows on the boxes. To add a Category to a box, select the Category in the Available Categories list, then click the right arrow on the box. To return a Category to the Available list, select the Category in a box and click the left arrow on the Available Categories box.

As Categories are moved from the Available Categories List to their assigned roles in the experiment, A table is constructed in the lower pane of the Configure Sample Organization and Statistical Analysis dialog. Samples are placed into rows and columns to indicate how they will be grouped for evaluation in statistical testing.

The Design Matrix can help the user to visualize the experiment and verify that the experiment has been set up correctly. When viewed from the Statistical Analysis tab, the column and row headers also contain check boxes. Using these boxes, the user may select which rows and columns should be included in the test. Unchecking a box excludes that row or column from consideration when the test is applied.

When a statistical test has been applied, if the summarization level is set to the level representing Biological Samples, colored bars appear in the column headers in the Samples

View to indicate the Comparison Groups used in the statistical test. In the Samples Report, the comparison groups are indicated by numbers in parentheses in the column headers.

For the Basic Design:

The user should specify:

- Which Categories Should be Studied: if no Categories are moved into this box, it will not be possible to apply a statistical test and no summarization above the level of the biological samples will occur. If one Category is selected for study, statistical comparisons between groups of samples corresponding to the different Attributes of that Category may be made, and values may be summarized to the Category level. If more than one Category is selected, statistical comparisons will operate on groups representing each possible combination of attributes for those Categories, and summarization may proceed up through the levels specified.

  For example, if both Extraction Method and Roast Time were selected for study, ANOVA would compare Complete Digestion for 0 min., Complete Digestion for 5 min., Soluble Digestion for 0 min, Soluble Digestion for 5 min. etc. Data could be viewed at the MS Sample level, the Replicate level, the Roast Time level or the Digestion Method level.

Selection of two Categories to be studiedSelection of two Categories to be studied

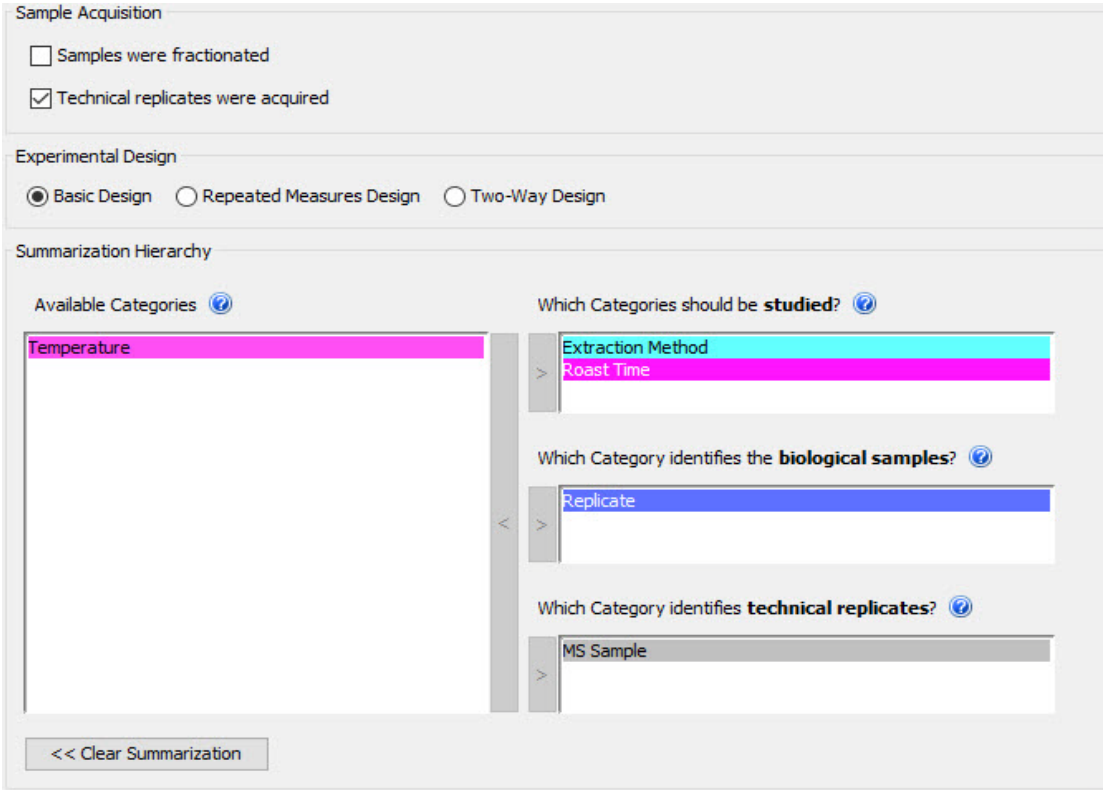
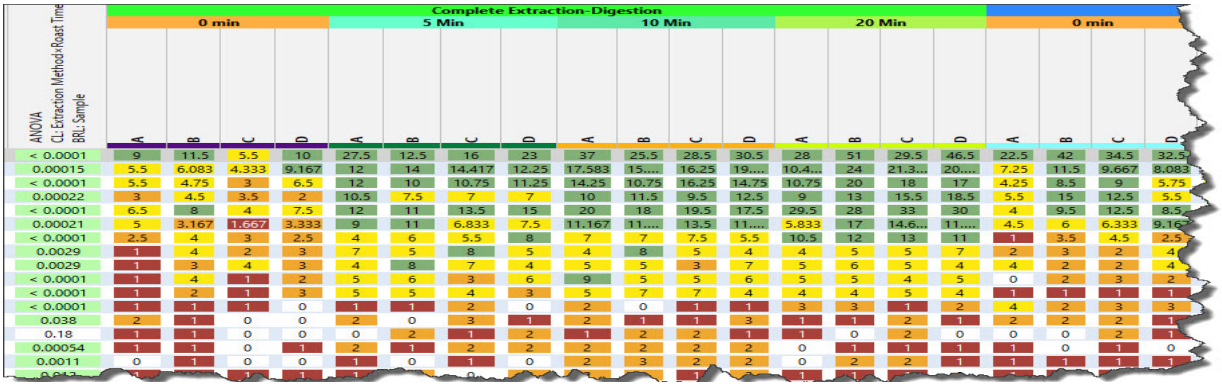


Figure 4-12: The resulting Samples View, shown at the Replicate level



- Which Category identifies the biological samples: the Category that identifies the biological subject should be placed into this box. This Category will be used as the blocking level in statistical analysis. This means that the values as they appear when rolled up to this level are used in statistical test calculations. In the Design Matrix, the biological subject defines the cells or blocks. More than one Category may be entered into this box, in which case each combination of the values in these Categories will define a biological sample.
- Which Category identifies technical replicates: if each biological sample has been divided and the resulting sub-samples have been analyzed in the mass spectrometer separately, the sub-samples are technical replicates. The Category which names these sub-samples should be entered here. In some cases, the sub-samples may have undergone different treatments or have been obtained separately from the same biological subject, but in Scaffold Quant, they should be considered as technical replicates if the intent is to use them as multiple measurements of the same biological subject.

Which Category identifies fractions: this field appears only if the fractionation check box is checked. This indicates that the MS Samples should be combined and treated as if they were a single MS Sample. For example, if biological samples were separated on a 2-D gel and individual gel spots were analyzed separately, the MS Samples should be classified as fractions so that the peptides detected in them can be combined for purposes of protein identification. This produces the same effect as the classic MuDPIT technique.

Clicking Apply finalizes the creation of the summarization level pull down list appearing in the summarization pane in the Scaffold Quant main window.

For the Repeated Measures Design:

- The first Category to be specified is the Time or Repeated Measure Category. The Time Category defines which of the repeat groups a sample represents. For instance, the Time Category could be Time Point, with values of 0 hr, 1hr, 2 hr and 3 hr. It need not represent time, however. For example, in a study that measures each subject's reaction to various treatments, it might be Treatment.

- Other Categories to be specified are similar to those described in "For the Basic Design:".

For the Two-Way Design:

Two categories should be selected for study. One will be designated as the Primary Analysis Category, while the other will be termed the Secondary Analyis Category. Even though one is designated as primary, the user may choose which category to assess with a Two-Way ANOVA test without changing the sample hierarchy. As a result, the user should specify:

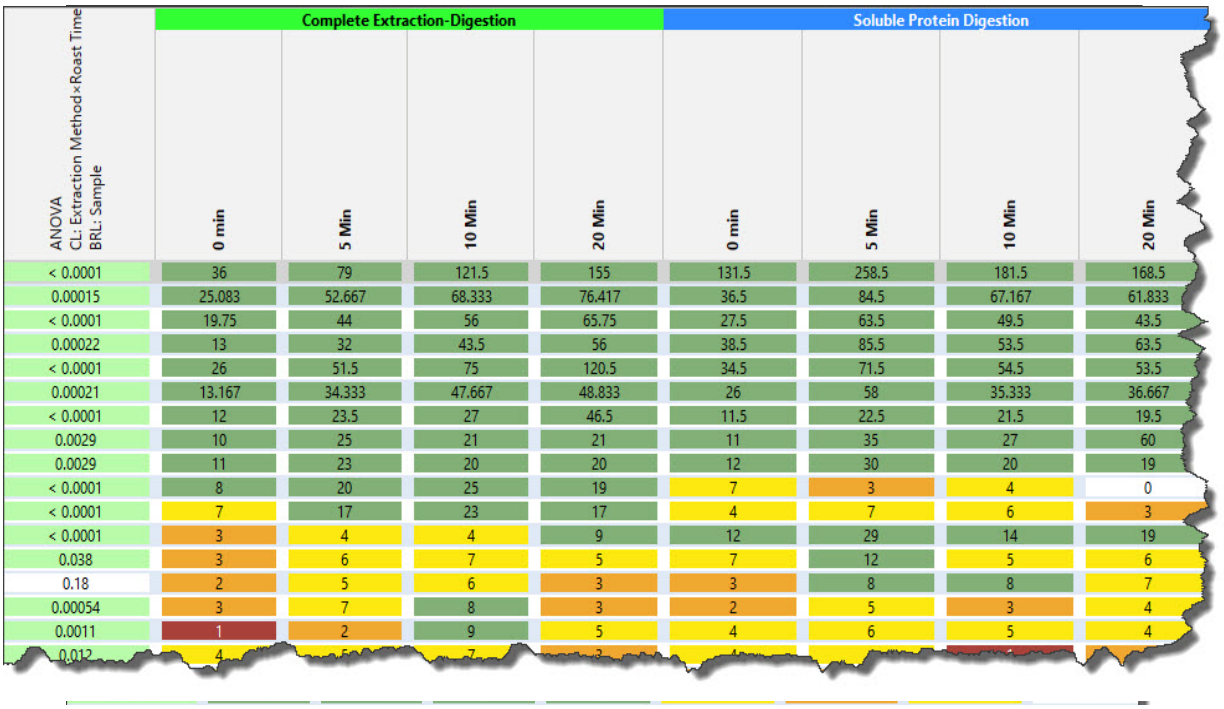
- Which Category should be considered as the Primary Analysis Category. This is generally the treatment or condition that is the main focus of the experiment.
- Which Category identifies the Secondary Analysis Category. Often this will be a category that defines a condition which is to be controlled for in the experiment.

Other Categories to be specified are similar to those described in "For the Basic Design:"

Summarization Level

The level of summarization at which the data is to be grouped can then be selected from the summarization drop-down list. The user can choose the level from the most detailed (MS sample) to any higher level shown in the summarization list. Selecting ‘Biosample’ in this example and then looking at the Samples View, causes the data to be rolled up as shown in Figure 4-13.

Figure 4-13: Data grouped by Extraction Method>Roast Time>Replicate, summarized at the Roast Time attribute level



ANOVA CL: Extraction Method×Roast Time BRL: Sample	Complete Extraction-Digestion				Soluble Protein Digestion			
	0 min	5 Min	10 Min	20 Min	0 min	5 Min	10 Min	20 Min
< 0.0001	36	79	121.5	155	131.5	258.5	181.5	168.5
0.00015	25.083	52.667	68.333	76.417	36.5	84.5	67.167	61.833
< 0.0001	19.75	44	56	65.75	27.5	63.5	49.5	43.5
0.00022	13	32	43.5	56	38.5	85.5	53.5	63.5
< 0.0001	26	51.5	75	120.5	34.5	71.5	54.5	53.5
0.00021	13.167	34.333	47.667	48.833	26	58	35.333	36.667
< 0.0001	12	23.5	27	46.5	11.5	22.5	21.5	19.5
0.0029	10	25	21	21	11	35	27	60
0.0029	11	23	20	20	12	30	20	19
< 0.0001	8	20	25	19	7	3	4	0
< 0.0001	7	17	23	17	4	7	6	3
< 0.0001	3	4	4	9	12	29	14	19
0.038	3	6	7	5	7	12	5	6
0.18	2	5	6	3	3	8	8	7
0.00054	3	7	8	3	2	5	3	4
0.0011	1	2	9	5	4	6	5	4
0.012	4							

The Summarization Bar allows the user to combine samples at various levels of categorization. The drop-down list displays the Categories in the Summarization hierarchy. The Samples View table will display samples combined at the level of the selected Attribute Group, with values rolled up to that level appropriately. The last item in the list is the command Edit..., which, when selected, opens the Edit Experimental Design dialog.

Available Categories

This column initially lists all of the Categories that have been created in the Organize View. These categories may be assigned various roles in the experiment by moving them into the boxes to the right. To assign an available category to a specific role, click on the category and either:

- drag the category into the appropriate box at the right.
- click on the right arrow button in the appropriate box at right.

To return a category to the Available Categories list, select it in the box to which it has been

assigned and either:

- drag it back to the Available Categories list.
- click on the left arrow in the Available Categories box.

Categories to be Studied

The category or categories that will define the groups to be compared in ratios or statistical tests. If no Categories are moved into this box, it will not be possible to apply a statistical test and no summarization above the level of the biological samples will occur. If one Category is selected for study, statistical comparisons between groups of samples corresponding to the different Attributes of that Category may be made, and values may be summarized to the Category level. If more than one Category is selected, statistical comparisons will operate on groups representing each possible combination of attributes for those Categories, and summarization may proceed up through the levels specified. For example, if both Condition and Sex are selected, statistical tests will compare Male Control, Female Control, Male Treated and Female Treated.

The Time Category

In a repeated measures experiment, the same subjects are measured at various time points or under different conditions. The Time Category defines which of the repeat groups a sample represents. For instance, the Time Category could be Time Point, with values of 0 hr, 1hr, 2 hr and 3hr. It need not represent time, however. For example, in a study that measures each subject's reaction to various treatments, it might be Treatment.

Primary Analysis Category

In a two-way experiment, the Primary Analysis Category should be the grouping that is the primary focus of the experiment. For example, in an experiment that compares protein levels with and without drug treatment, but wants to consider the possibility of a differential response to the drug in males and females, the Primary Analysis Category would be set to Treatment, while the Secondary Analysis Category would be set to Sex. Note that in a Two-Way ANOVA, however, the user may select whether to assess the Primary Factor effect, the Secondary Factor effect, or the Interaction effect, so the choice of Primary vs. Secondary Analysis Category is not extremely important.

Secondary Analysis Category

In a two-way experiment, the Secondary Analysis Category is a second grouping that may have an effect on the outcome and should be considered along with the Primary Analysis Category in comparisons and statistical tests. Often it is a grouping that may represent a variable that should be controlled for in testing the Primary Factor effect. Note that in a Two-Way ANOVA, however, the user may select whether to assess the Primary Factor effect, the Secondary Factor effect or the Interaction effect, so in cases where there is not a clear Primary factor, the two categories to be studied may be presented in either order.

Biological Samples

This level indicates the Category that defines a biological sample or subject. There may be more than one sample representing one biological sample if technical replicates or fractions are collected.

Technical Replicates

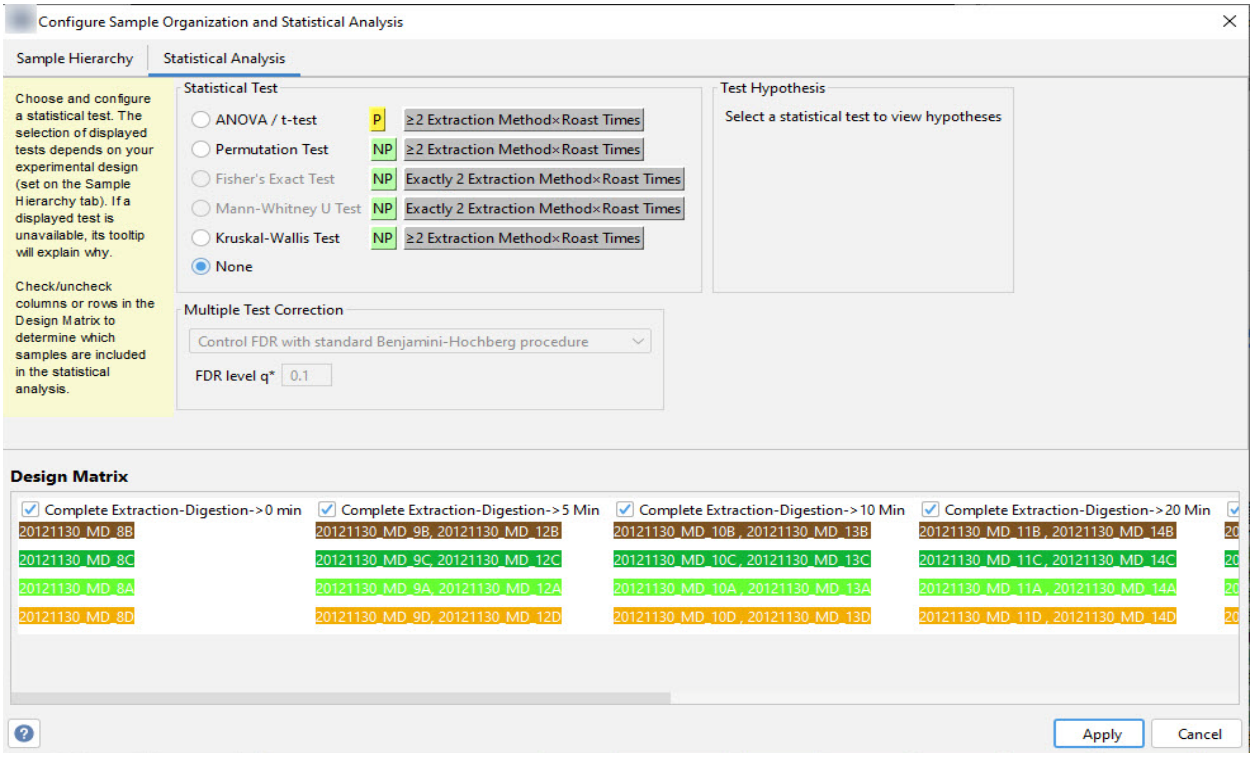
If the Technical replicates were acquired box is checked, the user must specify a category that represents the technical replicates. If each biological sample has been divided and the resulting sub-samples have been analyzed in the mass spectrometer separately, the sub-samples are technical replicates. The Category which names these sub-samples should be entered here. In some cases, the sub-samples may have undergone different treatments or have been obtained separately from the same biological subject, but in Scaffold Quant, they should be considered as technical replicates if the intent is to use them as multiple measurements of the same biological subject. Technical replicates may be the MS Samples if fractionation has not been performed.

Fractions

Fractions are generally the MS Samplesif the biological samples were separated on a 2-D gel. Specifying the samples as fractions allows quantitative values for peptides detected in different samples to be combined as they would be in a MuDPIT experiment.

Statistical Analysis Tab

Figure 4-14: The Statistical Analysis Tab



The Statistical Analysis Tab consists of several sections:

Instructions -- Helpful text displayed in the box with a yellow background.

Statistical Test - The tests appropriate for the selected experimental design are shown. If a test may not be performed, the test name is grayed and its selection is disabled. An explanation is provided.

Unchecking some boxes in the Design Matrix may make additional tests available. For example, the Mann-Whitney U Test requires exactly two categories. If the data has three, the test is unavailable, but unchecking the box in the header of one category allows the user to compare the other two categories using this test.

Multiple Test Correction - When testing the significance of a large number of protein inferences, it is advisable to apply a multiple test correction to the statistical test. The dropdown allows selection of a correction method.

The **FDR Level q*** allows the user to select a significance level.

Test Hypothesis - Displays the null hypothesis being tested and the alternative hypothesis, which would be accepted if the test result is significant. In the case of a two-way analysis, several hypotheses may be tested. A dropdown allows the user to select which of the possible hypotheses should be tested, and the test hypothesis text adjusts accordingly.

Hypothesis options for the Two-Way ANOVA

- Interaction Effect - measures whether the Primary and Secondary analysis categories are related. A significant result for the Interaction Effect for a protein means that the Primary and Secondary categories are not independent, but rather that the combination of these factors has an effect on the level of the protein.
- Primary Factor Effect - measures whether the Primary Analysis Category has a significant effect when controlling for the Secondary Analysis Category.
- Secondary Factor Effect - measures whether the Secondary Analysis Category has a significant effect when controlling for the Primary Analysis Category.

Hypothesis options for the Randomized Block Design

- Treatment Effect - measures whether the Primary Analysis Category has a significant effect when controlling for the Secondary or Blocking Category.
- Block Effect - measures whether the Secondary Analysis Category (which is the Blocking Level in a Randomized Block experiment) has a significant effect.

Chapter 5
The Samples View

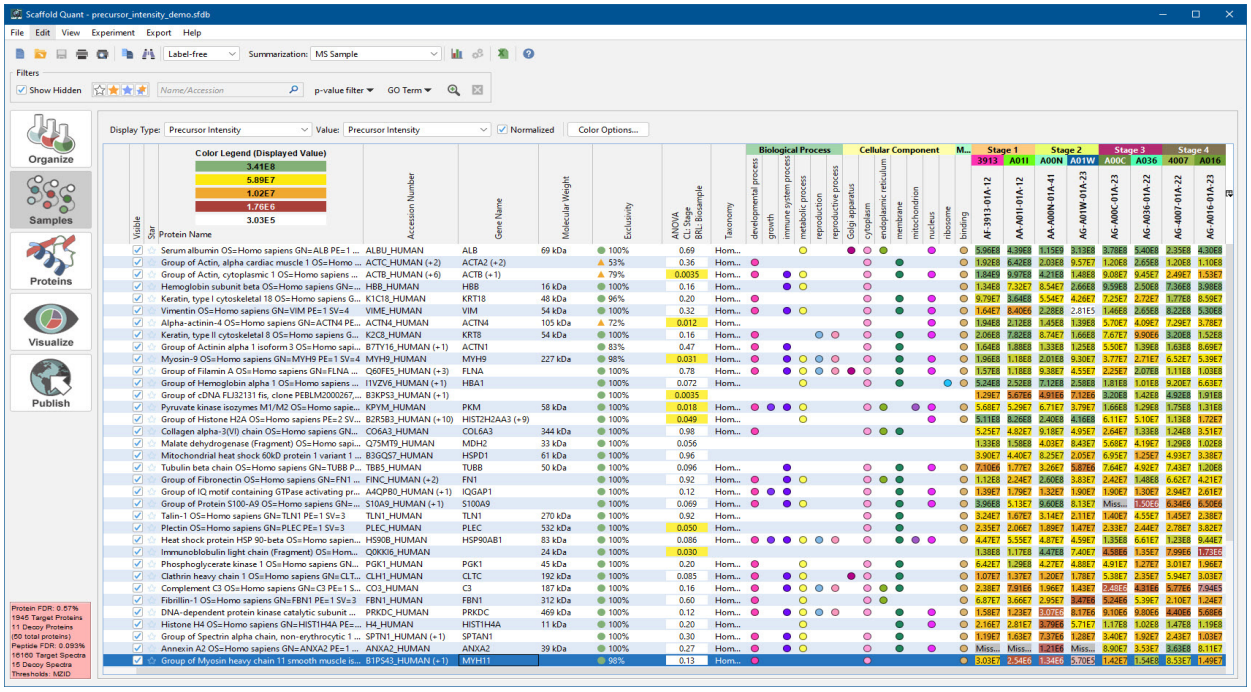
The Samples View provides a series of tools to help the user summarize and interpret experimental results at the protein level. The Samples View is the first view displayed in the Scaffold Quant main window when the application finishes loading data or when the user opens an Scaffold Quant *.SFDB file. While very similar to the Scaffold Samples View, the Scaffold Quant Samples View provides improved tools for custom summarization and visualization of large data sets. The user may select a View other than the Samples View as the default upon loading or opening through Edit>Preferences>User Interface>Views.

The Samples View organizes all of the information about the proteins identified in the experiment to provide an overview of the experimental results, as well as protein-level quantitative information.

The Quantitative Method selector, at the top of the Scaffold Quant main window allows the user to select whether quantitative analysis will be performed on MS Samples or quantitative samples associated with a multiplex quantification technique if the data supports this type of analysis.

The Summarization Bar, found in the main Scaffold Quant window, allows the user to change the level of summarization at which quantitative values should be displayed in the Samples View. The levels available depend on the sample organization established inThe Organize View and the experimental design set in the Configure Sample Organization and Statistical Analysis dialog (see Specifying the Design of an Experiment).

Figure 5-1: The Scaffold Quant Samples View


Scaffold Quant - precursor_intensity_demo.sfdb
File Edit View Experiment Export Help
Label-free
Summarization: MS Sample
Filters
Show Hidden
p-value filter
GO Term
Organize
Samples
Proteins
Visualize
Publish
Display Type: Precursor Intensity
Value: Precursor Intensity
Normalized
Color Options...
Color Legend (Displayed Value)
3.41E8
5.89E7
1.02E7
1.76E6
3.03E5
Protein Name
Biological Process
Cellular Component
Stage 1
Stage 2
Stage 3
Stage 4
Protein FDR: 0.57%
1945 Target Proteins
11 Decoy Proteins
(50 total proteins)
Peptide FDR: 0.093%
16160 Target Spectra
15 Decoy Spectra
Thresholds: MZID

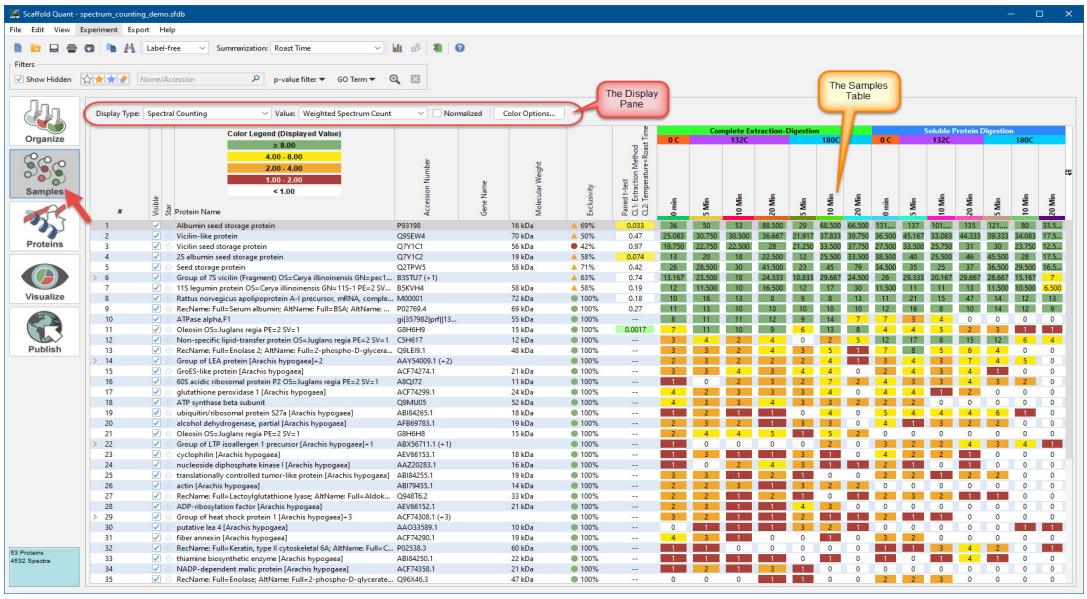
Samples View Features

The purpose of the Samples View Table is to show at a glance, how levels of each protein vary among the various Mass Spectrometry (MS) samples in the experiment . A variety of Display Type options are available, including Total Spectrum Count, Exclusive Unique Peptide Count, Precursor Intensity, TIC, etc.

Labeled

When labeled, multiplexed samples have been loaded into Scaffold Quant running with a Labeled Quant license, the Samples View Table can also display levels of proteins varying among quantitative samples. In this case, the Display Type options include reporter ion intensities when an isobaric labeling technique is chosen as the Display Method.

Figure 5-2: The Scaffold Quant Samples View Table Tab



The main elements of the Table tab are:

- Samples View Features
- Samples Table Display Bar

Samples Table

The Samples View presents an overview of the experiment. It provides the user with a list of identifiedproteins. It also provides scoring information and other characteristics to facilitate assessment of the reliability of the identifications, and quantitative information indicating the level of expression of each protein in each sample.

In addition, the Samples View provides tools such as filters, flexible summarization and statistical analysis that are designed to help answer the fundamental questions underlying the experiment. Through the Samples View, a researcher can identify the proteins that distinguish various classes of samples, such as treatment groups, tissue types, time since treatment, etc. Other Scaffold Quant Views then provide additional resources for further examination of the preliminary results presented in the Samples View.

The level of data summarization is chosen from the Summarization Bar in the Scaffold Quant main window. The Display Type pull-down menu determines the type of quantitative values reported in the table and the Value pull down menu sets the specific display option. For instance, the Display Type might be “Precursor Intensity” and the Value might be “Exclusive Log10 Precursor Intensity”.

The protein list can be filtered using the tools of the Filtering control bar.

General characteristics of the Samples Table:

- Samples Table Features
- Initial thresholding
- Initial Sorting of Columns
- Summarization Level in the Samples Table
- Color Legend

Samples Table Features

Like any table in Scaffold Quant, the Samples Table makes use of the features and tools described in the Display pane section.

The following columns, initially ordered as follows, appear in the Samples Table:

- **# --** Order number of each row at the current ordering conditions.
- **Visible --** Shows a list of selected check boxes. Deselecting a box hides the corresponding row unless the “Show Hidden” check box in the Filters pane is selected.
- **Star** -- Initially shows a grayed out star for every row. Clicking the star activates it, tagging the protein group. Clicking repeatedly causes the star to loop through its four possible states. The color of the star goes from gray to orange, to blue and to orange and blue then back to gray. For more information see “When a row includes a cluster of proteins, the row

number is preceded by a small clickable expansion icon and a pie-icon, see Figure .” on page 88.

- **Protein Name** -- Name of the protein group.
- **Accession Number** -- The protein identification number.
- **Gene Name** -- The gene associated with the protein as specified in the protein’s description in the Uniprot Database. This column is populated by selecting the Populate Gene Names option in the **Experiment Menu**. This operation parses the gene name from the Protein Name field if the Protein Name is in Uniprot format and contains the string “GN=...”. If the Protein Name does not contain the gene name in this format, no gene names will be extracted and the column will remain empty.
- **Molecular Weight** -- Provides the theoretical molecular weight.
- **Exclusivity** - the percentage of the peptides identifying this protein that are not associated with any other protein in the experiment.
- If **GO terms** and/or Pathways have been applied, a Taxonomy column and a series of columns indicating whether or not the protein is associated with each GO term or Pathway.

The rest of the columns represent, depending on the selected level of summarization, either the MS samples or the Categories defined in the and included in the hierarchical summarization.

The order of the columns can be changed by the user, see “Display pane” on page 71.

Initial thresholding

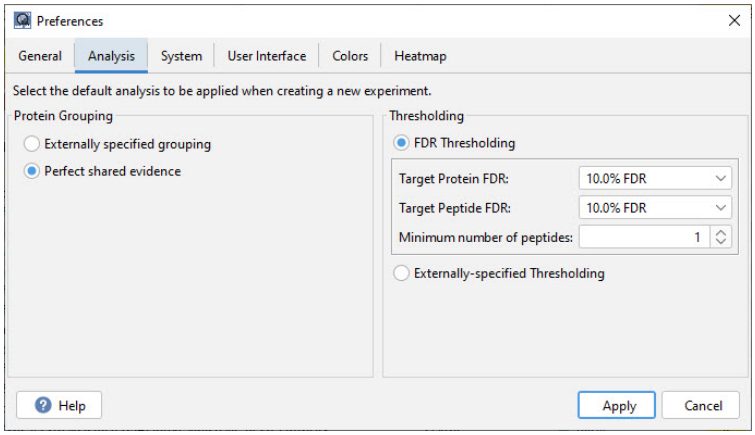
Each time data is loaded into Scaffold Quant, protein grouping and thresholds are applied. Grouping and thresholding options may be selected through the Preferences, Analysis tab. The initial default settings are the following:

- **Protein grouping:** Perfect shared evidence
- **Thresholding:** FDR thresholds are applied as follows:

  Target protein and peptide FDR 10%, minimum number of peptides 1.

The user may modify these settings through the Analysis tab in the Preference dialog which can be reached from the **Edit > Preferences...** menu item. When different preferences are applied, they become the user’s new default settings and are retained even when the program is closed and restarted.

Figure 5-3: Initial default setting for thresholding of loaded data



Initial Sorting of Columns

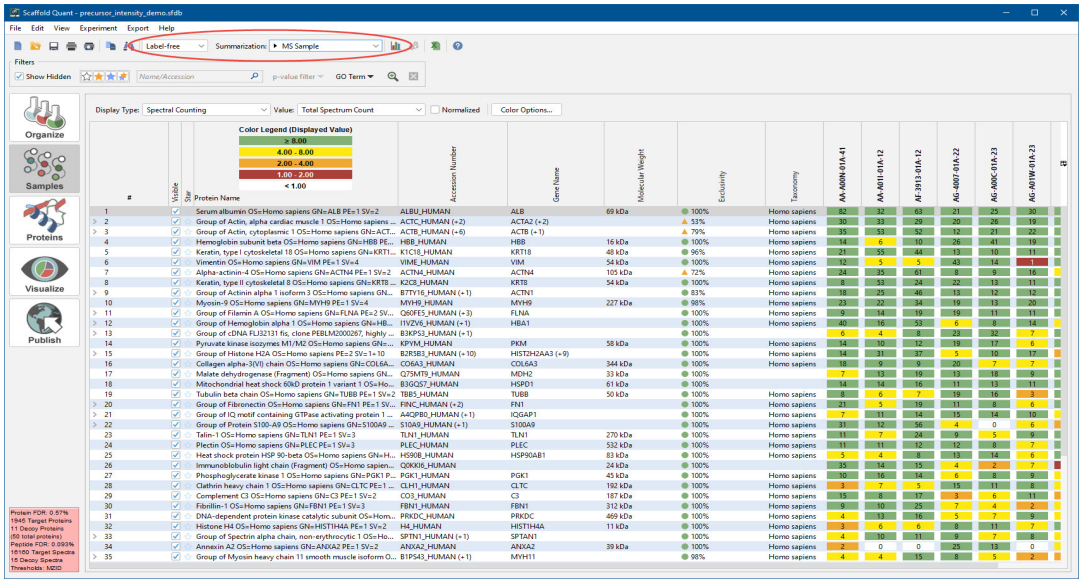
When the Samples View first opens, all of the protein groups in the protein list are sorted by the sum of the Normalized Spectrum Counts at the lowest summarization level, which is the MS Sample level, in descending order. This sum gives an estimate of the total amount of evidence present in the experiment for a protein, so ordering the list of protein groups by this value provides a list that shows the “best” proteins at the top of the list.

Each column is provided with a tri-state sorting feature. Clicking on any column header will reorder the table according to the values in that specific column. The first click sorts the column in ascending order, the second in descending order and the third returns the column to the original order.

Summarization Level in the Samples Table

After a new experiment is completely loaded, the Scaffold Quant Samples View appears with the Samples Table initially summarized at the lowest level of summarization, which is the MS Sample levelFigure 5-4. If the experiment is a labeled quantitative experiment, the user may use the Display Method drop-down to switch to viewing Quantitative Samples rather than MS Samples.

Figure 5-4: Samples View with summarization at the MS Sample level

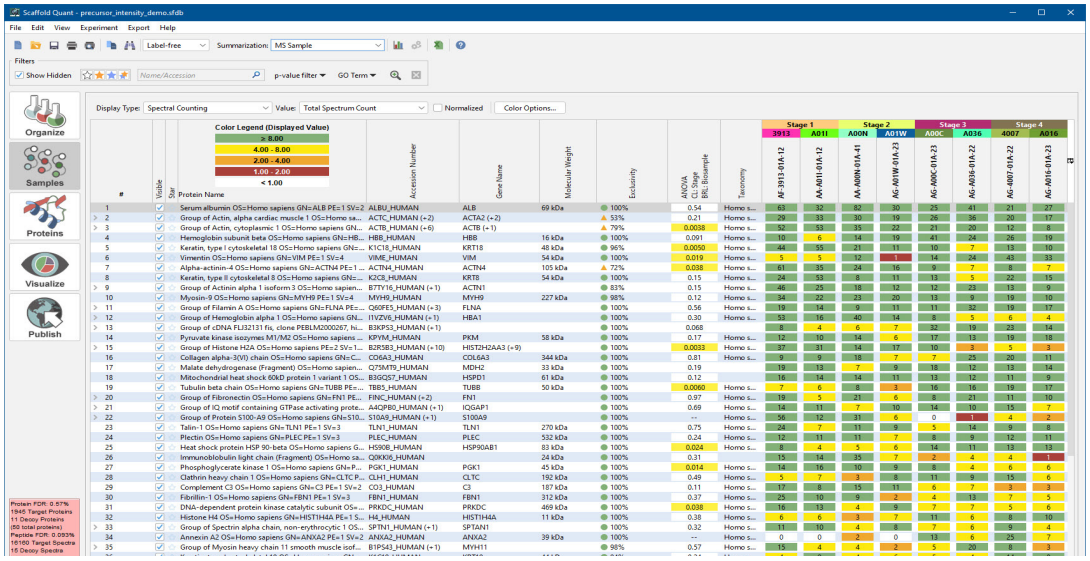


The user can add Categories and classify the MS (or Quantitative) samples within them through the Organize View. It is then possible to set up the desired Summarization Hierarchy using any of the defined Attribute Groups by choosing **Edit...** in the Summarization Bar.

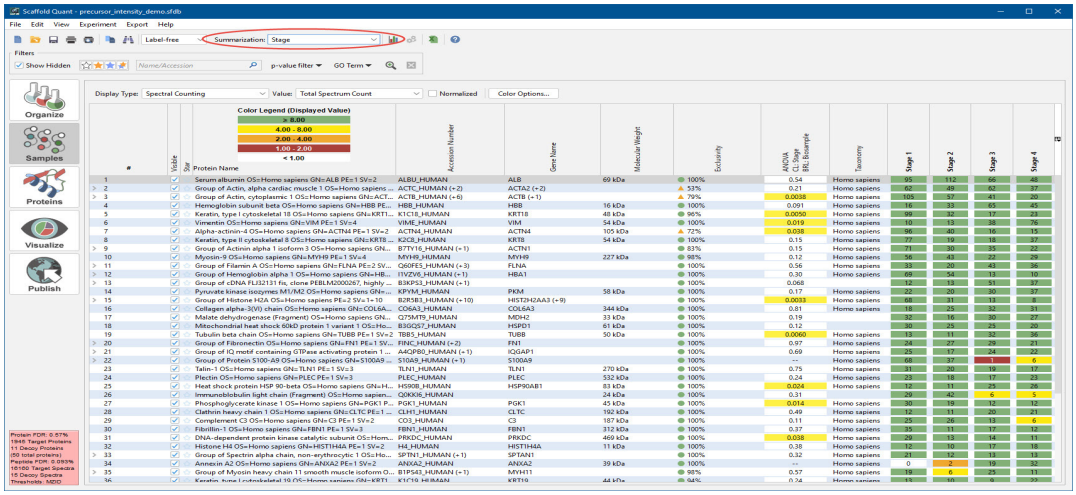
The selected Summarization Hierarchy and the choice of the summarization level will be reflected in the Samples Table column headers.In Figure the selected level is still MS Sample but a more complex hierarchy has been set up in the Summarization Pane.

The headers of the samples columns show the Attributes for all Categories in the Summarization Hierarchy. They are colored according to the values assigned to the Attributes through the Organization View.

Scaffold Quant



Selecting a higher level of summarization will hide the lower ones.



Rolling up of quantitative values to higher summarization levels

The quantitative values shown in the Samples table depend on the selected Display Type and the chosen summarization level.Quantitative values are combined (or rolled up) to higher levels of summarization differently depending on the Display Type selected and the experimental design, see Rolling up of quantitative values.

Color Legend

Located at the top of the Samples Table in the proteins column header, the color legend defines the color coding associated with the selected Display Type. The color legend can be customized through the Edit Coloring for Display Type dialog opened by selecting Color

Options... from the View Menu.

Tools for Limiting the Protein List

Protein lists can at times be very long. Scaffold Quant includes a series of tools that allow the user to simplify the list and filter it to the proteins of interest. Protein grouping and clustering proteins that share similarities reduces the number of independent rows in the list. Thresholding by protein and peptide FDR allows the user to disregard less than optimal candidates, and several filters allow the user to select only the proteins that are likely to be of biological significance in the experiment.

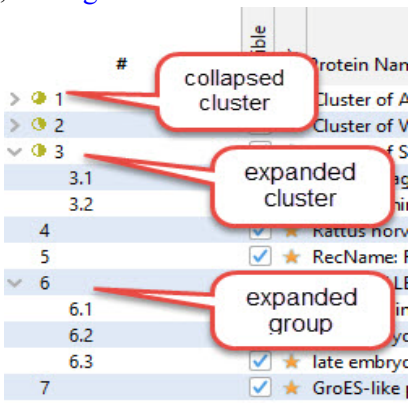
The following filtering and thresholding tools are available:

- Hidden Proteins
- Applying Confidence Thresholds to the Protein List

Applying filters to the Protein ListExclusivity RatioProtein groups and clusters appearing in the initial Proteins List

Scaffold Quant loads all of the proteins in the original input files (e.g.*.MZID files) in an experiment. For a protein or protein group to be included in the Proteins list it must have a valid Exclusive peptide in at least one of the MS Samples. The proteins are grouped across samples, clustering is applied, and the thresholds specified in the Analysis Preferences dialog are applied.

When a row includes a cluster of proteins, the row number is preceded by a small clickable expansion icon and a pie-icon, see Figure .



Tagging Protein of Interest, the star function

The user can mark proteins in an experiment that are of special interest by clicking the Star icon ☆ in the **Star** column for the protein. Three different colored stars, blue, orange and a

combination of the two colors may be applied by clicking multiple times on the same star or by selecting the star option in the right click menu. By using a combination of different stars it is possible to create four different sets of proteins of interest. The user can then bring these items to the top of the display by clicking twice on the **Star** column header. To return to the default display order, click the column header twice more.

Sets of proteins can be starred at the same time by selecting multiple proteins and using the star option in the right click menu.

Star filters are included in the "Filtering control bar" of the Scaffold Quant Main Window. Clicking a specific colored star in the filters bar removes all proteins tagged with that star color from the table.

Hidden Proteins

The user can easily remove proteins displayed in the Samples Table that are of no interest and/or are contaminants. An entire row in the Samples Table can be eliminated by simply clearing the Visible option in that row. This can be done for a single protein by clicking the Visible check box, or for a group by using the right click menu.For example, to eliminate Trypsin products from the table, the user can carry out a search for all proteins that contain "Trypsin" in their names, hover over one of the selected proteins, right click and choose **Show/Hide > Hide** to clear the Visible option for all of the proteins that meet this search criterion. After this, only those proteins that do not have "Trypsin" in their names will be displayed. Alternatively, the user could display only the Trypsin products, by selecting all of the Trypsin proteins, right clicking and selecting **Show/Hide> Hide Others.** This will clear the Visible option for all except the selected proteins.

The check box Show Hidden in the Filtering pane in the Scaffold Quant Main window toggles the display of hidden proteins

Applying Confidence Thresholds to the Protein List

When data is initially loaded into Scaffold Quant, the list of proteins in the Samples Table reflects the thresholds set through the Preferences, Analysis tab.

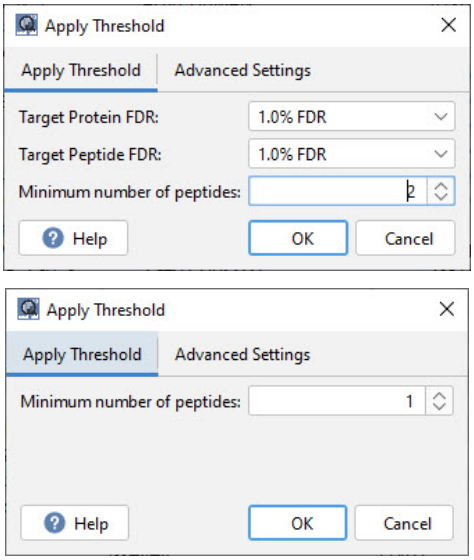
Within Scaffold Quant, if the loaded data were searched using decoys, it is possible to apply thresholds to set minimum characteristics for identification confidence based on the Protein and Peptide False Discovery Rates (FDR) and the minimum number of peptides in a protein through the Apply Threshold dialog box.

- Confidence thresholds can be applied only to data that was searched using decoys.
- FDR filtering should be applied only when all samples use the same scoring system. To combine results from different search engines, it is recommended that the files first be loaded into Scaffold and the data exported as mzIdentML files. The Scaffold probabilities can then be used as a common scoring system for all samples.

Apply Threshold dialog box

By selecting the menu option **Experiment > Apply Thresholding > Apply Threshold...** the user can access the dialog **Apply Threshold.** If the data has been searched using decoys, the dialog will show three possible thresholds to be adjusted; if no decoys were included in the search only one thresholding option is provided.

Figure 5-5: Apply Threshold dialog



When the data has been searched with decoys, the dialog contains the following options:

- **Target Protein FDR**, a scroll down list of FDR values to choose from and a “No bound” option.
- **Target Peptide FDR**, a scroll down list of FDR values to choose from and a “No bound” option.
- **Minimum number of peptides**, a text box where the minimum number of peptides included in a protein can be defined.

The thresholds specified in the Apply Thresholds dialog represent upper bounds values for the FDR Scaffold Quant computation. The thresholds are applied as described in the Appendix , “Computation of protein and peptide FDR in Scaffold LFQ”. The actual computed values after thresholding are shown in the FDR Information Box located in the lower left corner of the Scaffold Quant’ main window. The number of target and decoy proteins and the number of target and decoy spectra are also reported in the box, along with the total number of proteins that pass thresholds. The background of the FDR info box is colored in pink to highlight the fact that the data set loaded in the experiment was searched with decoys.

Selecting the option **Experiment > Apply Thresholding > Restore MZID Threshold** reinstates the original thresholds reported in the loaded MZID files.

A more detailed description of the method by which Scaffold Quant computes the number of target and decoy proteins and peptides for a set of FDR thresholds is given in the appendix Computation of protein and peptide FDR in Scaffold LFQ

When no decoys are present in the loaded data, the Apply Threshold dialog provides only the opportunity to threshold the data according to the minimum number of peptides identifying a protein. In this case, the FDR Information Box is colored blue and shows only the numbers of identified peptides and spectra and the threshold setting.

The peptide count that is considered for filtering the min # of peptides is the number of peptides over the whole experiment and it can be seen in the Samples table by selecting the highest level of summarization.

The Advanced Settings Tab

Selecting **Experiment>Apply Thresholding>Apply Threshold...** also provides access to the **Advanced Setting Tab**. This tab contains the **Choose Primary Score** button which opens a dialog that allows the user to select specific primary scores for the proteins and for the peptides. This allows a user to choose a score common to all samples if the user has loaded samples processed with different software, but which share a common score. For example, if one sample had been processed with Mascot, then loaded into Scaffold, it would contain both Mascot scores and Scaffold probabilities. If another sample were loaded which had been searched with Mascot but loaded directly, it would contain only Mascot scores. Scaffold Quant does not have a mechanism for computing probabilities based on a combination of different scores, so it will choose primary scores based on the first sample loaded.

When FDR Thresholding is performed, Scaffold Quant must be able to rank the proteins based on their scores. If subsequent samples do not have the type of scores which were designated as primary for the first sample, the program will be unable to rank the peptides and proteins in those samples, and will ignore them. A warning message is displayed to the user in this case, and the user may choose a different set of scores to be designated as the primary scores so that all samples may be included in FDR thresholding. Note that great care must be exercised when choosing primary scores, as many scores provided by search engines are not actually measures of the quality of the peptide-spectrum match. Choosing a score which is a poor measure of PSM quality will produce very bad results.

If there is no common set of scores among the samples, FDR thresholding is not recommended.

Applying filters to the Protein List

A number of filter options are offered in the Filtering control bar located in the Scaffold Quant Main Window. The filters affect the number of visible rows in the Samples Table.

- The Show Hidden check box toggles whether rows that have been tagged as not visible can be seen in the list.
- The Star Filter filters out all the rows that are not tagged as selected in the filter.

- The P-Value Filter filters out all proteins for which the statistical test result is not significant based on the selected criteria.
- Advanced Filter offers a variety of filter criteria. The Samples Table shows only proteins that meet the criteria specified in the filter. In the case of peptide filters, the Samples Table shows only proteins which contain peptides that meet the filter criteria.

Exclusivity Ratio

The exclusivity ratio, ER, provides a way to judge how a protein can be used to extract reliable quantitative information. The purpose is to indicate how many of the peptides included in a protein are shared with other proteins present in the experiment.

The Exclusivity Ratio is expressed as a percentage and given by:

$$ER = \frac{\text{Exclusive Peptide Count}}{\text{Total Peptide Count}} \times 100$$

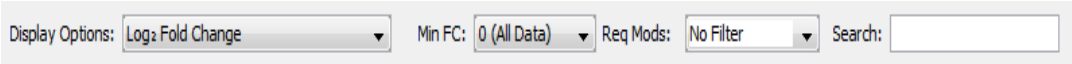
The larger the ER value is the more reliable the protein will be for quantitation.

In the samples table under the Exclusivity column, the ER value for each protein is reported along with a colored dot that visually displays the degree of exclusivity.

The dots are green when ER is between 80%-100%, yellow when ER is between 50%-80% and red when ER is between 0%-50%.

Samples Table Display Bar

Through the Samples Table Display Bar, the user can specify the type of values (for example, the Number of Assigned Spectra) that are displayed in the Samples Table for every protein group. The bar also contains filtering options for limiting the display to only those proteins that meet specific criteria.



The Samples Table Display bar contains the following features:

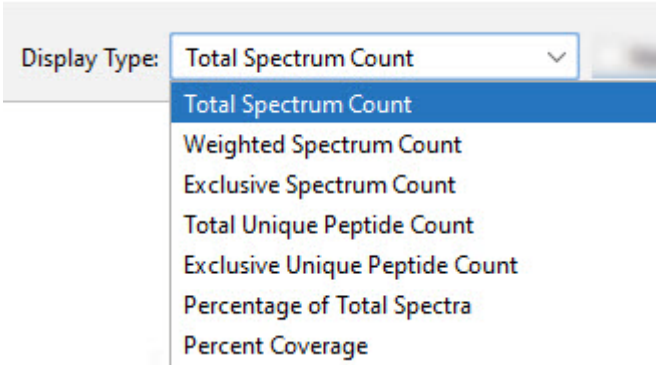
- Display Type
- Normalized check box and/or Ref: pull down list
- Color Options button
- Search Box

Display Type

The Display Type drop-down selects the type of quantitative values that will be available for display in the Samples Table. The possible display types depend on which Quantitative Method has been selected. In Label-free mode, options include Spectral Counting, Non-Quantitative values (like % Coverage), Precursor Intensity and TIC. The specific manner in which, e.g., the spectral counts or precursor intensities should be displayed is selected in the Value drop-down.

For multiplex quantitative modes, the types include various representations of reporter intensities, fold change of reporter intensities, and Coefficient of Variation. In quantitative modes, no Value drop-down is shown and the specific display format is selected directly through the Display Type drop-down.

Figure 5-6: List of Display Type Value options available in spectral counting mode



- **Total Spectrum Count ---**The total number of spectra associated to a single protein group, including those shared with other proteins.
- **Weighted Spectrum Count --** Number of spectra associated with only a specific protein group plus the apportioned number of spectra shared with other proteins, see Weighted spectrum count.
- **Exclusive Spectrum Count** -- The number of spectra associated only with a specific protein group.
- **Total Unique Peptide Count** -- The number of different amino acid sequences that are associated with a specific protein including those shared with other proteins
- **Exclusive Unique Peptide Count** --The number of different amino acid sequences, regardless of any modification that are associated with a single protein group.

Note: When this option is selected and a protein group is expanded in the Samples table, each protein in the group will have zero exclusive unique peptides associated with it.

- **Percentage of Total Spectra**-- The number of spectra matched to a protein, summed over all MS Samples, as a percentage of the total number of spectra in the sample.
- **Percent Coverage --**The percentage of all the amino acids in the protein sequence that were detected in the sample.
- **Log_2 Fold Change (weighted spectrum count)** -- The fold change is calculated by dividing each weighted spectrum count in a row by the values appearing under the column selected through the Ref: pull down list, and the base 2 log is calculated.
- **CV (Weighted Spectrum Count) --** The coefficient of variation among the replicate values which are summarized to provide the weighted spectrum count for a protein within a specific column. Note that this value will be 0 if samples are displayed at the individual MS Sample level.

The following Display Type Values are available only when precursor intensity data is loaded in Scaffold Quant, for more information see Loading precursor intensity data.

- **Precursor intensity** -- It shows the precursor intensity if this information is included in the loaded data files.**Log_2 Fold Change (precursor intensity)** -- This value is calculated by taking the base 2 log of the ratio of the rolled up Precursor Intensity value for a cell to the corresponding value for the column selected through the Ref: pull down list.
- **CV (Precursor Intensity) --** The coefficient of variation among the replicate values which are summarized to provide the precursor intensity value for a protein within a specific column. Note that this value will be 0 if samples are displayed at the individual MS Sample level.

.

The coloring in the Samples Table reflects the selected Display Type. For each type the coloring can be customized by selecting Color Options from the View Menu.

TIC

One of the options in the Display Type drop-down list in Label-free mode is TIC. TIC values displayed in Scaffold Quant are actually approximations of the Total Ion Current computed by summing the intensities of all fragment peaks in each MS2 spectrum.

.

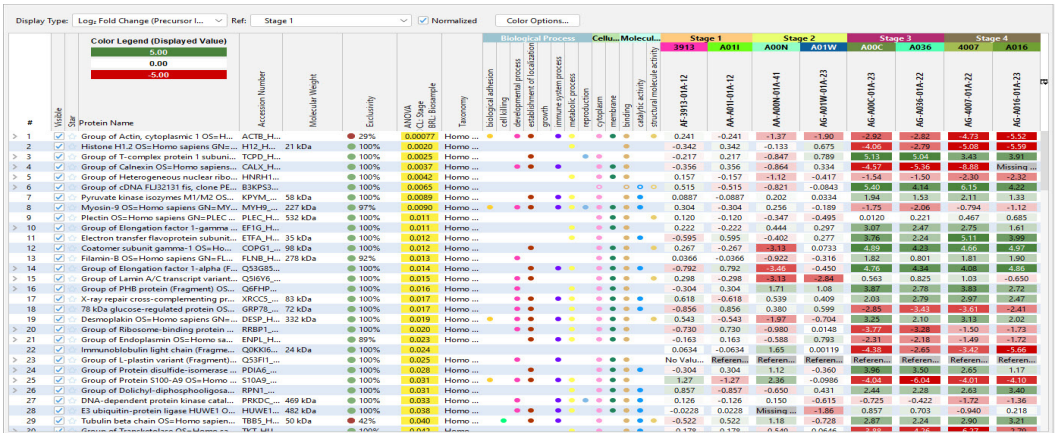
In older SFDB files, created before the implementation of TIC in Scaffold Quant, it may be necessary to trigger the population of TIC values. To do this, choose Add/Remove Quant Type... from the Quantitative Method drop-down at the top of the main window, then click ***Enable TIC****.*

Log_2 Fold Change and Precursor Intensity missing value tags

When selecting a Display type that represents the log_2 Fold change of a quantity, or a rolled-up log_{10} intensity value, three different missing values tags might appear in the samples table's cells:

- **Missing Values** -- the log of a quantity that is zero, which ultimately refers to a protein that has not been detected in a particular MS sample or group of samples belonging to the selected level of summarization.
- **No Values** -- the log of the ratio between two missing values.
- **Missing Ref.** -- the log of the ratio between a value and a missing reference value.

Figure 5-7: Samples table log_2 fold change and log_{10} Intensity tags.



Rolling up of quantitative values

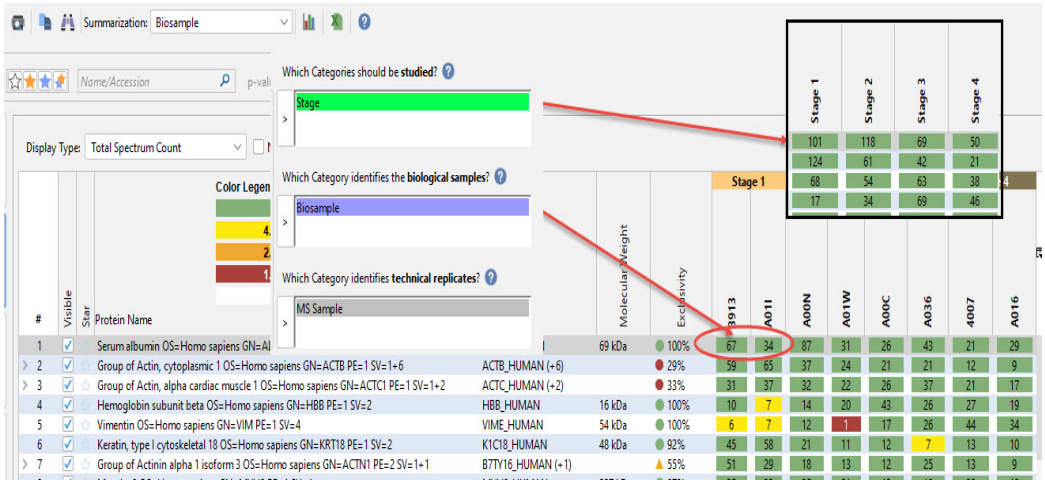
The quantitative values shown in the Samples table depend on the Display Type selected and the level of summarization chosen from the Summarization Bar. At the lowest level of summarization, the values shown relate to the amount of protein present in each of the loaded MS samples in the Scaffold Quant experiment. Each MS sample corresponds to a column in the Samples table. Changing the level of summarization groups the MS samples according to the categorization created through the The Organize View and hierarchically ordered using the Experimental Design. When transitioning from one level to the next, the columns in the lowest level are subsumed into a new column representing quantification at a higher level of summarization. The methods by which values are rolled up from one level of summarization to the next depend on the selected Display Type.

Rolling up of Spectrum Count display types

For spectrum count values, such as **Total Spectrum Count**, **Weighted spectrum Count**, **Exclusive Spectrum Count** and **Percentage of Total Spectra**, are rolled up by simply summing the counts shown in the columns included in the corresponding higher level Attribute group. For example Figure 5-8the picture below shows the Samples table at the Biosample level of summarization, where each Biosample is represented by a column. The

level above it, as shown in insert A depicting the experimental design, is the Stage level. Insert B shows a section of the columns that appear when level Stage is selected. The value shown for Stage 1 is equal to the sum of the values for biosample 3913 and A01I appearing at the Biosample level.

Figure 5-8: Rolling up quantitative values: Spectrum counts

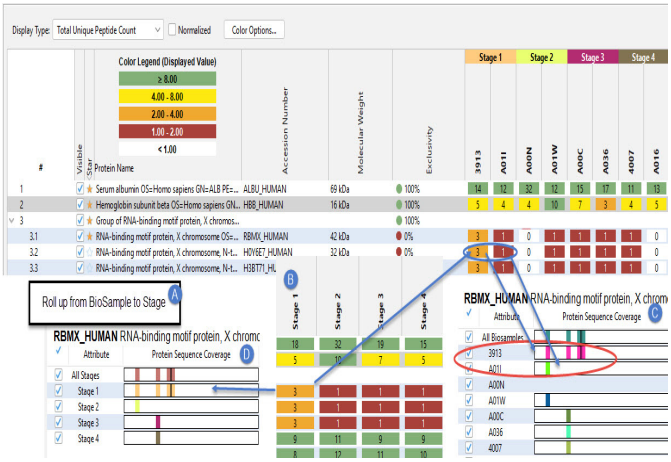


Rolling up of Peptide counts display type

Peptide count values appearing in the Samples table like **Total Unique peptide count, Exclusive Unique peptide count** and **Percent Coverage** are rolled up by taking the union of the set of peptides included in the lower level of summarization groups. Some peptides might be in more than one group so that when the group are rolled up to the higher level of summarization the peptides in common will be counted only once.

For example Figure 5-9t shows the Samples table at the Biosample level of summarization, where each Biosample is represented by a column. The level above it, as shown in insert A depicting the experimental design, is the Stage level. Insert B shows a section of the columns that appear when level Stage is selected. The value shown for Stage 1 is given by the union of the peptides included in Biosample 3913 and A01I appearing at the Biosample level. Inserts C and D show the Sequence coverage comparisons for both levels of summarization. Insert C clearly shows that Biosample A01I contains only one peptide which is also included in Biosample 3913 together with other two peptides. Insert D shows the result of the union of the two group which then corresponds to a peptide count of 3.

Figure 5-9: Rolling up quantitative values: Peptide counts

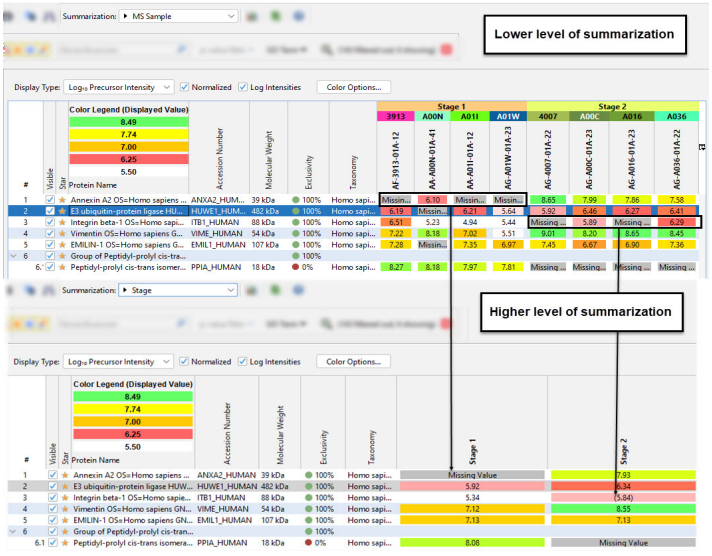


Rolling up of intensity values

Intensity values appearing in the Samples table are rolled to the upper level of summarization by taking the median of the values in the corresponding lower level group. If more than 50% of the values are missing in the group that is to be rolled up, the QRILC algorithm is used to impute missing values (see).

At the lowest level of summarization, missing values are highlighted using the Missing Value tag. At a higher level of summarization, when QRILC is used, the values are reported in parentheses. They are tagged as Missing Value when the lower level group of samples does not include any value at all, as shown in the example in Figure 5-10

Figure 5-10: Rolling up quantitative values: Log_{10} Precursor Intensity



Note:

The **Log_2 Fold change (... Intensity)** is calculated using the Log_{10} Intensity values with the Log_{10} value of the selected reference subtracted.The result of the subtraction is then transformed into the $Log_2$of the difference.

Normalized check box

When the Normalized check box is checked, the values shown in the Samples Table for the selected Display Type will appear normalized. When the chosen Display Type is the Log2 Fold Change (... intensity) and Normalization is selected, the display type Log10 Intensity will also appear normalized and vice-versa.

When the chosen Display Type is the Log2 Fold Change (weighted spectrum count) and normalization is selected, the display type Weighted Spectrum Count will also appear normalized and vice-versa.

The normalization algorithms applied are described in section.

Ref: pull down list

When Display Types containing a fold change are selected, the **Ref:** pull down list appears between the **Display Type** control and the **Normalized** check box. From this list, the user may select the Attribute or combination of factors to be used as the reference or denominator for the fold change calculation. The pull down list includes all of the Attribute Groups available in the summarization list, plus a list of all possible combinations of factor levels available at the selected summarization level.

Color Options button

Selecting the Color options menu item in the View menu opens the Edit Coloring for Display Type dialog. This dialog offers coloring adjustment tools for each Display Type.

Figure 5-11: Edit Coloring for Display Type dialog

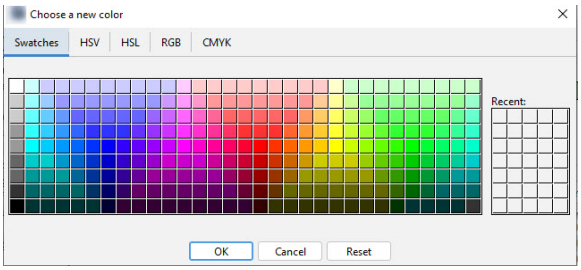
The Adjust Colors pane, included in the dialog, allows the user to create a custom color legend for the currently selected Display Type. Various features are available to set the intervals for a specific color as well as the definition of the overall range.

Sliding one of the colored squares located above the legend changes the range of the selected color. The color ranges can also be set by typing a value in the selected value box.

It is also possible to use a color gradient by clicking the color gradient check box.

Double clicking a color on the legend opens the **Choose a new color dialog** where the user can pick a different color to be added to the legend using either swatches, HBS or RGB methods. Double clicking a specific colored square also opens this dialog and allows the user to change the color of the selected square.

Figure 5-12: Choose a new color dialog



Chapter 6
The Proteins View

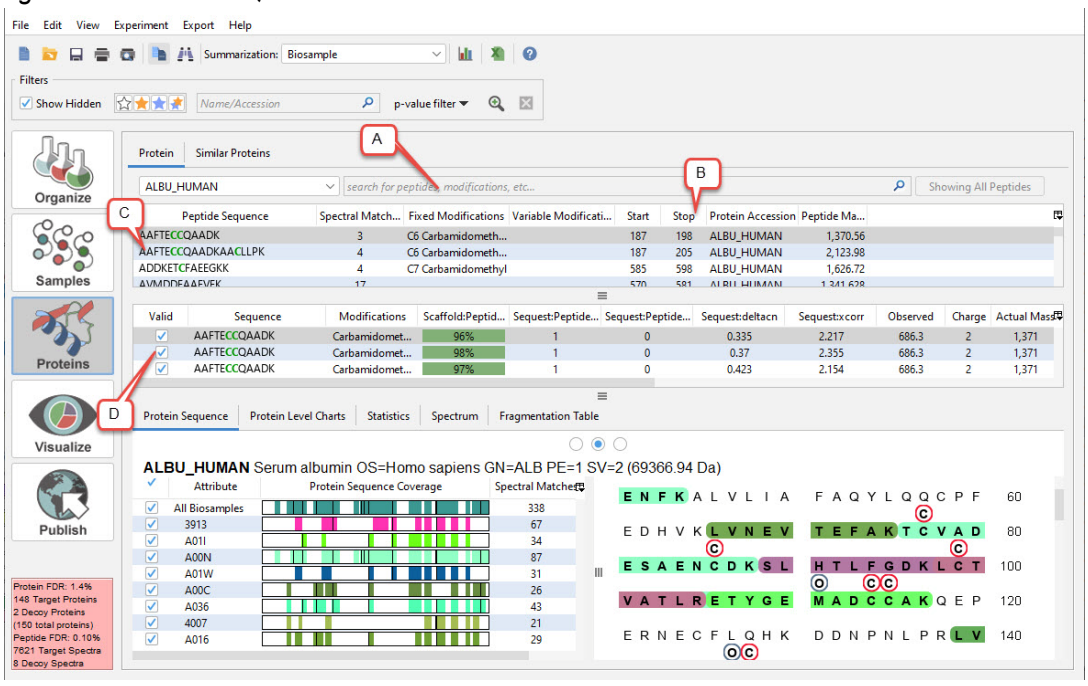
The Scaffold Quant Proteins View provides the opportunity to examine individual proteins or protein groups in greater detail. It allows examination of the peptides and spectra which comprise the evidence for the protein, and offers a visual representation of the protein sequence coverage.

Inspection and Validation of peptides: the Proteins View

The Proteins View builds on the validation and visualization features of the other Scaffold Suite programs, but has been enhanced to display information summarized by Categories as well. Figure 6-1 shows an example of the Proteins View. The View consists of four major panes:

- The Peptides Filtering Pane (A)
- The Peptides Pane (B)
- The Validation Pane (C)
- The Visualization Pane (D)

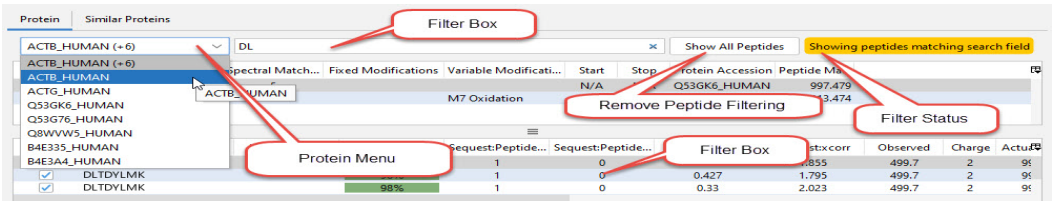
Figure 6-1: Scaffold Quant Proteins View



The Peptides Filtering Pane

At the top of the Proteins View, this line contains: a drop down Protein menu presenting the list of proteins appearing in the Samples View proteins table, a Filter box to specifically search for peptides and modifications within the Peptides View, and an indication of filtering status.

Figure 6-2: Peptides View filtering pane



Protein menu

Scrolling through the protein list allows examination of various proteins without toggling back to the Samples View. If a protein cluster is selected, the Peptides Pane includes all of the peptides from all proteins in the cluster, and the Protein Sequence Tab is empty, as there is no single sequence common to all proteins in the cluster.

Filter box

The filter box allows filtering of the displayed peptides. It accepts amino acid sequences, modification names or, if a cluster is currently displayed, protein accession numbers. When a string is entered into the search box, the Peptides Pane is filtered to display only those peptides that match the criteria, and in the case of amino acid sequences, the corresponding locations are highlighted in the Protein Sequence tab at the bottom of the window. The other Views are unaffected by this filtering.

Filtering Status

At the upper right of the Peptides View filtering panel is a button that, by default, shows the text “Showing All Peptides”. There are two methods of filtering peptides: through the Filter Box, and by selecting a specific amino acid in the Protein Sequence at the bottom of the Proteins View. When these filters are applied, the filtering status text changes to indicate the filtering conditions and the button text changes to “Show All Peptides.” Clicking on this button clears the filter(s).

The Peptides Pane

The Peptides Pane lists all of the identified peptides associated with the selected protein or protein cluster that meet the current threshold settings.

Displayed information

For each peptide, this pane displays the peptide sequence, number of spectral matches, fixed and variable modifications, start and stop positions in the protein sequence, the Protein Accession and the theoretical peptide mass. Any of these columns may be hidden through the table's Column control. The pane can be expanded vertically by pulling down the handle below the table.

Double clicking a peptide

Double-clicking on a peptide in the Peptides Pane opens the Samples View and filters the entire experiment so that only those proteins which contain that peptide are shown. The peptide sequence appears in the Identified Peptide filter box in the Advanced Filters dialog and the filter can be cleared using that control.

Figure 6-3: Peptides Pane in the Proteins View

Peptide Sequence	Spectral Match...	Fixed Modifications	Variable Modificati...	Start	Stop	Protein Accession	Peptide Ma...
AGFAGDDAPR	2			N/A	N/A	B4E335_HUMAN	975.441
DLTDYLMK	5			N/A	N/A	Q53G76_HUMAN	997.479
DLTDYLMK	6		M7 Oxidation	N/A	N/A	Q53GK6_HUMAN	1,013.474
DSYVGDEAQSK	12			N/A	N/A	Q53GK6_HUMAN	1,197.515
DSYVGDEAQSKR	40			N/A	N/A	Q53G76_HUMAN	1,353.616
EITALAPSTMK	19			N/A	N/A	Q53G76_HUMAN	1,160.611
EITALAPSTMK	18		M10 Oxidation	N/A	N/A	Q53G76_HUMAN	1,176.606
EKLCYVALDFEQEMATAASSSSLEK	1	C4 Carbamidomethyl		N/A	N/A	B4E335_HUMAN	2,806.304

The Validation Pane

The Validation Pane provides detailed scoring information and attributes for each spectrum that matches the peptide selected in the The Peptides Filtering Pane above.

- Displayed information
- Manual Validation

Displayed information

Columns available for display include the peptide sequence; modifications; search scores; the observed mass (m/z) and the actual mass which is calculated from the observed mass and the charge; the charge; the calculated (theoretical) mass; the delta mass in Daltons and in ppm; the number of enzymatic termini (abbreviated NTT); the MS/MS Sample name; the Spectrum ID and the Attributes that have been applied to this spectrum. The columns to be displayed may be selected using the Column control. In multiplex quantitative experiments, the intensities of the various reporter ions are also shown.

Mass Calculations

Actual Mass = (Observed m/z - mass of one proton)* Charge

Calculated Mass = The uncharged peptide mass calculated from the amino acid sequence

Delta Da = Actual Mass - Calculated Mass

Delta ppm = (Delta Da / Calculated Mass) * 10^6

In some cases the Delta ppm value is marked with an asterisk (*). This indicates that the Delta Da value was approximately equal to the mass of a neutron, causing Scaffold Quant to infer that a triggering error caused by misidentification of a monoisotopic peak has occurred. As a result, the program has adjusted the actual peptide mass by the mass of a neutron when calculating the Delta ppm. When this happens, there is a discrepancy between the displayed Delta Da and Delta ppm values.

Manual Validation

The Spectrum and Fragmentation Table tabs in the Visualization Pane below provide information to allow manual checking of the spectral matches. A spectrum can be invalidated by unchecking its Valid box. If a group of spectra are selected, they can all be checked or unchecked simultaneously by using the green plus or red minus buttons which appear when a bulk selection is made.

Changing the set of valid peptides requires recalculation and regrouping which may be time-consuming, so Scaffold Quant allows the User to review multiple spectra and accumulate a set of changes which can then be applied in a single operation. Spectra awaiting update are clearly marked to aid in the review process. To apply the changes, the User clicks **Apply Pending Changes**.

If the **Show Hidden** box at the top of the screen is not checked, invalidated spectra disappear when the changes are applied. If **Show Hidden** is selected, the spectra are shown, but with their boxes unchecked. This allows the User to restore previously invalidated spectra.

The User may return to the initial identifications by using the option **Clear User Peptide Validation** in the Experiment menu.

In the example illustrated in Figure 6-4, the spectra matching the peptide sequence “ADDFAEEGKK” were manually marked invalid and the changes are pending.

Figure 6-4: Evaluate individual spectrum for manual validation.

The Visualization Pane

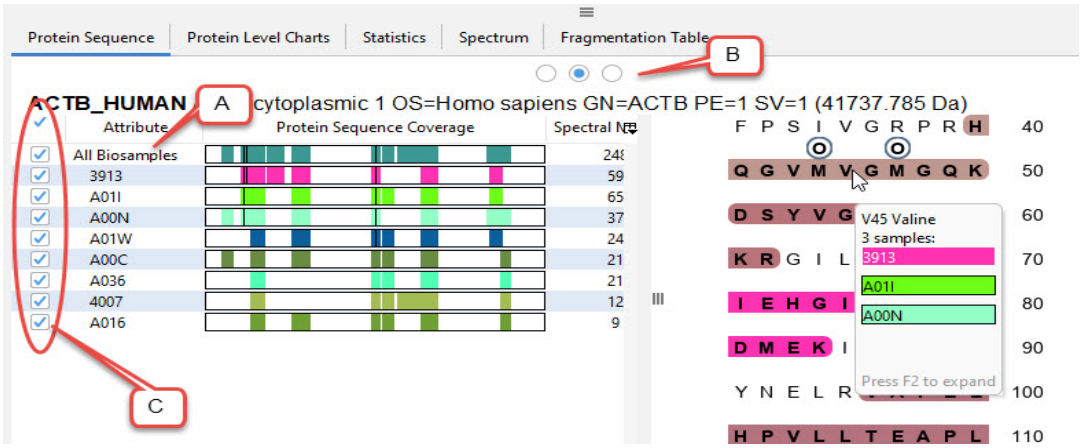
The lower Proteins pane consists of five tabs described in detail below:

- The Protein Sequence tab
- The Protein Level Charts tab (active only with precursor intensity data)
- The Statistics Tab
- The Spectrum tab
- The Fragmentation Table tab

Protein Sequence tab

This tab consists of two different displays, which may be viewed together or individually. The three radio buttons at the top of the tab select whether to display the sequence coverage diagrams, the protein sequence or both.

Figure 6-5: Protein Sequence tab in the Proteins View. All MS samples shown. Controls allow the choice of the level of summarization (A), display of just the left or right view or both (B), and which groups should be represented by coloring in the sequence view (C).*The right-click context menu also provides options for highlighting modifications and performing an automatic BLAST search of the protein sequence.*



Sequence coverage comparisons

The sequence coverage diagrams allow comparison of peptide coverage between samples or categories at the level of summarization selected at the top of the display. Each diagram graphically depicts the portions of the protein sequence covered by identified peptides in a specific MS Sample or other categorical grouping. The Attribute column gives the name of the sample or Attribute, and the number of Spectral Matches and Percent Coverage are also shown.

The check-boxes at left determine whether the corresponding peptides should be highlighted in the protein sequence view. The peptides corresponding to the colored regions for each checked line are highlighted in the same color in the protein sequence. Unchecking the All MS Samples row unchecks peptides for all samples. This row must be checked in order to display and sequence coverage.

Sequence Display Options

By default, when sequences are covered in more than one sample or category, the colors are blended. This is called Overlay, but the right-click context menu provides two other display options as well. The User may select to stack the colors above the sequence (Stacked), or to switch to a gray-scale display in which the shade represents the number of spectral matches to the region (Spectral Coverage).

Hovering over an amino acid in the sequence gives its name, position and any modifications it

might contain and lists each sample in which that amino acid was identified. When the Spectral Coverage display option is selected, it also shows the number of spectral matches at that point in the sequence. The peptide selected in the Peptides Pane is indicated by brackets, and if the context menu option Use Blinking Cursor is enabled, it also blinks. Modifications are indicated by colored circles, either solid or outlined depending on another option in the context menu. Each modification is indicated by its own color, and these colors may be selected by the User. Clicking on an amino acid in the sequence display filters the tables above to show only peptides covering that region. Clicking on Show All Peptides clears the filter.

Context menus

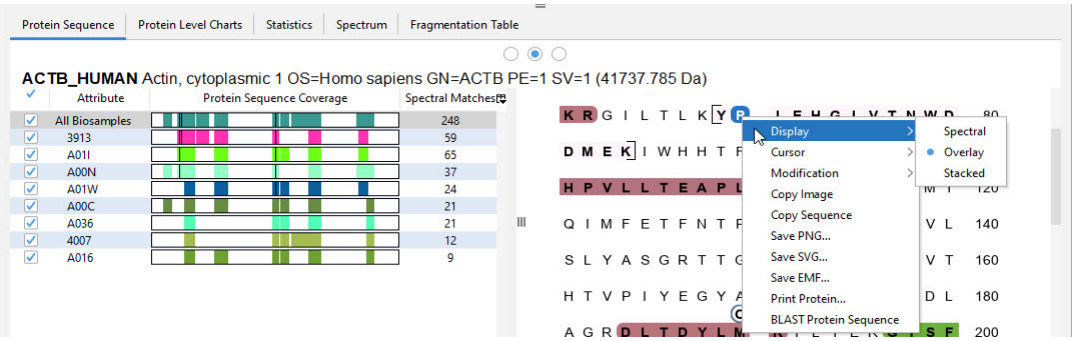
The User can right-click on the sequence display to open a context menu that has the following options:

- Display - Selects the format in which to display the sequence coverage (see Sequence Display Options)
- Cursor - Offers two checkboxes that control how the cursor is displayed. If Animation is checked, the cursor blinks, and if Show Cursor is checked, brackets indicate the selected peptide. These two controls operate independently.
- Modification - Offers options to edit the modification colors and to either display modifications as an outlined (open) circle if Outline Modifications is checked or a fully colored circle if it is unchecked.
- Copy Image - Copy a vector-based image to the clipboard which you can then paste into a third party tool such as Microsoft PowerPoint for easy editing and manipulation.
- Copy Sequence—Copies the protein sequence in text format to the Clipboard so it may be pasted into a third party tool such as Microsoft Word.
- Save PNG... —Saves the currently displayed spectrum in Portable Network Graphic format, which is a bitmap format, and opens the a file dialog box in which you can specify the name and directory for this saved PNG file.
- Save SVG... —Saves the currently displayed spectrum in Scalable Vector Graphic format and opens the a file dialog box in which you can specify the name and directory for this saved SVG file.
- Save EMF... —Saves the currently displayed spectrum in Windows Metafile format and opens the a file dialog box in which you can specify the name and directory for this saved PNG file.
- Print Protein...—Opens the Print dialog box in which you can specify the options for printing (printer, number of copies, and so on) the currently displayed spectrum.
- BLAST Protein Sequence—Select this option to automatically open an Internet browser session and display the Standard Protein BLAST page (blastp) for the selected protein.

The Power of Summarization

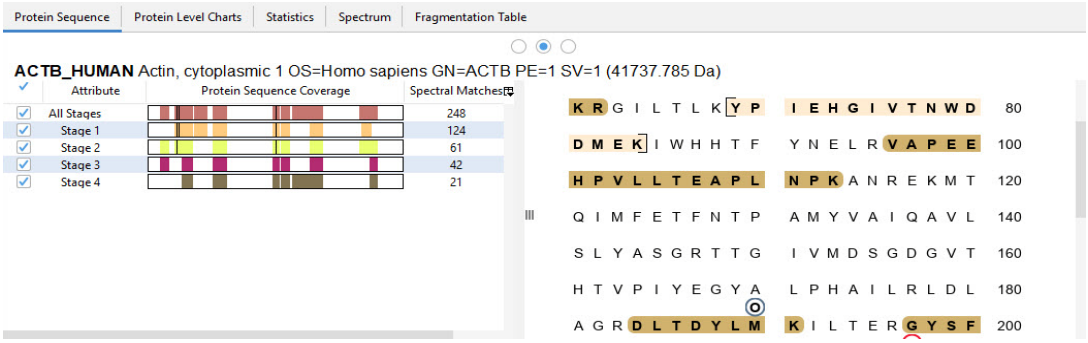
Using summarization in the Protein Sequence tab readily provides new insight. Identified sequences can be evaluated at the level of any Category. In the example shown in Figure 6-6, some sequences are observed only under one experimental condition (20-400-1200 HPLC conditions), while other sequences are observed under a different combination of conditions. (The colors are overlaid when different attributes share the same evidence.)

Figure 6-6: Easily compare differences in identified peptide sequences by Summarization Level.



In Figure 6-7, the coverage was summarized by Category which represents the lab. This same technique can be used for any attribute that could be applied to a particular dataset, for example, tissue types, treatment types, demographic differences, measurement conditions and more.

Figure 6-7: Choosing a different Summarization level (by Category), provides a different view of the data



In Figure 6-7, sequences found in both groups are colored with a blend of pink and green (a purple-blue color). Sequences found only in Lab 28 are green.

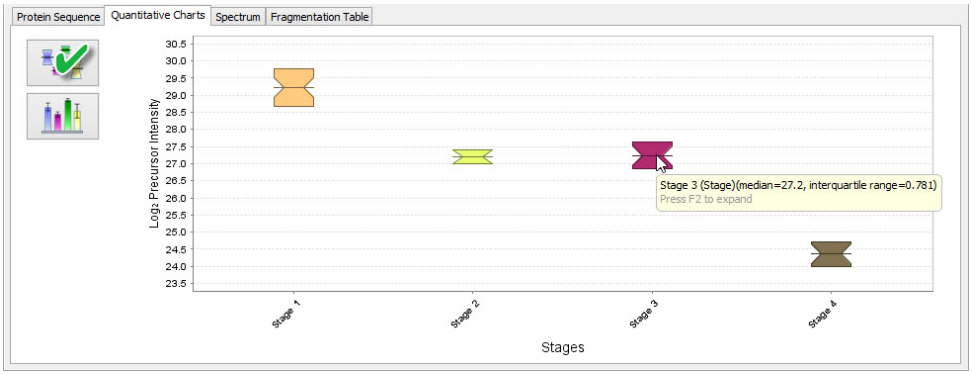
Protein Level Charts tab

The **Protein Level Charts tab** displays a graph that shows the Precursor Intensity for a protein in each MS samples or summarized to the selected level of summarization as chosen from the Summarization Bar in the Scaffold Quant main window.

Four formats are available for this graph—a Box plot, a Bar chart, a Line Graph and a Violin Plot. Different colors are randomly assigned to each MS sample or selected summarization.

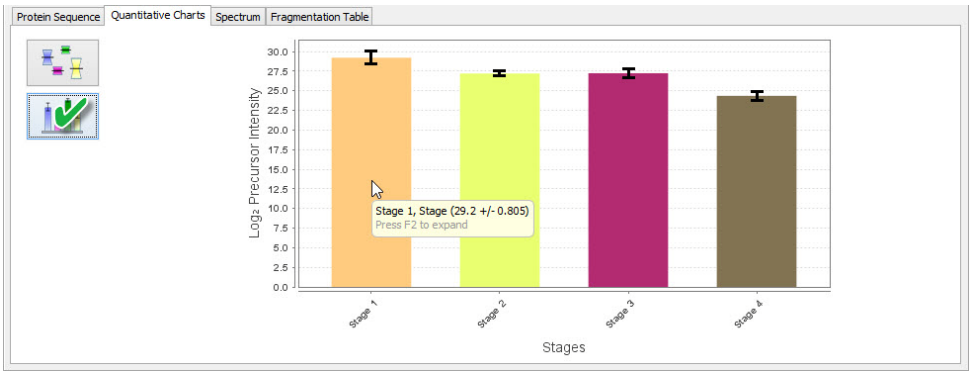
- Box plot—(The default plot) A box plot is a convenient way of graphically depicting the spread and centers of groups of numerical data. In Scaffold Quant, the box plot displays the median value and range for the Intensity at the selected level of summarization. Placing the cursor on any box plot displays the information about the median and interquartile range for the corresponding quantitative value.

Figure 6-8: Proteins View: Quantitative Charts tab - Box plot.



- Bar chart—The Bar plot displays the median value and range for the Intensity at the selected level of summarization. Placing the cursor on any bar displays information about the median and range for the corresponding quantitative sample or level of summarization.

Figure 6-9: Proteins View: Quantitative Charts tab - Bar chart.

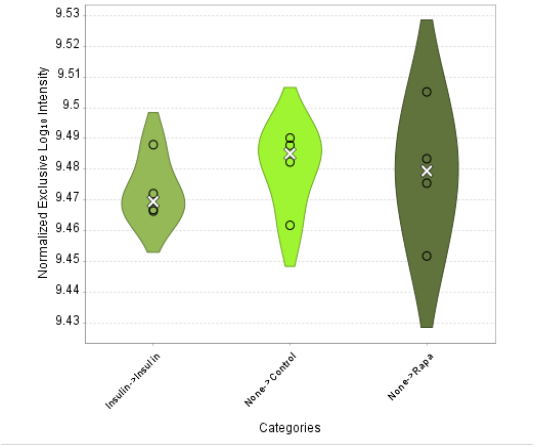


The charts plot the values that are shown in the Protein table when the Log_{10} Intensity Display Type is chosen.

- Line Graph — This graph plots the Intensity value in each sample or summarization group. This results in a line depicting the pattern of expression for each group.
- Violin Plot - The Violin Plot is similar to the box plot, but it also displays the probability density of the data at various values. In the Violin plot, the median value is depicted as an "X". If the plot depicts five or fewer points, each point is shown; if more than five points, a bar is shown, indicating the 50% confidence interval. The body of the violin plot

indicates the 95% confidence interval of the values, as determined by Kernel Density Estimation (KDE) using a Gaussian kernel and Silverman's rule.

Figure 6-10: Proteins View: Violin Plot



Statistics Tab

Statistics Table - The contents of this table depend on which statistical test has been applied to the experiment. It presents all of the values relevant to the calculation of that particular test. For example, for the ANOVA/t-test, columns displayed are: Sum of Squares, degrees of freedom, Mean Square, F-statistic, and significance of F-statistic.

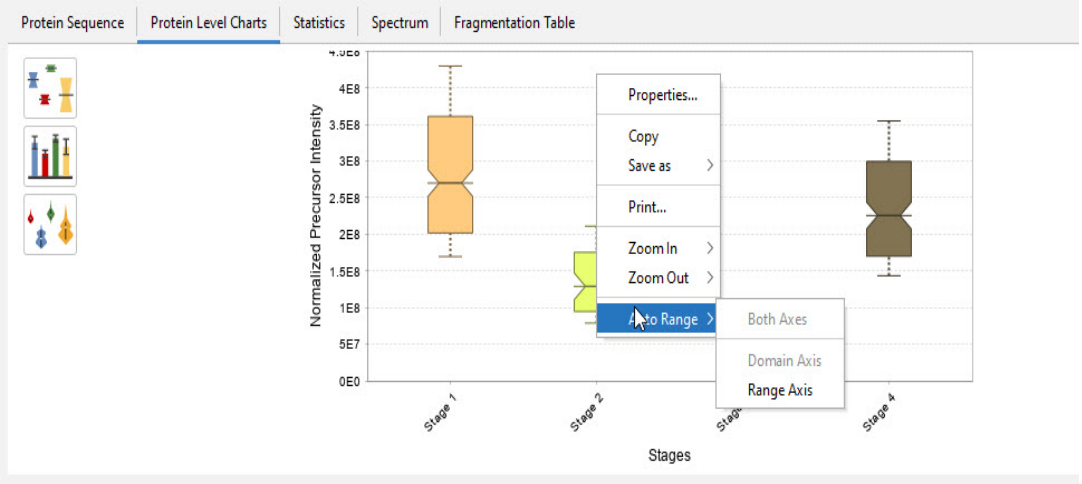
Interaction Chart - This chart is designed to help the user interpret interaction effects between primary and secondary variables. The chart consists of a series of graphs with one line for each Attribute in the secondary comparison Category. Each of these plots a series of points, one point for each Attribute in the primary comparison Category, representing the average quantitative value of all technical replicates with the indicated combination of Attributes.

If there is no interaction between the primary and secondary variables, the lines would be roughly parallel, with the slope of the lines indicating the effect of the primary variable. If the lines are not parallel, but do not cross, there is an interaction effect but it is still possible to draw conclusions about the effect of the primary variable with the understanding that the secondary variable affects the size of that effect.

If, on the other hand, the lines cross, it means that the two variables exhibit significant interaction and it is impossible to draw general conclusions about the effect of the primary variable.

Context Menu in the Charts

Figure 6-11: Right click context menu



- Copy Image - Copies the currently displayed chromatogram in a JPEG format to the clipboard. The user can then paste this copied image, which is appropriately pre-sized for publication, into a third-party application such as Microsoft Word.
- Save as - Allows the user to save bit map or vector images with a level of resolution appropriate for publication purposes.
- Print — Opens the Print dialog box through which the user can specify the options for printing (printer, number of copies, and so on) the currently displayed spectrum.
- Zoom Out—Returns the zoom out magnification to 100%.

This is identical to a single click of the left mouse button after using the Click and Drag feature.

- Auto Range —Rescale the graph to the original level in the selected dimension.

IRS Normalization Tab

Note: Feature available only with Labeled Quant license

In isobaric labeling experiments, channels may be normalized using a technique known as IRS normalization (see IRS Normalization Algorithm). This tab graphically displays information about the calculation.

Spectrum tab

The spectrum tab graphically depicts the MS/MS spectrum to assist in manual validation of peptide-spectrum matches. Dragging the mouse over a section of the image zooms the display. A variety of options for adjusting the display, copying, saving or printing the image, or performing a blast search of the peptide sequence are offered in the right-click context

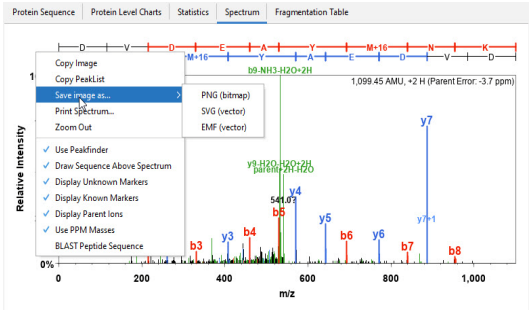
menu.

Figure 6-12: Proteins View Visualization pane Spectrum tab

Interacting with the Spectrum tab:

- Clicking anywhere on the spectrum displays the M/z value for the position.
- Clicking and holding the left mouse button anywhere on the spectrum and then dragging the mouse pointer to any position in the spectrum. While dragging the mouse pointer, the Start and Stop M/z values for the segment are displayed as well as the length for the segment. Releasing the button zoom in on the selected region. A single click of the mouse returns the zoom out magnification to 100%.
- Right-clicking anywhere on the spectrum opens a context menu that has the following menu options:

Figure 6-13: Right click context menu



- Copy Image - Copies the currently displayed spectrum in a JPEG format to the clipboard. The User can then paste this copied image, which is appropriately pre-sized for publication, into a third-party application such as Microsoft Word.

- Copy Peaklist—Copies the peak list for the currently displayed spectrum to the Clipboard which the User can then paste into a third party tool such as Microsoft Excel.
- Save Image as - Allows the User to save bit map or vector images with a level of resolution appropriate for publication purposes.
- Print Spectrum—Opens the Print dialog box in which the User can specify the options for printing (printer, number of copies, and so on) the currently displayed spectrum.
- Zoom Out—Returns the zoom out magnification to 100%.

This is identical to a single click of the left mouse button after using the Click and Drag feature.

- Use Peakfinder—Selected by default. When selected, a tool-tip opens that displays the ion designation, M/z value, and M/z error (in ppm) for the daughter ion that is closest to the current cursor position.
- Draw Sequence Above Spectrum — Selected by default. When unchecked it moves the sequence within the boundaries of the spectrum
- Display Unknown Markers—Selected by default. If a peak has not been assigned to any identified ion, then the peak m/z value is shown with a question mark (?).
- Display Known Markers—Selected by default. If a peak has been assigned to an identified ion, then the peak m/z value is labeled with the related type of ion.
- Display Parent Ions—Displays parent ions in the spectrum. If parent ions are much more prevalent than daughter ions, then clear this option so that you can more easily examine the daughter ions.
- Use PPM Masses—Selected by default. Displays the M/z error in ppm in the tooltip for a daughter ion. Clear this option to display the M/z error in AMU.
- BLAST Peptide Sequence—Select this option to automatically open an Internet browser session and display the Standard Protein BLAST page (blastp) for the selected peptide.

Fragmentation Table tab

This tab displays information about the peptide-spectrum match in tabular form. Options for copying or printing this data or saving the image of the table are also available in the context menu by means of the right-click

The Similar Proteins Tab

The Similarity Table

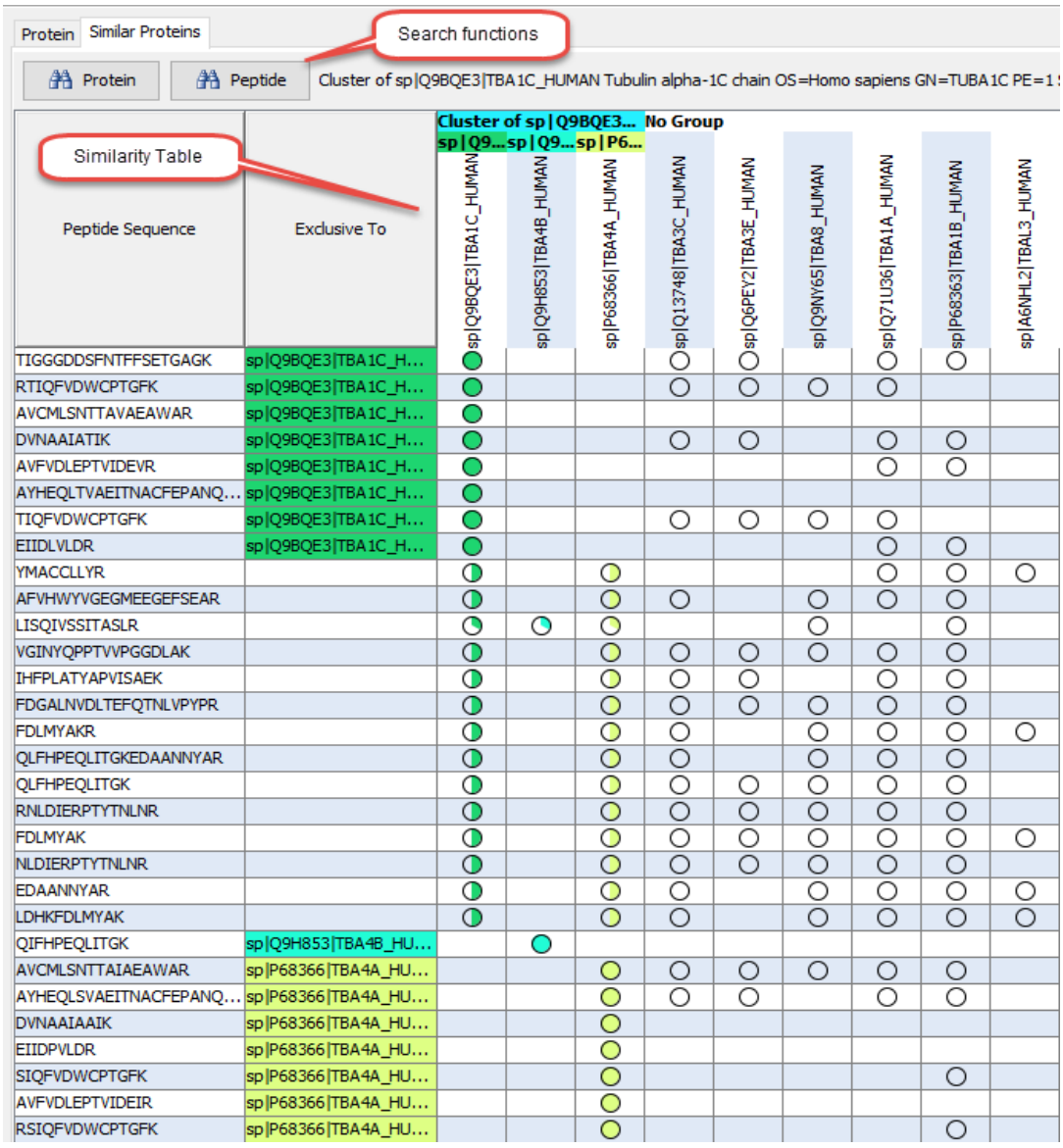
The Similar Proteins tab of the Proteins View allows the user to examine the role of shared peptides in protein identification and quantification. It provides insight into the results of parsimonious protein inference.

All of the peptides associated with the displayed protein or protein cluster and any proteins with which it shares peptides are displayed in the left column of the table. The next column shows which of the peptides are exclusively associated with one specific protein, with a distinct color assigned to each protein.

Each of the remaining columns represents a single protein. Colored headers at the top of the column indicate assigned proteins and clusters, or designate the proteins as “No Group” if the protein has been discarded from the identified proteins set by virtue of the principle of parsimony.

In each protein’s column, peptides that are associated with the protein are indicated by a colored circle. The degree to which the circle is colored indicates the degree of “exclusivity” of the peptide amongst the assigned proteins.

The Grouping Method is displayed at the top of the table. Note that if MZID grouping is used, parsimony is not applied and it may be possible to see proteins with no exclusive peptides.



Search Functions in the Similarity Tab

Although the Similarity Tab is reached through the Proteins View of a specific protein, search functions are provided which allow the user to access the Similarity Tab of any protein or peptide in the experiment. This is very helpful when trying to understand why a specific protein which the user expected to find may be missing from the experiment. Often these proteins will be found in the “No Group” columns. Searching for a specific peptide of interest in the Similarity View may provide insight into why a specific isoform or homologous protein was preferred in the analysis.

Chapter 7
The Visualize View

The Visualize View offers a variety of graphical tools to help the user discover quantitative trends and relationships between proteins and samples.It consists of three tabs: Principle Component Analysis, Quantitation, and a Heat Map of the filtered analytes list shown in the Samples table.

- “Quantitation Tab” on page 118, which provides a Volcano Plot to help identify which proteins exhibit significant differential expression, a Quantitative Scatterplot to show the relationships between values in different samples or categories, and GO annotation displays to help identify the biologically significant proteins in the experiment.
- “Principal Component Analysis tab” on page 124, which helps identify the underlying sources of variation in the data set.
- “Heatmap Tab” on page 127, which provides a graphical environment where a Heat map based on the findings listed in the Samples table is provided.

Quantitation Tab

The Quantitation tab consists of four sets of plots, which are visible under different circumstances: the Volcano Plot appears when a quantitative test comparing two groups has been applied; the Quantitative Scatterplot appears when the selected summarization level contains two or more groups; the Quantitative Trend Chart appears whenever there is precursor intensity data and two Gene Ontology plots appear below these figures when GO terms have been applied in the experiment.

Volcano Plot

The Volcano Plot makes it easy to identify proteins that exhibit significant quantitative differences among the samples. In a volcano plot, the y-axis represents the $-\log_{10}$ of the p-value and the x-axis represents the $\log_2$ fold change, calculated by comparing the rolled up intensity value of one category to that of another. As a result, the plot is only available when the following conditions are met:

1. The summarization hierarchy is arranged so that exactly two Attribute values have been selected for comparison in the statistical analysis, allowing calculation of a fold change.
2. A Statistical Test has been applied.

In the Volcano Plot, points with statistically significant p-values are colored green while points with p-values that would have been significant had a FWER not been applied are colored yellow. This corresponds to the coloring of the Statistical Test Result column in the Samples View.

The proteins that are most likely to be of biological significance are those which are statistically significant and also show large positive or negative fold changes.A few of these proteins are marked in Figure 7-1.

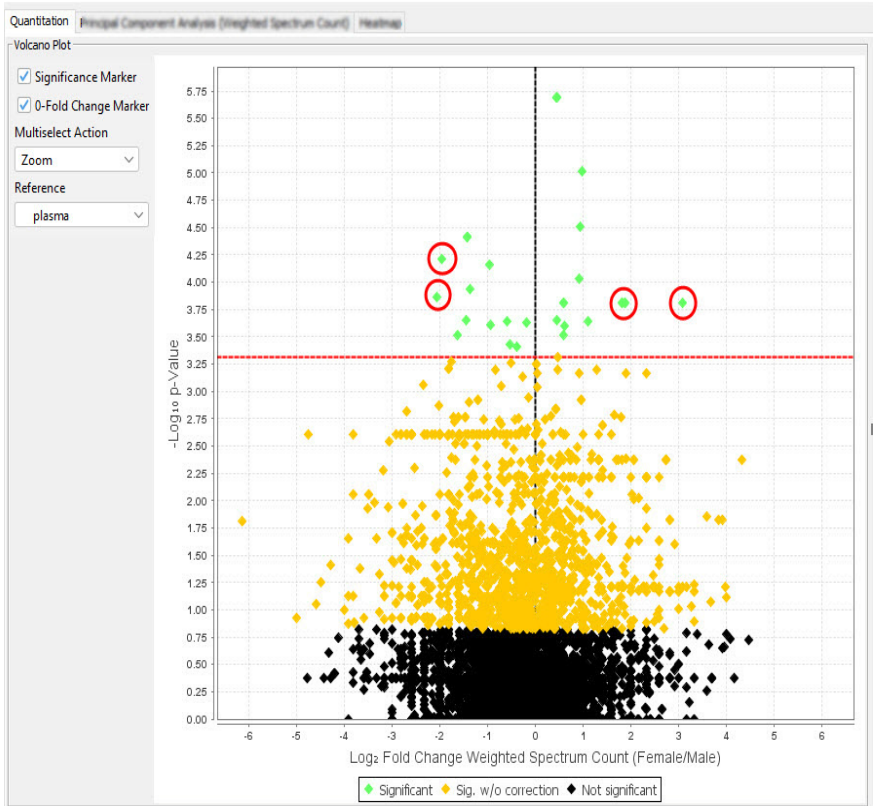
Plot Actions

A check box at the left of the plot toggles display of a horizontal dotted line that marks the significance threshold.

Another optional line, controlled by its own check box, marks the point at which there is no difference between the compared values, i.e. the zero fold change line.

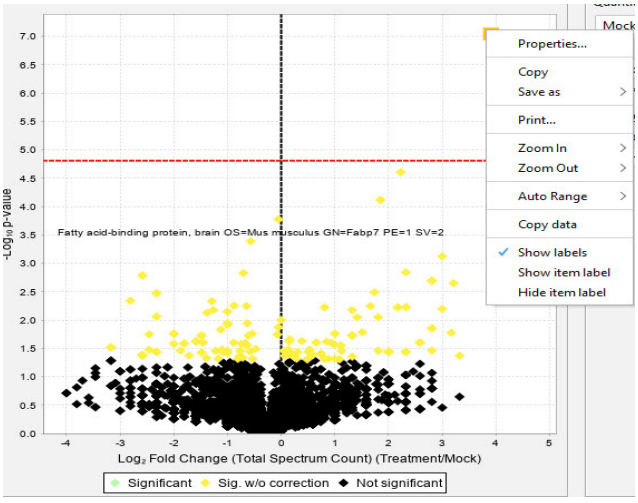
The **Multi-Select Action** pull down menu determines the behavior of the pane when the user selects a rectangular area in the plot by holding down the left mouse button and dragging the cursor. Depending on the option chosen, the graph may zoom in on the selected area or the proteins in the selected area may be tagged with stars. When stars are added or removed through the multi-select mechanism, a pop-up message informs the user of the action. The selected proteins will display the modified star status in the Samples View. This provides a convenient mechanism for filtering the proteins list based on the plot.

Figure 7-1: The Volcano Plo



Label Points in the graph, the user right-clicks to bring up a context menu. To label an individual point, select the point and choose **Show item label**. This can be done repeatedly to label a set of points. To label all points, turn off any individual labels and click **Show labels**. containing options to label only the currently selected point, label all points, or remove all labels. Clicking **Hide item label** removes the label of the currently selected point. Clicking **Show labels** when it is already checked clears all labels.

Figure 7-2: Labeling Points in a Chart



This menu also offers options that allow for copying the chart, saving the Image in various formats, adjusting the chart properties or copying the data displayed in the chart in order to recreate the graphic or analyze the data using a different program.

Quantitative Scatterplot

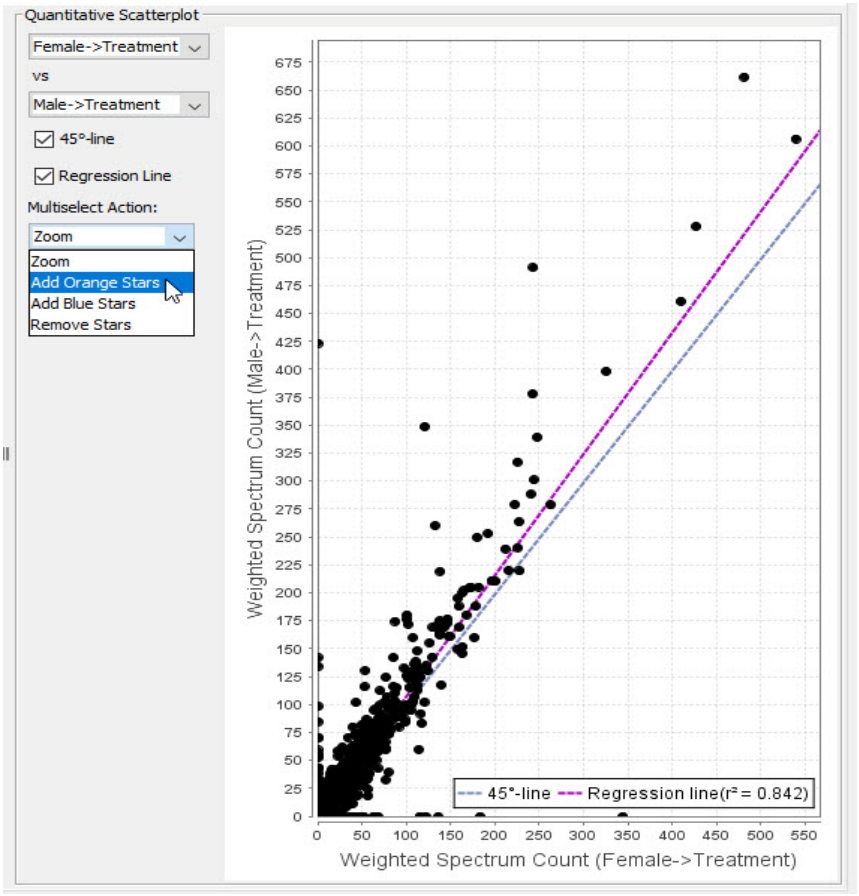
In the Quantitative Scatterplot, the quantitative values of proteins in one category are plotted against the corresponding values in another category, where the categories represent Attributes rolled up to the Summarization Level in the summarization hierarchy. The two categories are selected from pull down lists to the left of the graph.

Examination of the Quantitative Scatterplot assists the user in assessing the relationship between the two categories, and in identifying outliers, which may be proteins that are especially important in distinguishing the two groups.

Two lines may be displayed on the graph, each activated by a check box to the left of the plot. Points would be expected to cluster along the 45 degree line if the two categories are completely correlated. The regression line shows the result of performing a linear regression calculation of y on x. The correlation coefficient is shown in the legend when the regression line is displayed.

The Quantitative Scatterplot also features Labeling of Points and a Multi-select Action pull down. In the Quantitative Scatterplot, the color of the point indicates the star status of the protein in the Samples View. Orange indicates the protein has an orange star, blue a blue star, and purple indicates the protein has both orange and blue stars. Setting the star status through the Quantitative Plot allows the user to return to the Samples View and filter on characteristics recognized in the graph.

Figure 7-3: The Quantitative Scatterplot



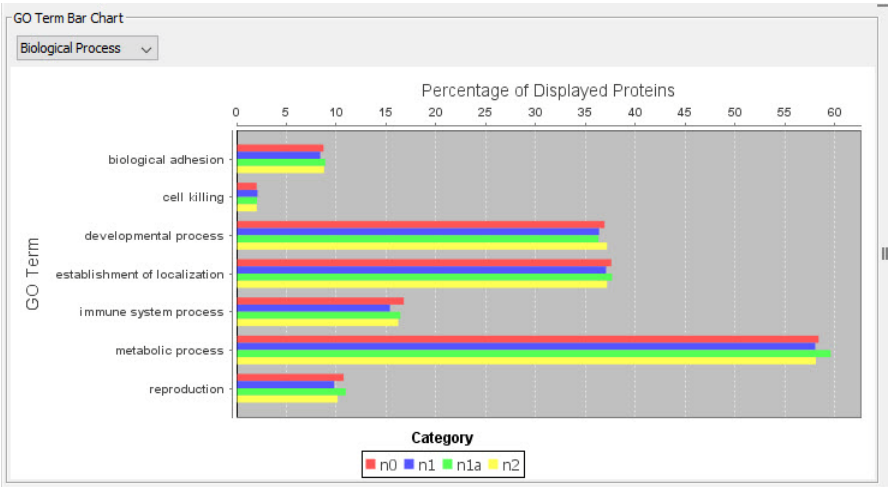
GO Annotations Bar Chart

The GO Annotations Bar Chart appears only if Gene Ontology Annotations have been applied to the Experiment (see Gene Ontology Annotations).

This chart allows the user to compare the percentages of proteins representing specific biological functions across samples. At the top of the pane is a drop-down to allow the user to select the desired Gene Ontology (Biological Process, Cellular Component, or Molecular Function). A group of bars is then displayed for each term in the selected ontology which has been chosen for display in the Displayed GO Terms dialog.

For each GO term, there is a bar for each category as defined by the currently selected comparison level in the summarization hierarchy. Each bar represents the percentage of the currently displayed proteins in that category that are tagged with the indicated GO term. Bars are color coded and a legend indicates the categories.

Figure 7-4: GO Annotation Bar Chart

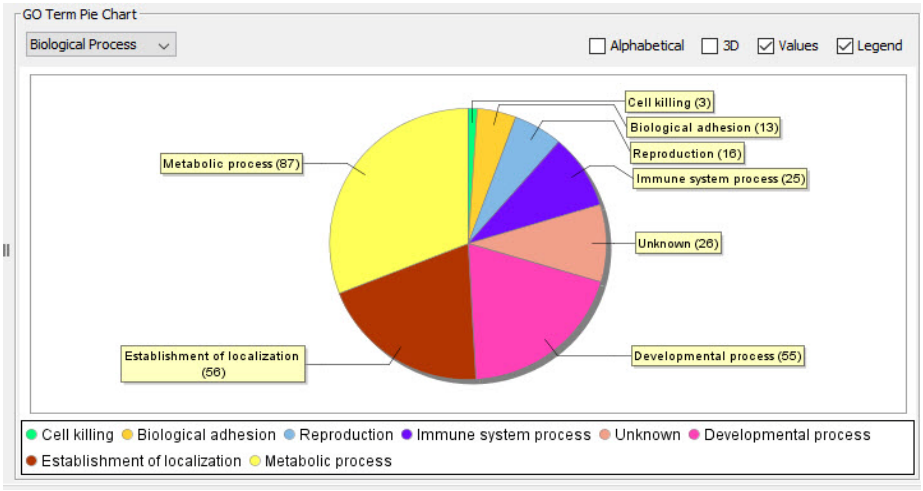


GO Annotations Pie Chart

The GO Annotations Pie Chart displays the relative proportions in of proteins which are associated with specific biological functions. The user may select a specific Gene Ontology (Biological Process, Cellular Component, or Molecular Function) from the drop-down at the top of the pane. The Pie chart then shows the percentage of proteins in the experiment which are associated with each GO term in that ontology which has been chosen for display in the Displayed GO Terms dialog.

A number of display options are also offered through check boxes at the upper right of the pane. These include options to sort the GO terms in the chart in alphabetical order, to switch to a three-dimensional visual presentation, to show or not show the percentage values, and to display or not display the chart legend.

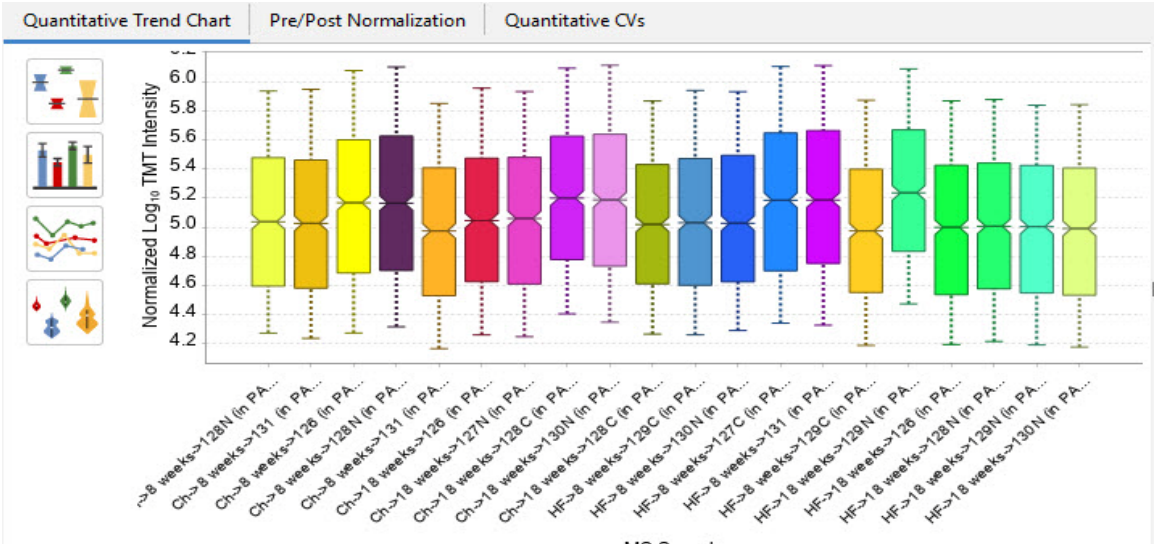
Figure 7-5: GO Annotations Pie Chart



Quantitative Trends Charts

The Quantitative Trends charts present the overall protein levels at the selected level of summarization. A variety of chart formats is offered, including a box plot, bar chart, line graphs and violin plots.

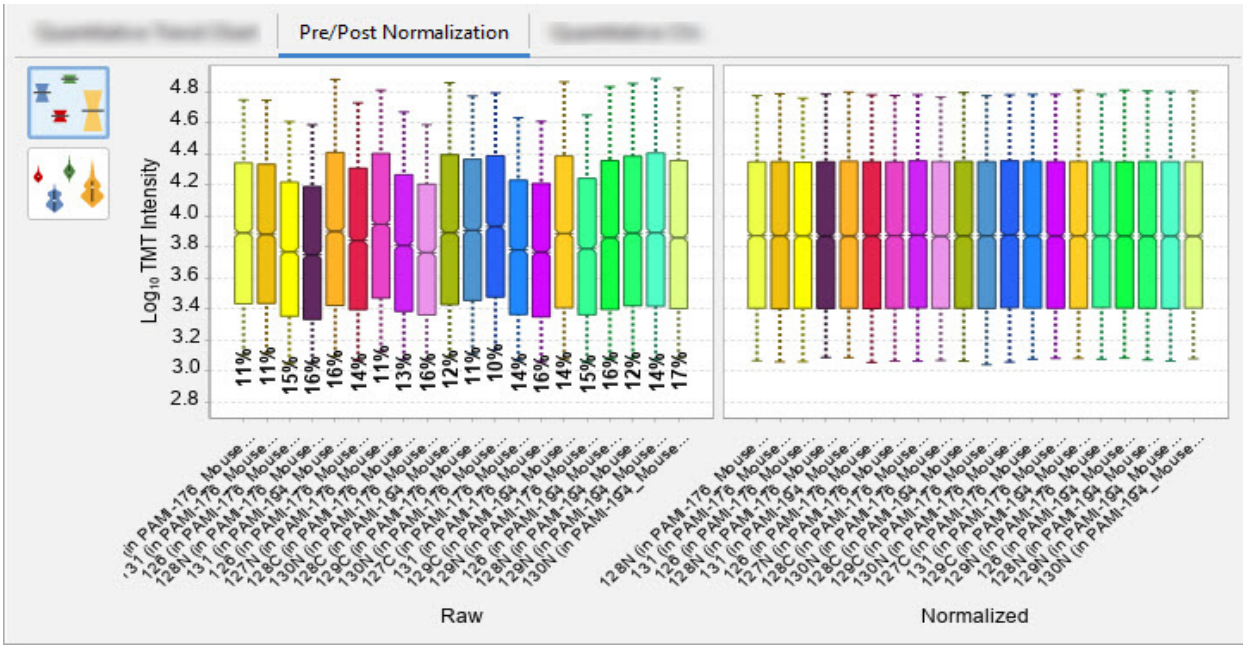
Figure 7-6: The Quantitative Trends Chart



Pre/Post Normalization Charts

These charts display, in either box plot or violin plot form, the overall protein levels of the various samples before and after TRL normalization is applied.

Figure 7-7: Pre/Post Normalization Box Plot



Quantitative CVs Chart

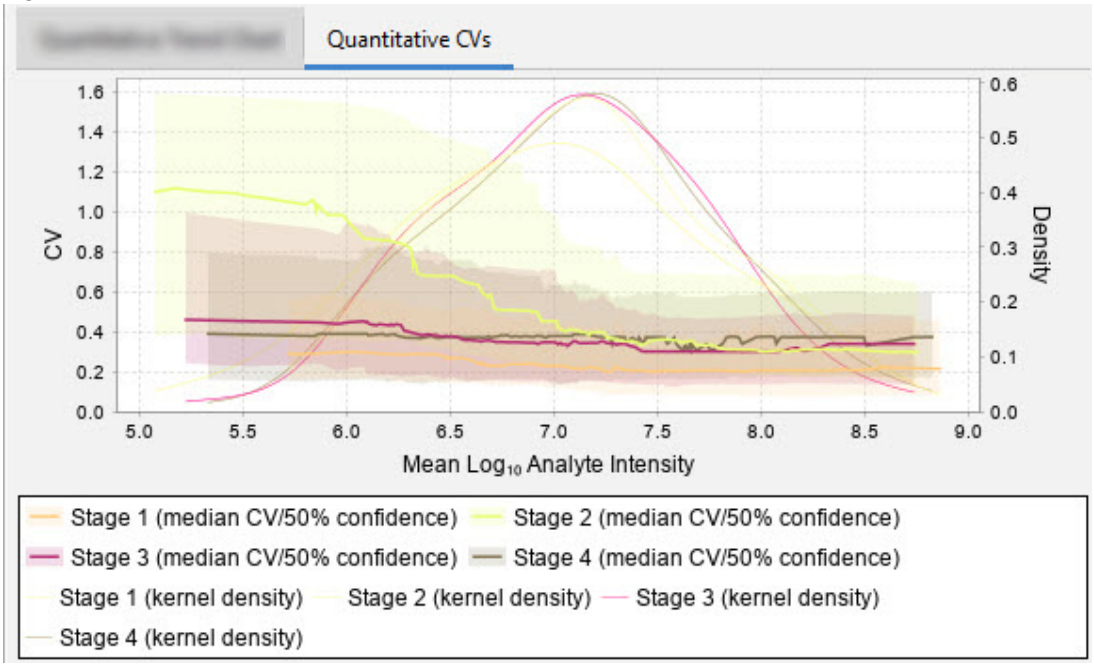
The Quantitative CVs chart contains two distinct types of plots, which, in combination, provide insight into the reliability of the quantitative values calculated in the experiment. The Quantitative CVs chart displays the relationship between mean protein intensity and Coefficient of Variation (CV) for each group of samples at the currently selected Comparison Level (see "Summarization Level" on page 73).

Shaded areas indicate the 50% confidence interval for CV, with the thick line showing the median value, computed for a sliding window of at least 50 proteins. The window size increases with increasing numbers of proteins. Note that the median line will appear flat if there are very few proteins in the experiment. The CV level is indicated in the y-axis displayed on the left of the chart.

A second set of plots is also displayed in the same figure. These plots show the distribution of protein intensities within each group. The intensity levels are indicated in the y-axis on the right of the chart. The intensities plotted here are computed with a Gaussian kernel density estimate with bandwidth set by Silverman's rule.

The chart is built from the unfiltered, thresholded set of analytes in the experiment. Values are gathered from the level of summarization directly below the Comparison Level, and analytes are ignored if any values at that level are missing or imputed.

Figure 7-8: The Quantitative CVs Chart



Principal Component Analysis tab

Principal Component Analysis (PCA) is a tool used to identify the underlying sources of variation in a data set. PCA looks for patterns of expression among the proteins that can be used to group samples in

meaningful ways. When used in combination with the flexible summarization offered in Scaffold Quant, this provides a powerful tool for exploring the biological meaning of quantitative differences observed in an experiment.

Figure 7-9: The Principal Component Analysis Tab

The PCA tab consists of four Plots:

The Overview

The Overview consists of a series of graphs where one Principal Component is plotted against another. The points in these graphs represent samples and the X and Y coordinates are the values computed from the corresponding Principal Component functions. Samples tend to cluster in different ways depending on the Principal Components applied. Clicking on a graph in the Overview selects that combination of Principal Components for display in greater detail in the Loadings and Scores Plots below.

The Scree Plot

The Scree Plot graphs the percentage of variance explained by each Principal Component. The lower (red) plot shows the percentage of variance explained by the individual Principal Components, while the upper (blue) plot is cumulative, so the first point shows the variance explained by PC1, the second by PC1 and PC2, etc.

The Scores Plot

The Scores Plot graphs one Principal Component against another. The points in the Scores Plot represent samples (rolled up to the Biological Replicate Level specified in the summarization hierarchy) and the X and Y coordinates represent the values of the Principal Components.

The Scores Plot has several associated controls, which are found on its left. Check boxes allow the user to toggle display of the legend and sample names on or off. Another option is to show confidence ellipses for the Attributes at the Comparison Level in the summarization hierarchy. A **confidence ellipse** is a colored ellipse that represents the area in which we can expect a sample with a certain Attribute to appear, with a certain level of confidence. The confidence level is adjustable through the Confidence (%) spinner. Note that if an Attribute is represented by two or fewer values, no ellipse is displayed.

The Scores Plot allows Labeling of Points through the right-click context menu and zooming in by dragging the mouse over an area of interest.

The Loadings Plot

In the Loadings Plot, each point represents one protein. The coordinates of each protein are a measure of the contributions of that protein to each of the Principal Components.in the plot. For example, if the plot displays PC1 on the x-axis and PC2 on the y-axis, points far to the left and right represent proteins that contribute strongly to Principal Component 1. The proteins near the top and bottom contribute strongly to PC2.

Options available to the left of the Loadings Plot allow the user to toggle on or off the display of protein names and vectors. The vectors connect each protein point to the origin, and the slope of the vector corresponds to the relative contribution that protein to each Principal Component.

Points in the Loadings Plot may be labeled through the right-click context menu The Loadings Plot also features a Multi-select Action pull down. As in the Quantitative Scatterplot, the color of the point indicates the star status of the protein in the Samples View. Orange indicates that the protein has an orange star, blue a blue star, and purple indicates the protein has both orange and blue stars. Setting the star status through the Loadings Plot allows the user to return to the Samples View and filter on characteristics recognized in the graph.

Further information about PCA

For details about how PCA is calculated in Scaffold Quant, see .

Heatmap Tab

Heat maps are an efficient method of visualizing complex data sets organized in two dimensional tables or matrices. Through the application of two independent procedures to a data matrix, heat maps make patterns more visible to the eye. The first procedure reorders columns and rows according to a "closeness" criteria which groups together in space highly similar data. The other procedure translates a numerical matrix into a color image[1].

In order to produce a figure that is meaningful, Scaffold Quant restricts the heatmap to a maximum of 1000 proteins. As a result, it may be necessary for the user to filter the protein set before accessing the heatmap. One method of accomplishing this is to use the star filters. For example, one might select the first 1000 proteins in the Samples View, right-click and choose Stars > Add Orange, then click on the empty star icon in the toolbar to filter out all unstarred proteins. Alternatively, one could filter on statistical significance, GO terms or some other criterion.

When the protein list includes less than 1000 proteins, Scaffold Quant constructs a heatmap from the data appearing in the Samples table and displays it in the Visualize View Heatmap tab.

Figure 7-10: Visualize View: Heatmap tab

The Heatmap tab includes the following three components, each containing a number of graphical tools:

- The**Heatmap Landscape pane** -- which shows the overall heat map.
- The Heatmap Details pane --Which shows a selected portion of the heat map and includes labels for the displayed columns and rows.
- The Heatmap Display controls--which lists the Display type selected from the Samples View and three toggle buttons

Heatmap Landscape pane

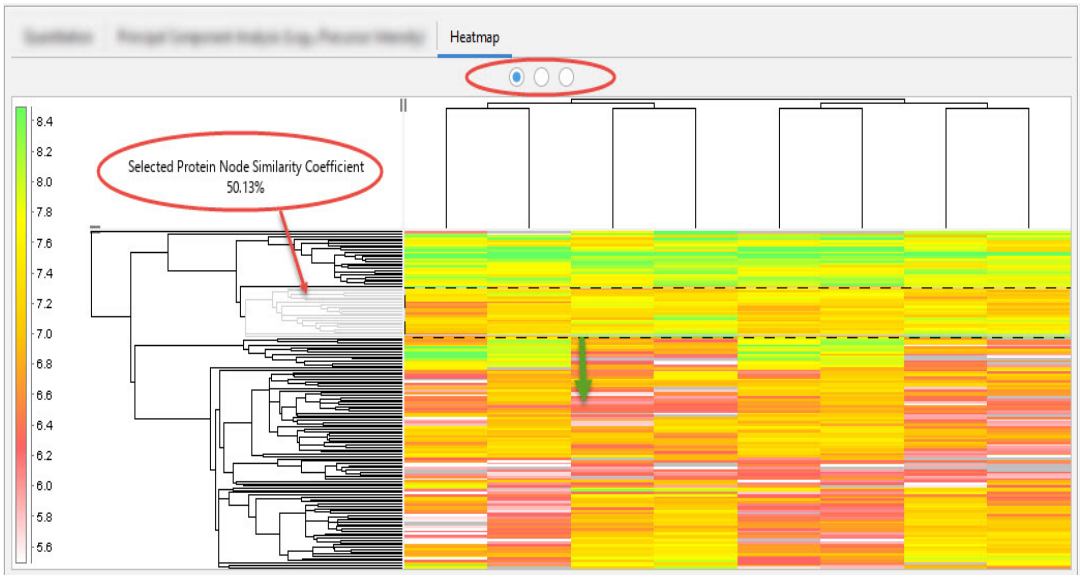
The heat map shown in this pane is created using the data summarized in the Samples Table, see Figure 7-11. As in the Samples table the rows in the Heat map represent protein groups, but not protein clusters as at times are shown in the Samples table when one of the options available in Experiment > Protein Clustering is chosen. Each column contains data from every MS sample or selected level of summarization as chosen from the Summarization Bar in the Scaffold Quant main window.

The Display Type listed in the Heatmap Display controls of this tab, determines the type of quantitative information used to reorder the data.

More information about the way the Heat map is constructed can be found in the appendix section Heat map clustering

1. Key, M., "A tutorial in displaying mass spectrometry based proteomic data using heat maps." BMC Bioinformatics 2012 13(Suppl 16):S10. doi:10.1186/1471-2105-13-S16-S10

Figure 7-11: Heatmap Landscape pane



The Heat map includes a colored landscape, color coded according to the legend shown on the left side of the Heat map pane. The type of color coding depends on the Display type chosen in the Samples View, which can be customized at will through the Color Options button.

Grays represent missing values.

Dendrograms representing the output of the hierarchical clustering are shown on the left and top sides of the Heat map landscape. The root of each dendrogram represents a single object or cluster of size 1.

Clicking and dragging the mouse over the Heatmap landscape allows the user to select a section of the map that is then shown in larger detail in the Heatmap Details pane

Sections of the map can also be highlighted and selected by clicking over the different nodes in either dendrogram or in both, thus allowing the user to select desired sets of protein groups and/or sets of samples. When doing so the calculated node distance is shown on the top left side of the pane.

Context menu

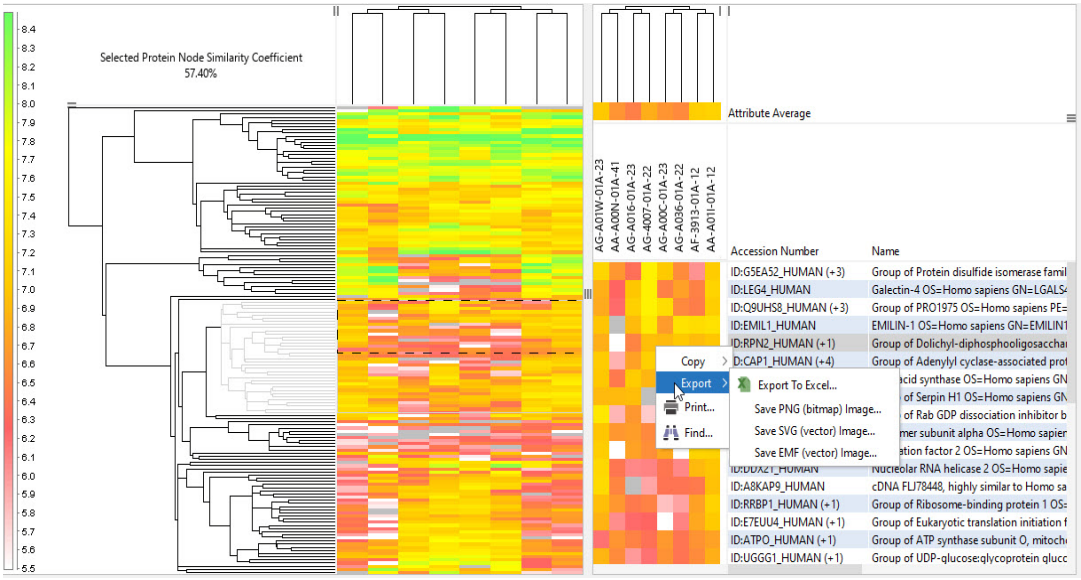
A right-click of the mouse while hovering over the Heatmap landscape provides a context menu with zoom and export options, see Figure 7-12.

Hot keys for zooming in and out of the heat map are also available:

- ZOOM IN: CTRL + NumPad +
- ZOOM OUT: CTRL + NumPad -

The Export PNG (Bitmap) Image... command opens an Export Preview dialog of the current heat map with options for toggling the inclusion or exclusion of some of the components of the exportable picture like, for example, any dendrogram or the colored landscape.

Figure 7-12: Heat map Context menu



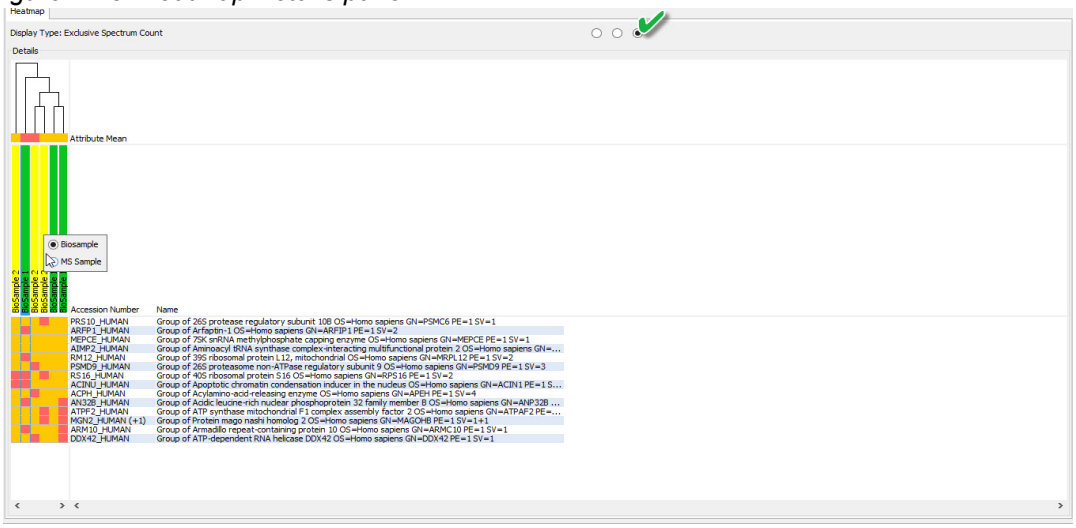
Heatmap Details pane

When sections of the Heat map are selected the Heatmap details pane is populated with the selected section of the map and the related information about the protein groups associated with each row and the related MS sample or selected level of summarization as chosen from the Summarization Bar available in the Scaffold Quant main window.

When the user right-clicks on the column headers, a context menu allows the selection of the level of summarization he/she wants to see represented within the column headers.

The table that represents the selected portion of the Heat map has the common properties of all Scaffold Quant tables, see Display pane.When no selection is active the pane will appear empty with instructions to help the user populate the pane.

Figure 7-13: Heatmap Details pane



Heatmap Display controls

These controls are located in the top section of the Heatmap tab. The three toggle buttons allow the user to determine which of the two heatmap panes display:

- The Heatmap Landscape pane only (left button)
- The Heat map and the Details pane are shown together (middle button)
- The Heatmap Details pane only (right button).

Heatmap Column Clustering Options

While a heatmap generally clusters the data in both dimensions (e.g. the samples or categories and the proteins), Scaffold Quant offers the option to view the heatmap with the proteins clustered, but with the columns displayed in the order in which they appear in the Samples View. This may be useful, for example, when viewing a time-series experiment, where it provides the ability to recognize groups of proteins that exhibit certain patterns of response to the treatment over time.

To turn column clustering on or off, select **Edit>Preferences>Heatmap** and check or uncheck the **Cluster Columns for Heatmap** box.

Chapter 8
The Publish View

The Scaffold Quant Publish View displays detailed information about the data loaded and the analytic methods applied in the current experiment. This is usually required for publication.

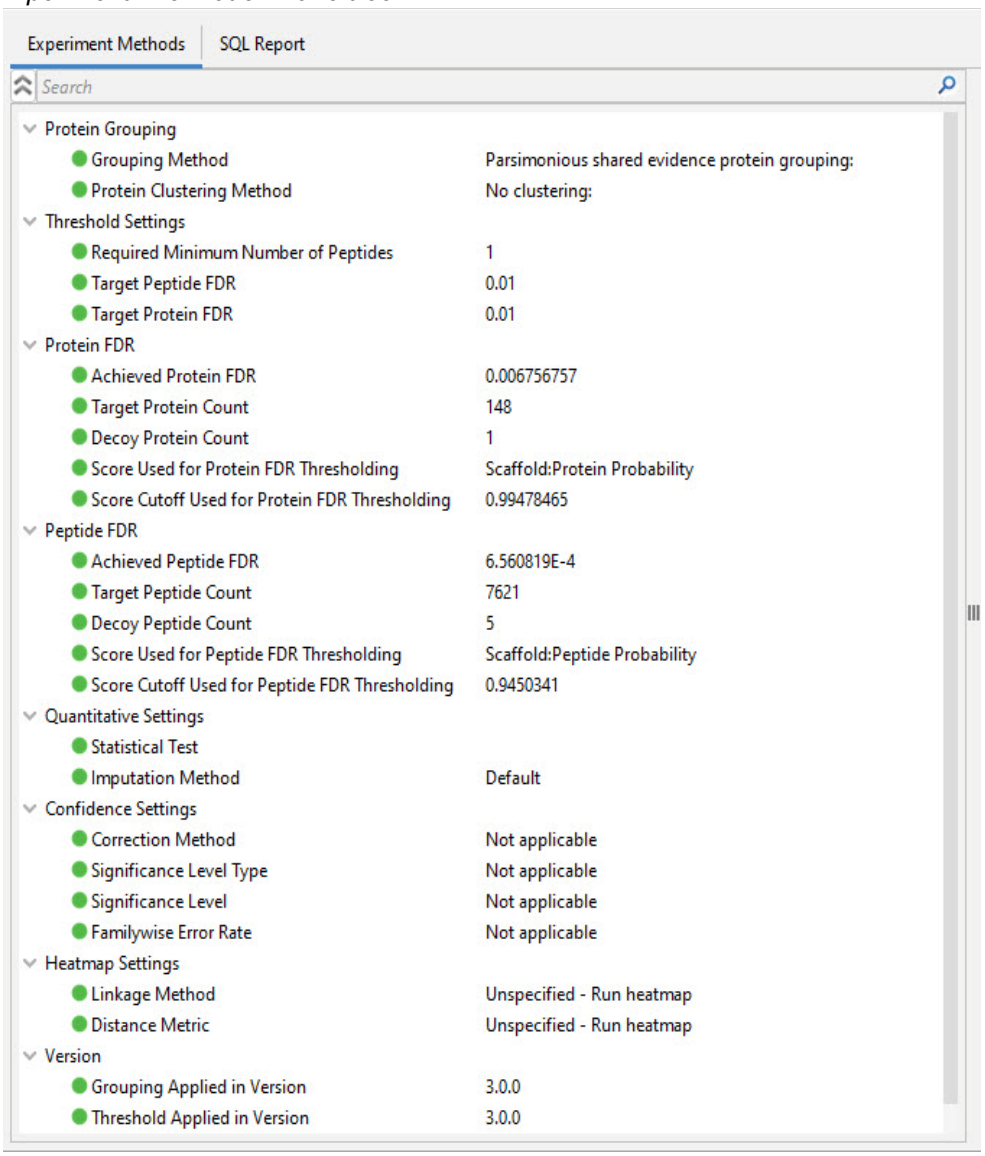
The Publish View has two tabs:

- Experiment Methods Tab
- SQL Export tab

Experiment Methods Tab

The left side of the Experiment Methods Tab contains a tree that lists the parameters characterizing the current experiment and their values. This information must be included in publications.

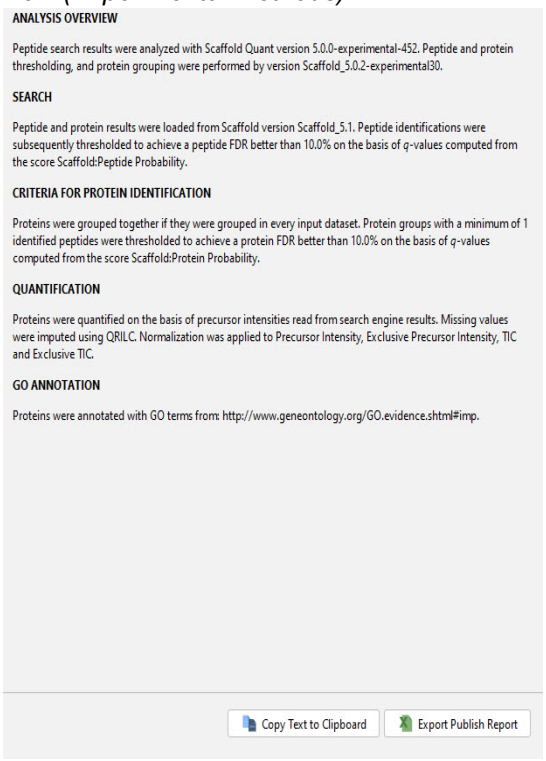
Figure 8-1: Experiment Information Pane tree



Analysis Overview

The right side of the Experiment Methods tab of the Publish View provides a draft version of the analysis parameters in text format to help the user in writing the Methods section of a publication or poster.

Figure 8-2: TheAnalysis Overview (Experimental Methods)



Report Buttons:

- Copy Text to Clipboard - copies the contents of the Analysis Overview pane for pasting into a text editor.
- Export Publish Report - exports the contents of the Experiment Methods table as a CSV file which can be opened in Excel.
- Export Supplementary Data - exports a CSV file similar to the Samples Report suitable for submission as a supplementary data table.

SQL Export tab

The experiment files created by Scaffold Quant, *.sfdb files, are essentially SQLite databases.

The SQL export tab is a SQLite graphical interface where a Scaffold Quant experiment file can be searched as a database using SQLite commands. This allows a User to create custom tables exportable to Excel.

A depiction of the schema of a *.SFDB file is shown in the Appendix in the Scaffold Quant User's Guide. The default SQL query which appears in the SQL pane when the program is opened displays all tables in the database.

Figure 8-3: The SQL tab

Figure 9: The SQL tab

The SQL Export tab contains four different panes:

- The SQL pane
- The Saved Queries pane
- The Results pane
- The Icon pane

The SQL pane

Through the SQL pane it is possible to directly explore the information stored in a Scaffold Quant file using SQLite queries.

- The SQL text box --where the user can enter, copy and paste SQL queries.
- The SQL Icon pane -- which contains the Run query button and a text box and a save button to save queries.

The results of the queries are shown in The Results pane. The saved queries are listed in the The Saved Queries pane

Example:

List of tables available in *.SFDB files.

SELECT name FROM SQLite_master WHERE type='table' ORDER BY name;

The Saved Queries pane

When a query is saved, with a name selected by the user, it will appear in this pane from where it is conveniently available to be launched again whenever needed.

The Results pane

When the run query button is pressed, if there are no errors, a table with the results of the query will appear.

To export the results the User needs to right click the mouse and select the menu option **Export > Export to Excel** and save the table to a tab delimited file that can be easily opened in Excel.

The Icon pane

The icon pane contains an icon to save new queries to a file that can later be retrieved and an icon to import previously saved queries.

Chapter 8
The Publish View

Chapter 9
Protein Grouping and Clustering

Gap Filling

When samples are loaded into Scaffold Quant, they contain associations between identified peptides and proteins. In some cases, peptides that are common to more than one protein may be associated with different proteins in different samples. This could result in a failure to group proteins which would have grouped had they been evaluated based on all of the evidence available in the combined experiment.

An example occurs when two or more Scaffold experiments are combined. Different exclusive peptides may have been detected in different experiments, with the result that when Scaffold's protein grouping method applied parsimony in each experiment, it discarded different proteins and thus eliminated their potential peptide-protein connections. When the samples are combined in Scaffold Quant, some of the evidence needed to properly group the proteins is missing. If not corrected, this could lead to overlapping sets of proteins or failure to align peptides for a common protein across samples.

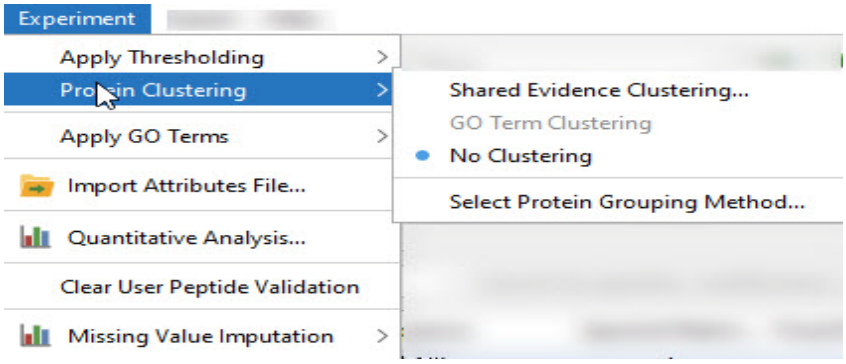
To remedy this situation, Scaffold Quant adds a peptide-protein connection in each sample to any identified peptide that is connected to an additional protein in any other sample. This allows the protein grouping algorithm to consider all of the possible protein associations for each peptide and to provide the optimal grouping solution for the experiment as a whole.

G*ap filling occurs only when loading data, not when protein grouping is done on an existing file.*

Grouping and clustering options

Options for grouping and clustering proteins are available through the menu **Experiment > Protein Clustering**.

Figure 9-1: Protein clustering menu options



Protein clustering

Scaffold Quant supports two algorithms that create clusters from peptide evidence:

- **Shared evidence Clustering --** This algorithm looks at identified peptides belonging to more than one protein and creates a metric that defines whether these proteins should be clustered. This approach creates a protein group whenever all identified peptides belonging to at least two different proteins are shared in each MS Sample. The computational details of the algorithm are reported in the appendix Terminology.

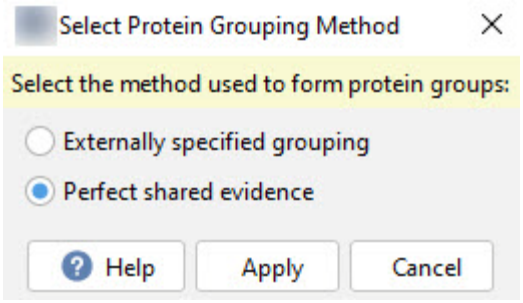
  GO Clustering - Proteins are clustered functionally, based on their sharing of GO annotations (see GO Terms Clustering).

- **No Clustering** - Proteins are grouped by the selected grouping method, but no additional clustering is performed.

Experiment >Protein Clustering > Select Protein Grouping Method...

When this menu item is selected, the **Select Protein Grouping Method** dialog opens, offering two options for the type of grouping Scaffold Quant will apply to simplify the protein list.

Figure 9-2: Select Protein Grouping Method



- **MZID specified grouping** -- This option uses the protein grouping indicated in the original input files loaded into Scaffold Quant or in the Scaffold experiment that created the SFDB file.
- **Perfect shared evidence** -- When this option is selected, Scaffold Quant groups proteins that share all of their peptides. **Parsimony** is applied, and potential protein identifications whose peptide evidence is better explained by another protein are eliminated (see Perfect Shared Evidence Protein Grouping).

Perfect Shared Evidence Protein Grouping

Protein grouping proceeds as follows:

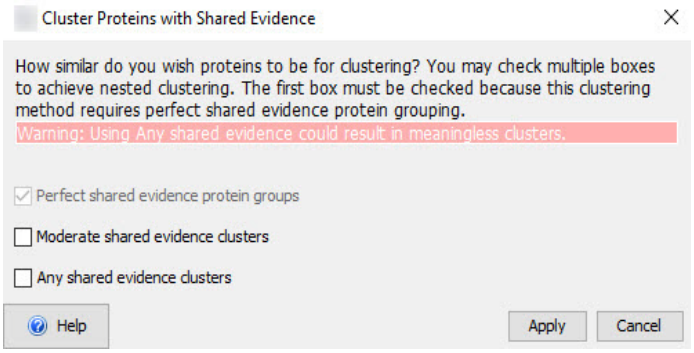
1. Each protein, is matched to its peptide sequences.
2. Sets of proteins with the same identified peptides are formed into protein groups.

3. Proteins are selected for inclusion by iteratively choosing proteins by the following procedure:
    - If a protein has any peptide sequences that match exclusively to it, include that protein.
    - When all exclusive peptides have been accounted for, score each remaining protein as the sum of the scores over the unclaimed peptides, and include the protein with the maximal score.
    - Break ties using (i) the sum of the score function over all peptides, then any remaining ties with (ii) proteins with known sequences and (iii) shorter protein sequences.
    - Remove all proteins from consideration whose peptides constitute a "subset" of peptides already included in chosen proteins.

Experiment >Protein Clustering > Shared Evidence Clustering...

When this menu option is selected, the dialog box **Cluster proteins with shared evidence** opens. It contains three check boxes, the combination of which defines the level of peptide evidence clustering that Scaffold Quant will apply to the protein list. Selecting more than one level of clustering creates nested clusters.

Figure 9-3: Shared evidence clustering



- **Any shared evidence clustering** -- This option clusters proteins that share any peptides. It tends to create a smaller number of large clusters.

This level of clustering must be used carefully, as it may group proteins that have very little in common, producing meaningless clusters.

- **Moderate shared evidence clustering** -- This option clusters proteins that share about half of their peptide evidence with another protein in the cluster. It creates smaller, more meaningful clusters.
- **Perfect shared evidence protein groups** -- This option corresponds to protein groups where all proteins share all of their peptides. The option is grayed out, since it must always be selected in order to apply the other levels of clustering. If MZID grouping is active when shared evidence clustering is selected, Scaffold Quant automatically groups proteins using perfect shared evidence protein grouping.

When selecting both the **Moderate** and **Any** shared evidence clustering options, sub-clusters will be formed and displayed as a tree structure in the Samples View.

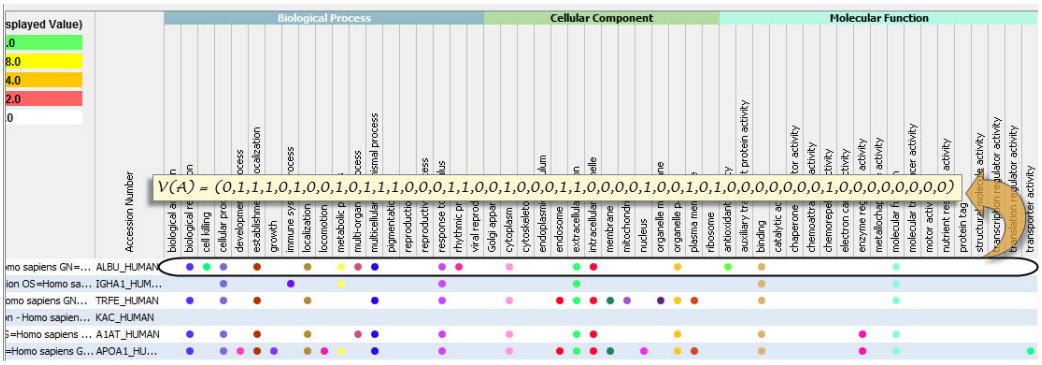
GO Terms Clustering

The option for GO Terms clustering is available when Gene Ontology (GO) Annotations have been applied. To apply GO annotations, first import and select a GO terms database through Edit GO Term Options. Once the GO database has been imported, the option **Experiment > Apply Go Terms** becomes available allowing the user to obtain GO terms for the proteins listed in the Samples Table. A number of columns, corresponding to the GO terms included in the Displayed GO Terms list (see The Displayed GO Terms Tab) are added to the Samples Table. Circles in these columns indicate that the protein in that row is associated with the GO term in that column.

Each protein can then be thought of as being characterized by the list of GO terms with which it is associated. In Scaffold Quant, this information may then be used to rearrange the protein rows and to form clusters of proteins that share similar GO terms.

The association of a protein to a particular GO term is shown with the presence or absence of a colored dot under a specific GO term column.The presence or absence of a dot is recorded as a one or a zero respectively. These values are then considered the components of a multidimensional vector, *v(A)* which summarizes the GO Association for any protein in the list.

Figure 9-4: Vector v(A) summarizing GO term associations for protein A



All the vectors associated with the proteins in the Samples Table can be represented as points in a multidimensional space. Clusters are formed by defining a distance metric and the computing a maximum radius for cluster formation as described in the appendix Distance Based Clustering.

The metric chosen in the Case of GO Terms clustering is the square of the Euclidean distance, which calculates differences in location, with zero indicating that two points occupy the same location, see the appendix .GO Term Clustering distance metric

Graphical Display

In the Samples Table, clusters are characterized by a small pie-icon displayed in the Samples View. The fraction of this pie that is filled in is the minimum shared evidence value over all of the pairs of proteins in the cluster.

Notice that even if the user is applying moderate shared evidence clustering it is possible for this pie to be empty.

For example, if A has peptides {x, y}, B has peptides {b, x}, and C has peptides {c, y}, then A and B share half of their peptides, as do A and C, so all three proteins will be in the same cluster, but B and C share no peptides, so the minimum shared evidence = 0 and the pie will be empty.

Currently the colors are:

$$0 \Leftarrow \text{red} < 1/3 \Leftarrow \text{yellow} < 2/3 \Leftarrow \text{green} < 1$$

When the shared evidence is exactly 1, the proteins are considered a 'group,' not a 'cluster,' and no pie-icon appears.

Chapter 10
Quantitative Methods and tests

Label-Free Quantitative Methods

Scaffold Quant supports label-free quantitative methods based on either spectral counts or on MS1 precursor intensity measurements. Values may be normalized across samples, and a variety of quantitative statistical tests are available to establish differential expression among combinations of factor levels or treatments.

Label- free quantitative methods directly use raw spectral data from parallel MS runs to determine relative protein abundances. The most commonly used approaches include Spectral Counting and Precursor Intensity based methods. Scaffold Quant supports both of these:

- Spectral Counting
- Precursor Intensity Quantification

Scaffold Quant offers the option of normalizing the levels of proteins detected in the samples or groups of samples at the current level of summarization, to appropriately adjust the values shown in the Samples table for comparison purposes. The normalization scheme works for the common experimental situation in which individual proteins may be up-regulated or down-regulated, but the total amount of all proteins in each sample or level of summarization is about the same. It is not appropriate if the total amount of protein varies widely from one sample to the next.

Note that Scaffold Quant computes normalization in slightly different manners depending on the type of quantitative method selected.

Spectral Counting

Spectral counting relates protein abundance to the number of peptide-spectrum matches (PSMs) for each protein detected in a sample or group of samples at a selected summarization level. This type of quantitation is based on the empirical observation that the more of a particular protein is present in a sample, the more Tandem MS spectra that are collected for peptides of that protein[1].

Scaffold Quant provides a number of different types of spectral counts, selectable through the Display Type pull-down list:

- **Total Spectrum Count** ---The total number of spectra associated to a single protein group, including those shared with other proteins.

1. *Quantitative mass spectrometry in proteomics: a critical review.* Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B., Anal Bioanal Chem. 2007 Oct;389(4):1017-31. Epub 2007 Aug 1. Review. PMID: 17668192 [PubMed - indexed for MEDLINE]

- **Weighted Spectrum Count --** Number of spectra associated with only a specific protein group plus the apportioned number of spectra shared with other proteins, see Weighted spectrum counts.
- **Exclusive Spectrum Count** -- The number of spectra associated only with a specific protein group.
- **Total Unique Peptide Count** -- The number of different amino acid sequences that are associated with a specific protein including those shared with other proteins.
- **Exclusive Unique Peptide Count** --The number of different amino acid sequences, regardless of any modification that are associated oly with a single protein group.
- **Percentage of Total Spectra** -- The percentage of the total number of spectra in the experiment that are associated with a specific protein group in a specific sample or group of samples.
- **Percent Coverage** -- The percentage of the number of amino acids in the protein sequence contained in one or more peptides identified in the sample or group of samples.

When counting, Scaffold Quant includes every PSMwhich is considered valid. ValidationQuantified status can be manually changed by the user as described in Manual Validation. Hiding proteins from the Samples table does not affect the reported counts.

Spectral counting normalization

Scaffold Quant normalizes spectral counting data across all samples in the experiment. This is done by adjusting the values appearing in the Samples table so that each column has the same total number of unique spectra or peptides. This type of normalization is independent of the number of rows appearing in the table.

Depending on the level of summarization selected, the Samples table will display a different number of columns and the counts will be rolled up accordingly. The normalization coefficient used to adjust the values for a specific column is given by the ratio of the average number of uniquespectra in all of the columns divided by the number of uniquespectra in a column.

For verification, the user can extract the distinct spectrum counts for each MS sample in the experiment running the following SQLite query in the SQL Export tab:

```
SELECT COUNT(DISTINCT SPECTRUM_ID) as ‘Unique Spectrum count’,
SAMPLE_NAME

FROM MatchToHypothesis mth JOIN MSSample ms ON ms.ID is mth.MSSAMPLE_ID

WHERE IS_VALID is 1

GROUP BY MSSAMPLE_ID
```

The average number of distinct spectra will vary according to the number of columns present in the table at a selected summarization level.

Precursor Intensity Quantification

An increasingly popular option for quantitative Proteomics is Precursor Intensity Quantitation, which offers a good compromise between the accuracy of labeled techniques and the simplicity and lower cost of label-free quantitation. This method relies on measuring the signal intensity of the peptide precursors representing a specific protein at the MS level and comparing these intensities across samples.

Scaffold Quant is designed to provide easy and confident validation, visualization and quantitation of search results. It does not read raw files and does not have direct access to precursor information; instead it reads intensity data already computed by the identification software. Currently, Scaffold Quant is able to obtain precursor intensity information from Thermo Proteome Discoverer, Mascot Distiller and MaxQuant files through SFDB or mzIdentML exported from Scaffold and Agilent Spectrum Mill mzIdentML files.

Scaffold Quant has the option to normalize precursor intensity values across samples and calculate fold changes at different summarization levels. Statistical tests of differences in the calculated intensities are also offered, including the T-Test, ANOVA and Coefficient of Variation as appropriate to the experimental design.

Computing Precursor Intensities

Precursor Intensity Quantitation is based on the principle that the area of a peak in the MS1 chromatogram provides a measure of the relative abundance of the corresponding peptide in the sample. Peptides are identified based on their MS/MS spectra, and then the corresponding MS1 peaks are identified in each LC-MS/MS run. The areas under these peaks are calculated and normalized and their ratios are used as a measure of the relative abundance of the peptides in different samples. Relative quantities of proteins are estimated by combining the precursor intensities of the constituent peptides in various ways. The mzIdentML files loaded in Scaffold Quant need to have this information already included in them since the program does not group the different peptides into proteins.

The following illustration of the typical LC-MS/MS analysis of a peptide is reproduced from Lai *et al*[2].

Figure 10-1: Identification of a peptide through LC-MS/MS analysis

2.Xianyin Lai, Lianshui Wang, and Frank A. Witzmann, *Issues and Applications in Label-Free Quantitative Mass Spectrometry*, International Journal of Proteomics, 2013, vol. 2013, Article ID 756039, 13 pages. DOI:10.1155/2013/756039

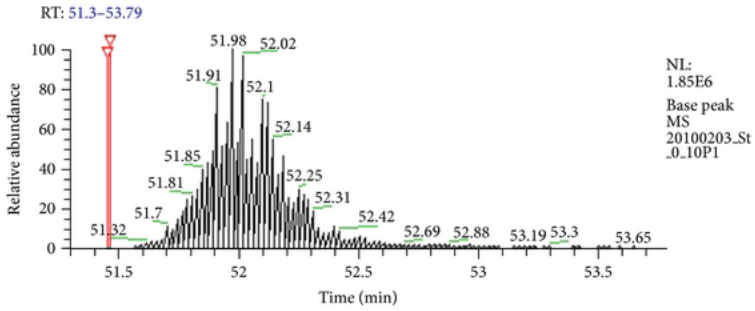


Figure 10-1-a: The peptide is eluted from the LC column and its ion intensity is plotted as a function of the retention time.

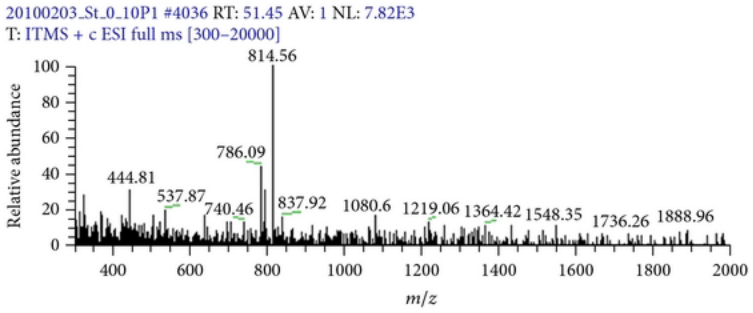


Figure 10-1-b: At the first scan time shown in red in (a), a full MS scan is performed. The ion with m/z 786.09 is selected as a precursor ion for MS/MS analysis.

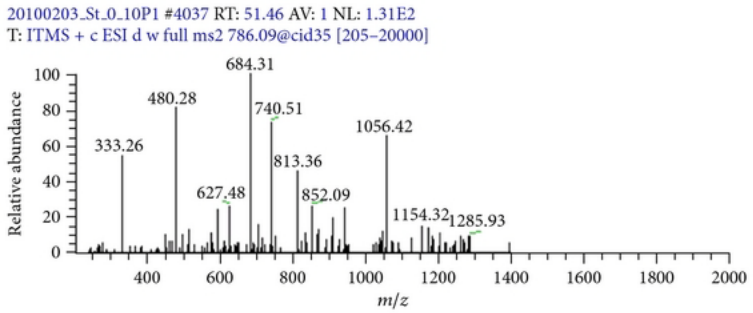


Figure 10-1-c: At the next scan (also shown in red in (a)) , an MS/MS scan is performed, providing peptide fragmentation information for peptide identification.

Once a peptide has been identified, a program can go back to the MS1 scans and find a series of spectra which contain peaks corresponding to the same peptide as it continues to elute from the column. These spectra are then aligned and the intensities of the peaks for the specific m/z value which represents the parent ion of this peptide are plotted against the retention time, giving an extracted ion chromatogram.

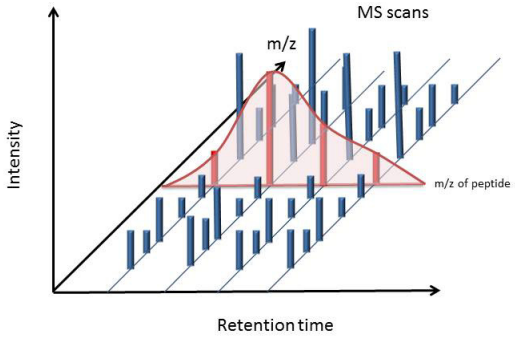


Figure 10-2: The intensities of the MS peaks at the same m/z value are plotted as a function of the retention time. The area under this curve (enclosed in red) is the precursor ion intensity.

In the extracted ion chromatogram, a curve is fit to the intensities at a specific m/z. The area under this curve represents the total amount of the specified peptide that eluted. Scaffold Quant reads these values from its input files and uses them to do quantitative analysis.

Preparing Data for Precursor Intensity Quantitation in Scaffold Quant

Scaffold Quant reads precursor intensity information from various identification programs provided that the user has requested this type of quantitation during the search and that the program used has a way to export the results into an mzIdentML file. Following are instructions for preparing input files for precursor intensity quantitation in Scaffold Quant:

Proteome Discoverer

Proteome Discoverer provides a workflow template for computing precursor intensity values. The template **WF_LTQ_Orbitrap_Sequest_Precursor_ions_Area_Detector** can be used as a starting point, and the search engine choice or instrument settings may be changed.

mzIdentML files exported directly from Proteome Discoverer do not load into Scaffold Quant. Precursor intensity data from Proteome Discoverer results must be analyzed first in Scaffold where the precursor intensities are read from the MSF files. The results can then be exported into an mzIdentML file which is loaded into Scaffold Quant.

Mascot Distiller

Mascot Distiller allows the calculation of precursor intensities but does not provide a way of exporting the results into mzIdentML files. This means that the data, once produced, need to be loaded into Scaffold first to be able to obtain the correct file format for Scaffold Quant. The user can find, hereafter, instructions on how to produce precursor intensities data using Mascot distiller and how to load the data into Scaffold and then exported as mzIdentML files containing precursor intensities for loading into Scaffold Quant.

When setting up a Mascot search the user has to select **Average[MD]** as the quantitation method. When the search is complete, in Distiller selecting **Analysis > Calculate XIC**, and then **Analysis > Quantitate** provides the precursor intensities used for quantitation. The results need then to be exported as an XML file using **Analysis > Quantitative Report > Save as XML**. Creating an ROV file by saving the project with **File > Save Project As….** is also needed**.** At this point the user has to place the ROV file and the XML file in the same directory, and if the DAT file is not accessible directly from the Mascot Server, also place that file in the same location. **The file that needs to be selected for Load into Scaffold is only the XML files.**

Once the data are loaded into Scaffold the menu **Export > mzIdentML...** allows the user to export the experiment to the correct file format for Scaffold Quant.

Spectrum Mill

No special settings are required. Load the entire Spectrum Mill results directory into Scaffold Quant.

MaxQuant

MaxQuant allows the calculation of precursor intensities but does not provide a way of exporting the results into mzIdentML files. This means that the data, once produced, need to be loaded into Scaffold first to be able to obtain the correct file format for Scaffold Quant. The user can find, hereafter, instructions on how to produce precursor intensities data using MaxQuant and how to load the data in Scaffold from where it is possible to create mzIdentML export files containing precursor intensities quantitation data.

MaxQuant 1.3 will only compute precursor intensity when two or more raw files are processed together. Each of the samples to be compared must be labeled with a different experiment name in the experiment.txt file.

Generally, all MaxQuant results in a single directory load into Scaffold as a single sample. For precursor quantitation, however, the samples to be compared must be loaded into different BioSamples. Accordingly, Scaffold has a special dialog that opens when the program recognizes the presence of an experiment file. To place each experiment into its own BioSample, from the loading wizard the user has to select the MaxQuant output directory and click **Add to Import Queue** then when the dialog appears, select the first experiment. Then the user has to click **Next**, then **Add another BioSample** and select the same directory, but choose a different experiment from the dialog box.

In **MaxQuant 1.4**, precursor intensity may be computed even when analyzing a single raw file if the user selects the Label Free Quantitation option. Individual results may then be loaded into separate BioSamples in the usual way and used for Precursor Intensity Quantitation.

If two or more raw files are analyzed together in MaxQuant 1.4 with the Label Free Quantitation option selected, and no **Experiment.txt** file is provided, they form a single combined folder which loads into Scaffold as a single sample. In this case, Scaffold and Scaffold Q+ are unable to perform Precursor Intensity Quantitation. It is possible, although not required, in MaxQuant 1.4 to create an experiment file. The experiments can be named through the MaxQuant 1.4 GUI, and then an experiment file can be exported by right-clicking

and choosing **Export**. The user should name the file **Experiment.txt** and then Scaffold will recognize it and loading can proceed as for MaxQuant 1.3 results.

Once the data are loaded into Scaffold the menu **Export > mzIdentML...** allows the user to export the experiment to the correct file format for Scaffold Quant.

Precursor Intensity Normalization

In Scaffold Quant, precursor intensity quantitative values in attribute groups are computed according to the following algorithm:

1. Within each MS Sample, all Protein Spectra Matches (PSMs) that identify the same ion, where by ion we refer to a peptide sequence with a list of modifications and a charge, are combined by quantitatively characterizing the ion with the maximum precursor intensity across all the considered PSMs for that specific ion.
2. Depending on the experimental organization defined by the user in the organize view, technical replicate (TR) groups might contain one or more MS Samples. For more information on how to define a TR attribute group see . Within each TR group an ion is then quantitatively characterized by summing all the precursor intensity values associated to it in all the MS Samples included in the TR group.
3. When normalization is not applied, the ion intensities within each TR group are then summed and then, if desired, the $\log_{10}$ is applied to the result. This is the value that is shown in the Samples table when Normalization is not selected.
4. When Normalization is selected, the summed intensities from the different MS Samples are first $\log_{10}$ transformed and normalized following the procedure described in Normalization procedure.
5. After the normalization is performed the ion intensities are exponentiated and the ion intensities are then summed for each protein group or cluster within every TR group in the experiment. The results are then $\log_{10}$ transformed if desired and shown in the Samples table.

Normalization procedure

The normalization procedure starts by building a histogram for each technical replicate group (TR group) of the $\log_{10}$ Raw Intensity of the ions included in the group. This is done by considering all the spectra found in the TR group. If a spectrum has no intensity, a value may be imputed to it by the QRILC method (see).

For each histogram, the three primary quartiles, 25%, 50% and 75% are calculated. For each TR group, the quartiles of its histogram are plotted against the quartiles of the entire experiment. A linear function is calculated that connects the 25% and the 50%, and another linear function is computed connecting the 50% and 75% points on this graph.

In each TR group, values below the median are transformed using the first function, and those above the median are transformed with the second.

Missing Value Imputation

Missing values are a major issue in proteomics, complicating quantitative analysis and statistical testing. Values may be missing for a variety of reasons. Missing values may be

random or may be the result of failure to detect very small quantities of a protein. The latter is often the case in DDA proteomics.

Scaffold Quant provides options for imputing missing values (see "Missing Values" on page 164).

Labeled Quantification

Note: Feature available only with a Labeled Quant license

Isobaric Labeling (iTRAQ® or TMT®)

Isobaric labeling is a technique for performing quantitative proteomics that allows multiple samples labeled with different isobaric tags to be mixed and analyzed together in a single MS Sample. The relative quantities of a peptide in the samples can then be determined based on the intensity of the reporter ion corresponding to each tag in its MS2 spectrum. This allows quantification of many samples with fewer LC-MS runs and reduces variation caused by differences between runs. More than one multiplex sample may be analyzed in a single experiment to accommodate more samples, but it is recommended that one or more pooled reference sample be included in each multiplex to aid in normalization between multiplex samples.

Purity Correction

Reporter ion masses sometimes differ from the expected mass because of isotopiic variation. This can cause inaccuracy in quantification based on the observed intensities. For this reason, manufacturers often provide tables showing the percentage adjustments that are needed to correct the observed reporter ion intensities. For iTRAQ®, Sciex recommends using a standard purity correction. For TMT®, Thermo Fisher provides a specific purity correction chart for each batch of reagents.

If data is loaded first into Scaffold and exported as an SFDB file, the purity correction can be applied in Scaffold and the adjusted values will be exported. For MS3 data, this is the only way to apply a purity correction for Scaffold Quant, and it cannot be changed.

If data is loaded from mzIdentML files and the quantitative values are extracted from the MS2 spectra, the purity correction can be entered into Scaffold Quant (see "Purity Correction" on page 49) and the adjusted values will be stored in the SFDB. To change the purity correction, it is necessary to repopulate the quantitative values.

Scaffold Quant has the option to normalize reporter ion intensity values across samples and calculate fold changes at different summarization levels. It also computes quantiative tests to determine the statistical significance of observed differences in protein levels across quantitative samples.

Entering a Purity Correction

ThermoFisher Scientific™ provides purity correction tables for individual lots of TMTpro Reagents. Values from these tables should be entered into Scaffold as follows:

For all except TMTpro 16-plex, choose "**View relative to Monoisotopic concentration**" in the top drop-down and type the percentages for each TMT reagent in the order listed in the Certificate of Analysis.

To enter TMT 16-plex purity corrections:

Enter only the values for which the adjusted masses correspond to other mass tags. These values are indicated in the table by the appearance of the corresponding mass tags in parentheses. These values should be entered into the -1 or +1 columns in the appropriate rows. A sample product data sheet with the values to be entered into Scaffold highlighted and the resulting correctly completed purity correction table in Scaffold are shown in the figures below.

Figure 10-3: Sample Product Data Sheet for TMTpro 16plex

ThermoFisher SCIENTIFIC
The world leader in serving science

PRODUCT DATA SHEET

Thermo Scientific™ TMTpro 16plex Label Reagent Set, 1 x 5mg

Product Number: A44520 **Lot Number:** UL297970

Form: TMTpro reagents are supplied dried, 5 mg/tube. Make a stock solution by reconstituting each tube with 200 µl dry acetonitrile.

Note: ** Reporter ion isotopic distributions (+1, +2) are primarily for natural carbon isotopes with reporter ion interference for each mass tag shown in parentheses. Incomplete stable isotope incorporation (-2, -1) are also reported for carbon and nitrogen isotopes. Reporter ion isotopic distributions can be used as isotope correction factors in TMTpro 16plex method template in Proteome Discoverer software version 2.3 and above.

****Reporter Ion Isotopic Distributions:**

Mass Tag	Reporter Ion Mass	-2		-1		M+	+1		+2	
		-2x 13C	-13C -15N	-13C	-15N	-	+15N	+13C	+15N +13C	
TMTpro-126	126.127726	N/A%	N/A%	N/A%	N/A%	100%	0.34%	9.31% (127C)	0.02%	0.
TMTpro-127N	127.124761	N/A%	N/A%	N/A%	0.71% (126)	100%	N/A%	8.98% (128N)	N/A%	0.
TMTpro-127C	127.131081	N/A%	N/A%	0.75% (126)	N/A%	100%	0.33%	7.98% (128C)	0.02%	0.
TMTpro-128N	128.128116	N/A%	0.00%	0.95% (127N)	0.79%	100%	N/A%	8.38% (129N)	N/A%	0.
TMTpro-128C	128.134436	0.00%	N/A%	1.47% (127C)	N/A%	100%	0.34%	6.91% (129C)	0.00%	0.
TMTpro-129N	129.131471	0.00%	0.00%	1.46% (128N)	1.28%	100%	N/A%	6.86% (130N)	N/A%	0.
TMTpro-129C	129.13779	0.51%	N/A%	2.74% (128C)	N/A%	100%	0.36%	6.15% (130C)	0.00%	0.
TMTpro-130N	130.134825	0.49%	0.02%	2.76% (129N)	0.62%	100%	N/A%	5.98% (131N)	N/A%	0.
TMTpro-130C	130.141145	0.04%	N/A%	3.10% (129C)	N/A%	100%	0.42%	4.82% (131C)	0.02%	0.
TMTpro-131N	131.13818	0.04%	0.04%	3.09% (130N)	1.36%	100%	N/A%	4.75% (132N)	N/A%	0.
TMTpro-131C	131.1445	0.06%	N/A%	3.81% (130C)	N/A%	100%	0.40%	3.29% (132C)	0.03%	0.
TMTpro-132N	132.141535	0.04%	0.00%	2.84% (131N)	0.79%	100%	N/A%	3.51% (133N)	N/A%	0.
TMTpro-132C	132.147855	0.10%	N/A%	4.14% (131C)	N/A%	100%	0.40%	1.80% (133C)	0.02%	0.
TMTpro-133N	133.14489	0.36%	0.01%	3.64% (132N)	0.82%	100%	N/A%	1.94% (134N)	N/A%	0.
TMTpro-133C	133.15121	0.88%	N/A%	4.70% (132C)	N/A%	100%	0.40%	1.01% (134C)	0.00%	N
TMTpro-134N	134.148245	0.40%	0.04%	4.92% (133N)	0.10%	100%	N/A%	1.05% (135N)	N/A%	N

Figure 10-4: Entries in the Purity Correction dialog for TMTpro 16plex

Configure Labeled Quantification

The Experiment menu contains an option to open the Configure Labeled Quantification dialog. The option is only available when a labeled quantification method has been selected as the Display Method; at other times it is grayed and inactive.

The Configure Labeled Quantification dialog allows the user to set various parameters that control how quantitative samples will be normalized.

Normalization of Reporter Ion Intensities

Two levels of normalization are performed in Scaffold Quant when analyzing multiplex experiments based on isobaric labeling. Technical Replicate Level (TRL) normalization primarily compensates for differences between the labeled samples that make up the multiplexes. Internal Reference Scaling (IRS) normalizes between multiplexes to minimize differences that result from combining different MS runs. Either level of normalization may be turned off by un-checking its checkbox in the Configure Labeled Quantification dialog.

TRL Normalization

TRL Normalization of labeled data is carried out just as it is in Precursor Intensity quantification (see Precursor Intensity Normalization) but operating on Quantitative Samples rather than on MS Samples. The distribution is built by first combining fractions (if any are specified) and then using the values from within each technical replicate.

The results of applying TRL normalization can be seen in the Pre/Post Normalization Charts in the Visualize View.

IRS Normalization is based on the Internal Reference Scaling[3] method developed by Phil Wilmarth at Oregon Health and Science University. For more information, see https://pwilmart.github.io/IRS_normalization/understanding_IRS.html.

IRS Normalization

IRS normalization is carried out at the protein level. It works by computing a scaling factor for each multiplex and adjusting the protein levels according to these factors. The scaling factor is computed by:

/

(average of MS Sample total intensities)/(sample total intensity)

The IRS Normalization Chart in the Proteins View provides a visual representation of this calculation. This chart is only displayed when an isobaric labeling technique has been selected in the Quantitative Method drop-down.

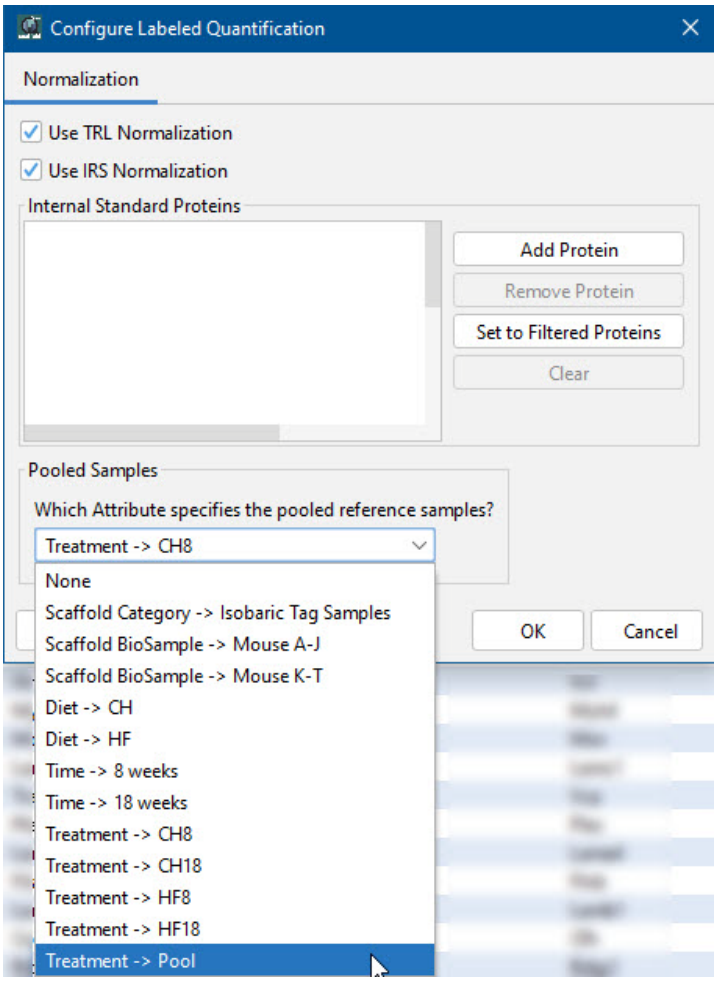
Pooled Reference Samples

When analyzing multiplexed data, Scaffold Quant supports the normalization of quantitative samples based on pooled reference samples included in each multiplex. To perform pooled reference normalization, it is necessary to create a category that includes an attribute that identifies the quantitative samples that represent pooled reference samples. For example, a Treatment category might include attributes such as “Control”, “Treatment 1”, “Treatment 2”, and “Pooled”. Once this has been done, the user may open the **Configure Labeled Quantification** dialog through the **Experiment** menu and select the category and attribute that identify the pooled reference samples. When this has been done, the pooled samples will be used for normalization and will not be available in the Configure Sample Organization and

3. Extended multiplexing of tandem mass tags (TMT) labeling reveals age and high fat diet specific proteome changes in mouse epididymal adipose tissue. Plubell, D.L., Wilmarth, P.A., Zhao, Y., Fenton, A.M., Minnier, J., Reddy, A.P., Klimek, J., Yang, X., David, L.L. and Pamir, N., 2017. Molecular & Cellular Proteomics, 16(5), pp.873-890.

Statistical Analysis dialog for use in comparisons.

Figure 10-5: Identifying Pooled Reference Samples



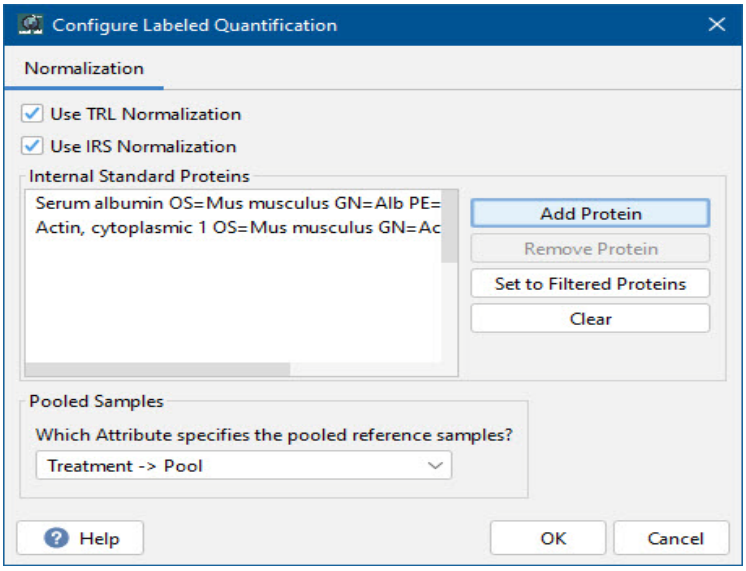
When a set of pooled samples has been designated, these samples are used in calculating the IRS scaling factor. Instead of using the average of all MS Sample total intensities Scaffold Quant uses the average of the total intensities of the pooled reference samples. Since these samples contain all of the proteins they exhibit less variation and provide more effective normalization.

Normalization to Internal Standards

Another option in performing normalization is to designate one or more proteins as internal standards. These may be proteins spiked into the samples in known quantities or proteins that tend to be present at similar levels in all of the samples.

When internal standard proteins are specified, Scaffold Quant calculates the ratios of the average total intensities of these proteins across samples and scales the other proteins using these ratios.

Figure 10-6: Designating Internal Standard Proteins



Quantitative Tests

Scaffold Quant provides a variety of statistical tests to identify proteins that show different quantitative abundances at any previously established level of summarization. The experimental design and the number of replicates dictates the most appropriate test to use.

The tests are available for selection through the Configure Sample Organization and Statistical Analysis dialog.

The tests are based upon the data that is being displayed in the Samples Table. Adjusting thresholds and filtering the data changes the number of proteins shown in the table and the tests may select different proteins as having abundance level changes.

Configure Sample Organization and Statistical Analysis dialog

Selecting the menu option ***Experiment > Quantitative Analysis*** opens the **Configure Quantitative Analysis** dialog. This dialog consists of two tabs: the **Sample Hierarchy Tab** which allows the user to specify the experimental design (For details, see "Specifying the Design of an Experiment" on page 66) and the Statistical Analysis Tab (see "Statistical Analysis Tab" on page 76) from which the user may choose a quantitative test to apply to the data appearing in the Samples Table and to define the significance level for the selected test.

The user must first specify the type of experiment and assign categories their appropriate roles in the analysis through the Sample Hierarchy Tab, and then may select and configure a statistical test through the Statistical Analysis Tab.

Not all factor levels or treatments need to be used for a specific test; only factor levels with a selected check box in The Configure Sample Organization and Statistical Analysis Dialog are used in computing the test. This can be useful if the user wants to exclude one or more treatments from the quantitative analysis. Sometimes this may be necessary in order to satisfy the constraints of the test. For example, the Two-Way and Repeated Measures tests require that the experiment be balanced. If one group has fewer samples than the others, it may be necessary to exclude that group from the analysis.

The user may also choose to apply a Multiple Test Correction, and must specify the desired significance level.

Once the experimental design has been specified, the quantitative test chosen, any multiple comparison correction has been selected and the significance threshold has been specified, clicking **Apply** starts the calculation. When the calculation is complete, a new column appears in the Samples Table showing the results of the selected quantitative test.

The heading of the added column lists the type of test applied and the comparison levels utilized. The p-values or q-values shown in the added column are highlighted in green if they are significant even with any multiple test correction, or yellow if they are significant only without the error correction applied. Values which are not significant under either condition are not highlighted.

Figure 10-7: Display of Statistical Test Results

When the selected summarization level corresponds to the chosen biological sample level, or blocking level, for the test the blocking variables involved in the test and belonging to the same treatment, or combination of factors levels, are tagged with a colored band. This helps the user recognize the groups of experimental units blocked together in the current test

Tests available for analyzing experiments of the Basic Design type

ANOVA/t-test

ANOVA (Analysis of Variance) is an analysis method for testing equality of means across treatment groups or categories. It tells if there are differences among categories. The result of the test is a p-value which ,when low, indicates a large probability of variation among the different categories considered for the test. It does not, however, indicate which of the categories are different from each other.

Scaffold Quant supports a two-tailed version of ANOVA. When only two treatments are selected from the combinations of factor levels available, the two tailed ANOVA is equivalent to a T-Test.

Permutation Test

This test, depending on the selected treatments used to perform the test, establishes if there are statistically meaningful differences among multiple groups (equivalent to ANOVA) or differences between just two groups (equivalent to a T-test). Rather than depending on assumptions about the distributions of the values, however, it performs the experiment of randomly assigning the observed values to the various categories and assessing how rarely the degree of difference between categories in the experiment is observed.

It is based on an F-statistic calculated on the original set of data points; the data points are then randomly permuted and a new F-statistics is calculated using the randomized values. This randomization and computation of an F-statistic is repeated 10000 times. Finally, a p-value is calculated by counting the number of times the randomized F-statistics were at least as large as the original F-statistics and dividing by 10000.

Fisher’s Exact Test

The Fisher’s exact test is a statistical significance test used to compare counts for a selected protein in two different factor levels or treatments against the background of the remaining proteins in the Protein List. It more effectively calculates statistical significance than a T-test when the number of replicates in the factor levels is low or when the number of spectra assigned to each protein in each factor level is low.

Scaffold Quant uses the two-tailed version of the test. In the Configure Quantitative Analysis dialog after selecting the Fisher Test option, a list of possible factor levels or treatments appears in accordance to the selected Summarization level. Only two factor levels within the list can be checked at the same time and by default Scaffold Quant selects the first two factor levels appearing the in the list.

The Fisher Exact Test cannot be applied if the **Normalized** box is checked in the Table Tab

Display pane. The test is exact and applies its own internal normalization. To make sure that normalization is not accidentally selected after the test is applied, the Normalized check box is grayed out.

The elements in the contingency table are determined as follows: the first row contains the protein sample counts and the second row contains the spectral count minus the total spectral count for the sample (see example below). If proteins share spectra, the Weighted Spectrum Count should be used to validate the calculation.

As an example, let's consider an experiment with a list of four proteins and create the contingency table for one of the proteins in the list.

(Weighted Spectrum Counts)	Factor Level 1	Factor Level 2
Protein A	**11**	**18**
Protein B	9	33
Protein C	9	28
Protein D	10	42
Total Count	39	121

To compute the Fisher's Exact test p-value the following 2x2 contingency table is formed:

	Factor Level 1	Factor Level 2	*Row Total*
Protein A	11	18	29
Background proteins	28 (39-11)	103 (121-18)	131
Column Total	39	121	160

where the 28 and 103 are chosen so that the column-sums are the totals 39 and 121.

To compute the p-value, Scaffold Quant generates the hyper-geometric distribution with parameters N = 160 (=39+121) the total sum for the columns and rows in the contingency table, K = 39, and n = 29 = 11+18. Then to get the two-tailed p-value, Scaffold Quant sums all probabilities from the distribution which are less than or equal to the probability occurring when k = 11. (See e.g. h*ttp://mathworld.wolfram.com/FishersExactTest.html*)

The first use of the Fisher's exact test with MS/MS spectrum counting results was discussed in Zhang et al, J. Proteome Res., 2006, 5 (11), pp 2909?2918. Please consult this article for more details on how the test is calculated.

Mann-Whitney U Test

The Mann Whitney test is a non-parametric statistical hypothesis test for assessing whether one of two samples of independent observations tends to have larger values than the other. It can also be defined as a distribution-free test of whether two medians are equal. The test uses the ranks of the data in the two samples. Although the Mann Whitney test compares well with a t-test, it is independent of the way the data is distributed. Because the Mann Whitney test is the non-parametric version of the t-test, it requires exactly two quantitative sample categories to be selected for testing.

Kruskal-Wallis Test

The Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing more three or more samples that are independent, or not related. (The parametric equivalence of the Kruskal-Wallis test is the one-way analysis of variance (ANOVA)). The factual null hypothesis is that the populations from which the samples originate have the same median. When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur. Because it is a non-parametric method, the Kruskal-Wallis test does not assume a normal distribution (unlike the analogous one-way analysis of variance); however, the test does assume an identically-shaped and scaled distribution for each group, except for any difference in medians.

Tests available for analyzing Repeated Measures experiments

Repeated Measures ANOVA/Paired t-test

The Repeated Measures Analysis of Variance test (rANOVA) is a parametric statistical hypothesis test for assessing whether the population means of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). The Paired t-test is similar, but compares exactly two samples. These tests may be used when the data being analyzed is normally distributed and has equal variances across the categories.

Wilcoxon Signed-rank Test

The Wilcoxon Signed-rank test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements have been taken under exactly two conditions (e.g., time points). It is a nonparametric alternative to the Paired T-Test, and may be used when the data being analyzed is not normally distributed. The Wilcoxon Signed-rank test does assume that the distributions in the two categories are independent and identically distributed.

Friedman Test

The Friedman test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). It is a nonparametric alternative to the Repeated Measures Analysis of Variance (rANOVA), and may be used when the data being analyzed is not normally distributed. The Friedman test does assume that the distributions in the categories are independent and identically distributed.

Tests available for analyzing Two-Way experiments

Two-way ANOVA

The Two-way ANOVA assesses the effect of two independent variables on protein levels. It measures the main effect of each of the independent variables, as well as whether there is significant interaction between the variables.

Randomized Block ANOVA

This test is applicable to experiments with a Randomized Block Design. It assesses the effect of the Primary Analysis Category while controlling for the effect of the Secondary Analysis Category, which in this case is the blocking factor. It does not, however, assess the effect of the Secondary Analysis Category itself. An option is available, however, to assess the blocking effect, which assesses whether there is significant interaction between the Primary and Secondary Analysis Categories.

Friedman Test

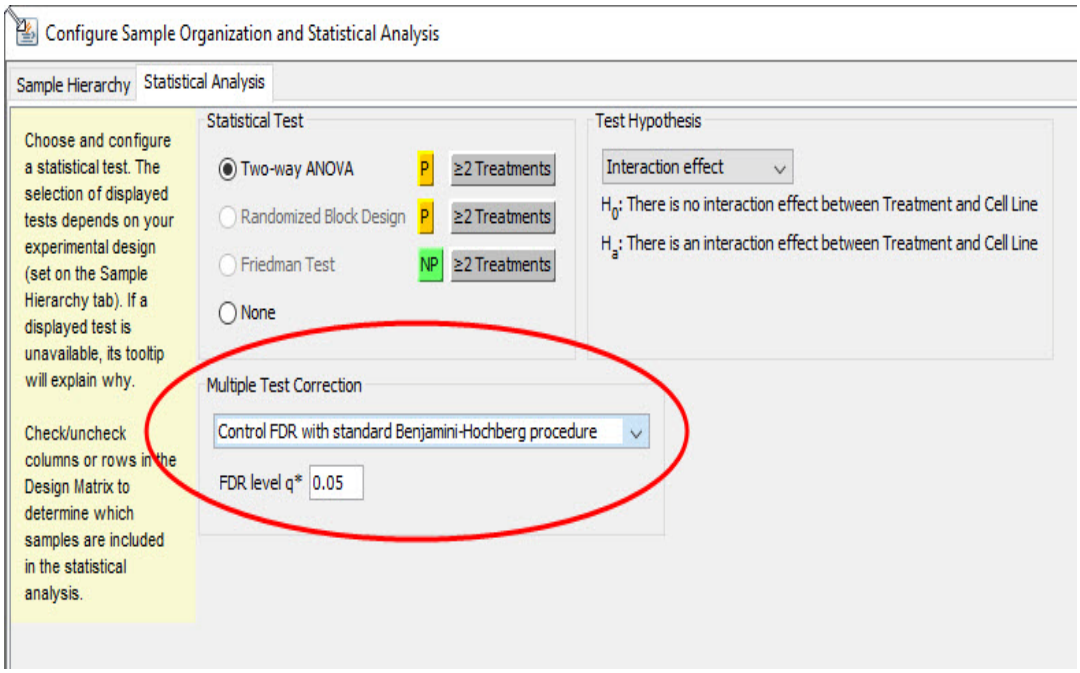
The Friedman test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). It is a nonparametric alternative to the Repeated Measures Analysis of Variance (rANOVA), and may be used when the data being analyzed is not normally distributed. The Friedman test does assume that the distributions in the categories are independent and identically distributed.

- *In choosing the statistical test to apply, it may be helpful to remember that log intensity values are more likely to be normally distributed while the intensities themselves are not. Parametric tests, therefore, are more suitable when analyzing log values, while it may be preferable to select a non-parametric test for intensities. Nonparametric tests are also better when the data contains outliers which may skew the results.*

Significance Level

Controls in the Statistical Analysis tab allow the user to set the significance level required for the selected inference test and to choose methods to control the family-wise error rate through a pull down list.

Figure 10-8: Significance Level tab



Multiple Test Correction

When considering a set of statistical inferences simultaneously and doing multiple comparisons, the risk of making one or more false discoveries (Type I error) grows quite quickly. In these cases it is common to adjust p-values for the number of hypothesis tests performed. A common method is to control the family-wise error rate, which is defined as the probability of making Type I errors. One of the initial and still quite common methods used to control this error is the Bonferroni correction where the significance level α for an individual test is found by dividing the family-wise error rate (usually 0.05) by the number of performed tests. Thus when doing 100 statistical tests, the α level for an individual test would be 0.05/100=0.0005, and only individual tests with P<0.0005 would be considered significant.

The Bonferroni approach is a fairly conservative one and for a very large number of independent comparisons it may lead to a high rate of false negatives.

To address this issue Scaffold Quant provides two different types of corrections:

- Control FWER with Hochberg's step-up and Holm's step-down
- Control FDR with standard Benjamini-Hochberg procedure

Control FWER with Hochberg's step-up and Holm's step-down

There are various methods described in the literature that control the Family-wise error rate (FWER) using less conservative corrections that are still based on the Bonferroni inequality. These methods are usually quite appropriate to control the FWER in experiments in which a limited number of comparisons are of interest and where the use of the False Discovery Rate is inappropriate. In these cases, such corrections guard against false positives being reported.

Scaffold Quant offers an option to use the following methods to calculate the corrected significance level:

- Holm's step-down method
- Hochberg's step-up method

For more information about these methods, see Techniques to Control the Family-wise Error Rate.

When this option is selected the significance level is expressed in terms of α o and the related text box appears underneath the pull down list. This text box allows the user to set the significance level α to the desired value. The default value is 0.05.

Control FDR with standard Benjamini-Hochberg procedure

This method of controlling the error rate is particularly useful in high-dimensional experiments where a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate (FDR) is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Scaffold Quant computes the FDR using the Benjamini-Hochberg procedure as developed in the original paper[4].

When this option is selected the significance level is expressed in terms of the FDR level q* and the related text box appears underneath the pull down list. This text box allows the user to set the FDR level q* to the desired value. The default value is 0.05.

Test Hypothesis

This section displays a statement of the null hypothesis, H_0 and the alternative hypothesis, H_a that will be tested by the selected test.

In the case of a two-factor analysis, several different measures are computed. This section then contains a drop-down menu for selecting which test measure should be displayed. Options are:

- Interaction Effect - measures whether there is a statistically significant interaction between the two variables.

4. Benjamini Y and Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, Series B (Methodological), Vol.57, No. 1: 289-300.

- Primary Factor Effect - measures whether the variable designated as the primary analysis category demonstrates a statistically significant effect on protein level when controlling for the secondary variable.
- Secondary Factor Effect - measures whether the variable designated as the secondary analysis category demonstrates a statistically significant effect on protein level when controlling for the primary variable.

Design Matrix

The Design Matrix at the bottom of the Statistical Analysis tab is provided to help the user verify the test design or adjust it by adjusting which cells of the matrix should be included in the test. At the beginning of each row and column is a checkbox. If the box is checked, the cells in that row or column will be included in the test. If it is unchecked, they will not. This feature allows the user to, e.g. remove all but two columns to allow performance of a T-test and creation of a Volcano Plot or remove a group that contains an inconsistent number of samples to create a balanced experimental design for a two-way ANOVA.

Multiple Test Corrections

When considering a set of statistical inferences simultaneously and doing multiple comparisons the risk of making one or more false discoveries or a Type I error grows quite quickly. In these cases it is common to adjust p-values for the number of hypothesis tests performed. There are many different methods that provide a way to perform this adjustment. A common one is to control the family-wise error rate, which is defined as the probability of making Type I errors. One of the initial and still quite common methods used to control this error is provided by the Bonferroni correction where the significance level α for an individual test is found by dividing the family-wise error rate (usually 0.05) by the number of performed tests. Thus when doing 100 statistical tests, the α level for an individual test would be 0.05/100=0.0005, and only individual tests with P<0.0005 would be considered significant.

The Bonferroni approach is a fairly conservative one and for a very large number of independent comparisons it may lead to a high rate of false negatives.

To address this issue Scaffold Quant provides two different types of corrections:

- Control FWER with Hochberg’s step-up and Holm’s step-down
- Control FDR with standard Benjamini-Hochberg procedure

Control FWER with Hochberg’s step-up and Holm’s step-down

There are various methods described in the literature that control the Family-wise error rate (FWER) using less conservative corrections than the Bonferroni one but are still based on the Bonferroni inequality. These methods are usually quite appropriate to control the FWER in control trial experiments in which a limited number of comparisons are of interest and where the use of the False Discovery Rate is inappropriate. In these cases these type of corrections

guard against any false positive occurring

When this option is selected Scaffold Quant uses the following methods to calculate the corrected significance level:

- Holm's step-down method
- Hochberg's step-up method

For more information on how the two methods are developed in Scaffold Quant go to the appendix Techniques to Control the Family-wise Error Rate.

When this option is selected the significance level is expressed in terms of α o and the related text box appears underneath the pull down list. This text box allows the User to set the significance level α to the desired value. The default value is 0.05.

Control FDR with standard Benjamini-Hochberg procedure

This method of controlling the error rate in multiple experiments is particularly useful in high-dimensional type of experiments where a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate (FDR) is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Scaffold Quant computes the FDR using the Benjamini-Hochberg procedure as developed in the original paper[5].

When this option is selected the significance level is expressed in terms of the FDR level q* and the related text box appears underneath the pull down list. This text box allows the User to set the FDR level q* to the desired value. The default value is 0.05.

5. Benjamini Y and Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," Journal of the Royal Statistical Society, Series B (Methodological), Vol.57, No. 1: 289-300.

Chapter 11
Reports

A variety of reports are available in Scaffold Quant to assist the user in interpreting and working with quantitative analysis data. The Current View may be exported using the right click context menu option Export> Export to Excel. It creates a CSVfile corresponding to the information displayed in the table or graph from which it is selected. A variety of programmed reports are also available from the Export option on the Scaffold Quant main menu. Each report is saved in a CSVformat.

The user cannot change the report format, but may select a different location in which to save the report. When the user saves an Excel report, a default name in the format <Report Name><Scaffold Quant File name> is provided for the report, but the name and location may be changed in the file browser. Finally, the user can open and view any Scaffold Quant report in Excel or another spreadsheet application, or using a text editor.

The following reports are available in Scaffold Quant:

- Export Current View to Excel
- Export Attributes File...
- Export Samples Report to Excel...
- Export Peptide Report to Excel...
- Export Statistical Report to Excel...
- Export Publish Report to Excel...
- Export Heatmap Report to Excel...
- Export Spectra Report...
- Export Spectra Report by Protein Group...
- Run SQL Query for Export...

Export Current View to Excel

This is accomplished by right-clicking in any pane of any View and choosing Export>Export to Excel. Exports the information contained in the current view to a comma-delimited text file that can be opened and viewed in Excel.

Export Attributes File...

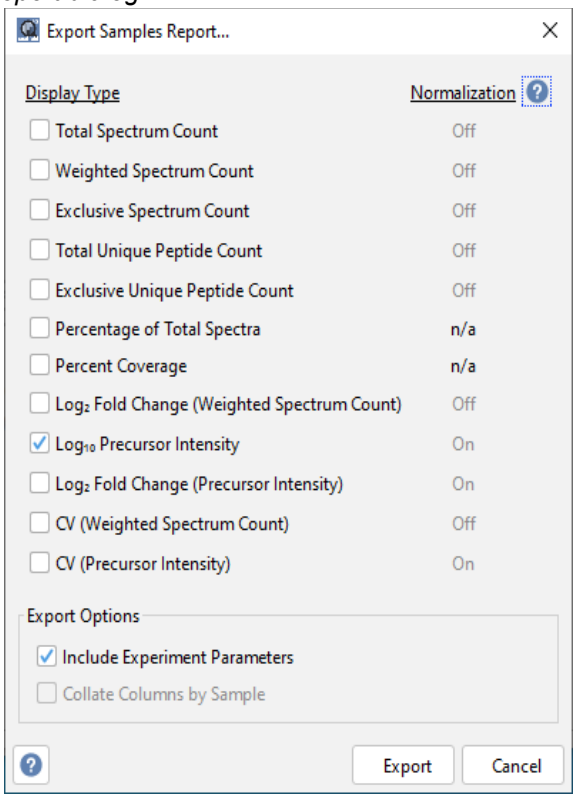
Generates an attributes file that captures the sample organization of the current experiment.

Export Samples Report to Excel...

Opens a dialog that allows the user to produce a customized protein-level report.This option generates a

comma-delimited Samples table similar to the one appearing in the Samples View, but allows the user to select whether or not to display each of the Display Types, whether to include a header specifying the experimental analysis parameters, and if more than one Display Type is selected, how to group the quantitative values. If “Collate Columns by Sample” is selected, the columns containing all quantitative values for a sample will be adjacent to each other, if it is not, all columns of a single Display Type for all samples will be adjacent, followed by columns for the next Display Type, etc.

Figure 11-1: Samples Report dialog



Export Peptide Report to Excel...

Generates a comma-delimited Peptide table for all proteins appearing in the Samples View.

Export Statistical Report to Excel...

Generates a comma-delimited file which presents the details of the currently applied statistical test. If no test has been applied, the report is empty. If a test results column appears in the Samples View, the report presents a section for each protein showing the specific values that went into the test calculation and the results.

For instance, for a basic ANOVA test, the details for one protein are as follows:

Data for Group of Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1+6							
Stage 1->3913	Stage 1->A01I	Stage 2->A00N	Stage 2->A01W	Stage 3->A00C	Stage 3->A036	Stage 4->4007	Stage 4->A016
38.696	55.893	34.223	31.656	23.898	23.125	12.466	12.439
Result for Group of Actin, cytoplasmic 1 OS=Homo sapiens GN=ACTB PE=1 SV=1+6							
Source	SS	df	MS	F	Sig. of F		
Between Groups	1,308.30	3	436.11	11.517	0.019475		
Within Groups	151.47	4	37.867				
Total	1,459.80	7					

Export Heatmap Report to Excel...

Generates a comma-delimited file containing a header section detailing the parameter settings and filtering applied when the Heatmap was generated, along with the information contained in the Heatmap.

Export Publish Report to Excel...

Generates a comma-delimited report of the information contained in the Publish View

Export Spectra Report...

Generates a comma-delimited file containing a header section detailing the parameter settings and filtering applied, followed by a table corresponding to the information in the The Validation Pane for each peptide in the experiment.

Export Spectra Report by Protein Group...

Generates a comma-delimited file containing a header section detailing the parameter settings and filtering applied, followed by a table corresponding to the information in the The Validation Pane along with the protein accession number for each peptide for each protein in the experiment. Shared peptides are repeated, appearing with each protein to which they are assigned.

Run SQL Query for Export...

Opens the SQL Query tab of the Publish View, see SQL Export tab.

Chapter 11
Reports

Appendix

Appendix A.Computation of protein and peptide FDR in Scaffold Quant

The User specifies FDR thresholds through the Thresholding dialog box as described in Scaffold Quant uses the following approach in applying these thresholds:

All peptide spectrum matches in the experiment are listed in descending order by their primary scores. Some of these matches are target peptides, while others are decoy peptides. Starting from the top of the list, for each row, Scaffold Quant calculates the current peptide FDR by dividing the number of decoy peptides in the list so far by the number of target peptides so far. This process continues until the entire list has been processed. The program then finds the last row in which the calculated FDR is at or below the selected Peptide FDR Threshold. All PSMs above that point are considered to meet the threshold. Those which fall below the cutoff are considered to be invalid and are excluded from the experiment.

This procedure has the effect of producing the largest set of peptides possible while maintaining a peptide FDR at or below the threshold. Next, the program extracts the list of proteins that are associated with peptides in the thresholded list.

This list of proteins is then filtered to include only those which are identified by at least the specified Minimum Number of Peptides. The protein list is sorted in descending order by protein score (or probability). The Protein FDR (the number of decoy proteins in the list so far divided by the number of target proteins seen so far) is calculated for each row. The last row in which the calculated Protein FDR is at or below the specified threshold is chosen as the cutoff point.

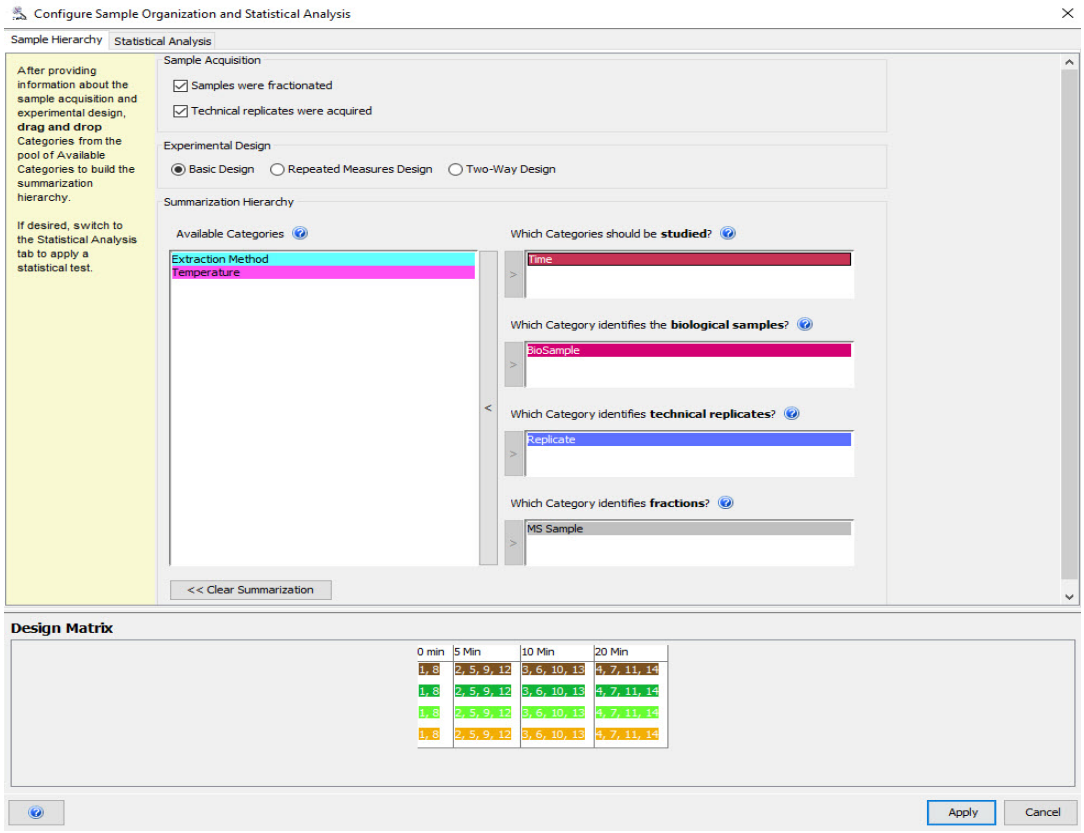
Once again, the effect is to produce the largest possible set of proteins that meets the specified Protein FDR Threshold. This set becomes the thresholded protein list, and only the “valid” peptides are included for each protein.

Any PSMs that were not associated with any protein in the thresholded protein list is then eliminated from the experiment. Because some peptides that passed threshold are not included in the proteins that pass threshold, Scaffold Quant recalculates the peptide FDR value from the final thresholded set. The results are displayed in the FDR Information Box. Note that occasionally this procedure might result in a peptide FDR that is slightly higher than the selected Peptide FDR Threshold.

Appendix B.Rolling Up Intensity Values

When precursor intensity data is loaded into Scaffold Quant, the **Precursor Intensity** and **Log2 Fold change (Precursor Intensity)** types become available for selection in the Display Type pull-down list in the Samples View.

Figure 1: When the Display Type **Precursor Intensity** is chosen and the level of summarization is set to any level above the MS Sample level, Scaffold Quant rolls up the values listed in a row, or group of proteins, to a higher level of summarization in one of two distinct ways, depending on whether the particular level of summarization is below the technical replicate level (fractions) or is at or above the technical replicate level, see Figure 1.*Summarization with fractionation and technical replicates*

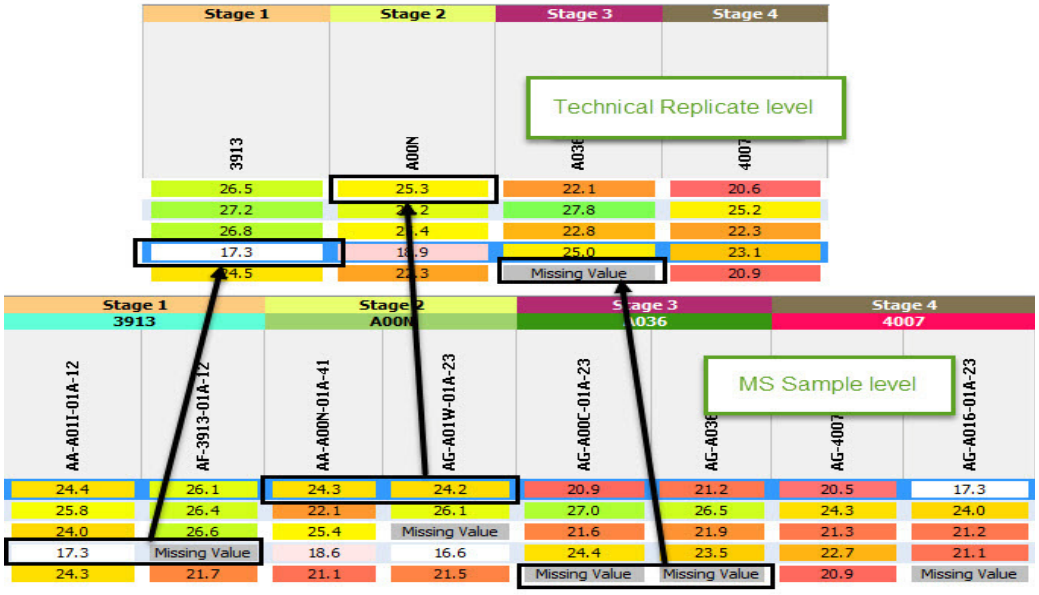


At the MS sample level, for every protein, Scaffold Quant reports the sum of the maximum precursor intensity values for each ion peak included in an MS sample. When a technical replicate summarization group includes more than one MS sample, Scaffold Quant rolls up the values from the MS sample level to the higher technical replicate summarization level by simply summing all of the precursor intensity values in the group.

When an MS sample does not include data for a particular protein but that protein is found in another sample, Scaffold Quant labels the corresponding cells in the Samples table with the tag “Missing Value”. When rolling up to a technical replicate level that includes some missing values, the program ignores the missing values and assigns a value which is the sum of the existing intensities. However, if all of the values are missing, “Missing Value” is assigned to the group as shown in .Figure 2.

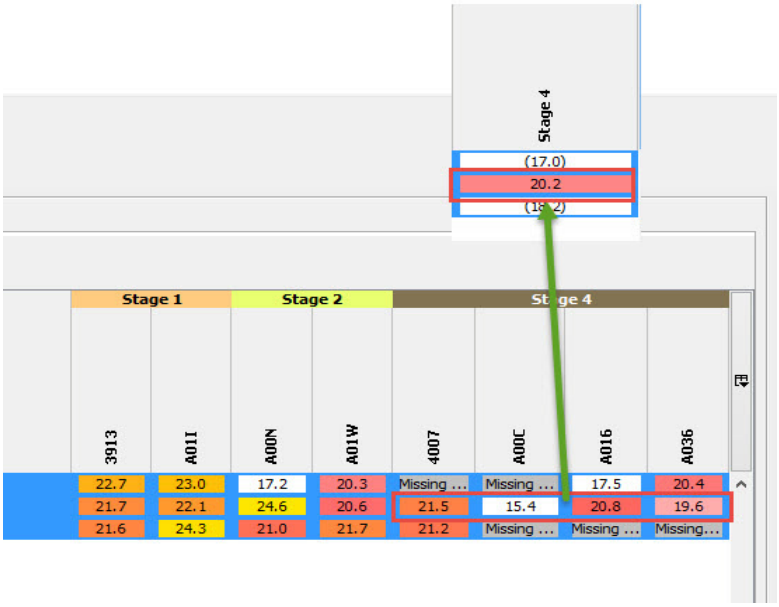
Note: This configuration is recommended when working with MS samples that are fractions of a MUDPIT experiment.

Figure 2: Rolling up to the Technical Replicate Level



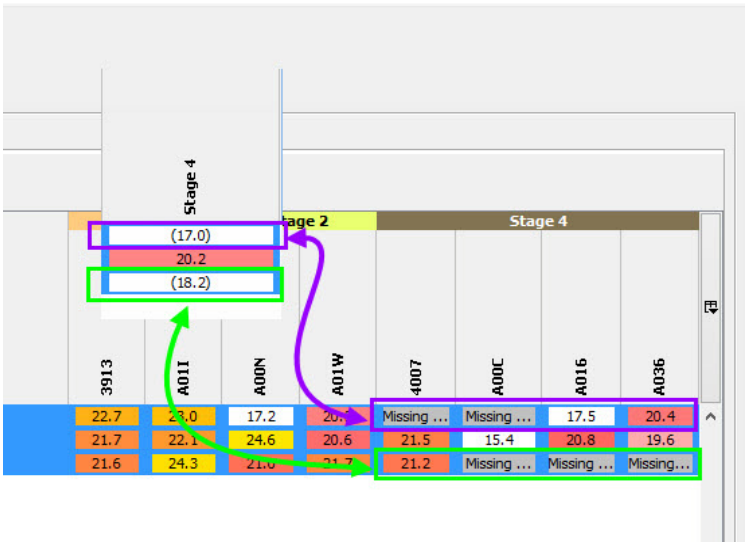
When the summarization level is switched to a level higher than the technical replicate level, Scaffold Quant rolls up the values using the median of the Intensity values of the technical replicates in an attribute group, see Figure 3 The picture shows the rolled up value of the $\log_{10}$ Precursor Intensity values of four biosamples that make up the attribute group Stage 4 and the $\log_{10}$ Precursor Intensity when the higher summarization level Stage is selected.The value corresponds to the median of the values appearing when the next lower level of summarization is chosen.

Figure 3: Rolling up values to a higher summarization level



If fifty percent or more of the values belonging to a group in a specific row are missing, a median cannot be calculated. In this case, the user has the option of choosing to have the missing values imputed and the imputed values used in the calculation of the rolled up value. When this occurs, the values so calculated are shown in parenthesis in the Samples Table, see Figure 4.

Figure 4: Rolling up of values to a higher summarization level with missing values



Missing Values

Missing values affect various computations in Scaffold Quant. In order to be able to roll up values to higher levels of summarization, to perform statistical testing and to perform Principle Component Analysis, it may be desirable to impute values when no measurement has been obtained. On the assumption that missing values in Scaffold Quant are generally a result of either absence of the peptide in a sample or presence at very low intensity, Scaffold Quant offers the option to use the method of "Quantile Regression for Imputation of Left-Censored data" (QRILC)[1].

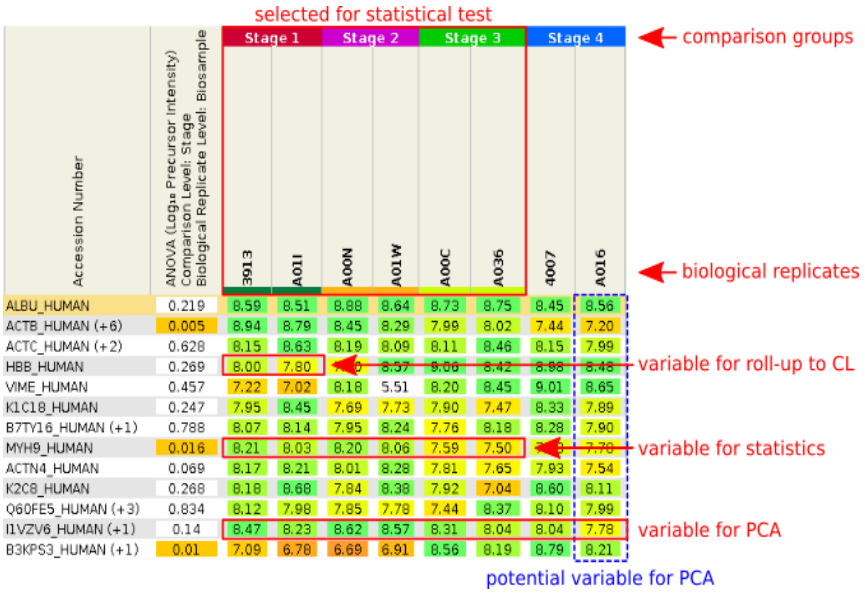
This option is selected through the menu item Experiment>Missing Value Imputation. **QRILC** is the default selection, but imputation may be turned off by selecting **None**.

QRILC imputes values by drawing from a truncated normal distribution whose parameters are estimated from the observed (non-missing) values and the number of missing and observed values, assuming that missing values are lower than any observed value.

When rolling up values where over 50% are missing, the mean estimated by QRILC is used as the rolled up value (which is treated as partially-missing, indicated by surrounding the value in parentheses).

When computing statistics where any number of values are missing (provided there are at least two observed values), QRILC is used to estimate a distribution for values across all comparison groups, and replaces missing values with values drawn from the estimated distribution, truncated at the proportion of missing values (so that if X% of values are missing, the imputed values fall below the X^{th} percentile of the estimated distribution).

PCA does the same, but across values for all biological replicates in the file.



Parameter estimation is done by a linear regression between the quantiles of the observed data and quantiles of a truncated normal distribution (if X% of values are missing, the distribution used for fitting is

1. Cosmin Lazar (2015): imputeLCMD v2.0". R package version 2.0

truncated below the X^{th} percentile, so that the 0^{th} percentile of the observed data is matched to the X^{th} percentile of the distribution, and so on up to the 95^{th} percentile). This gives an estimated mean and standard deviation for the distribution of values. For more information, see the documentation for the R package "imputeLCMD v2.0" which provides the reference implementation for QRILC (https://www.rdocumentation.org/packages/imputeLCMD/versions/2.0/topics/impute.QRILC).

Appendix C.Shared Evidence Clustering Algorithm

Level of sharing $L(A, B)$

The algorithm computes for each pair of proteins (A, B) a value that expresses the level of sharing of peptides between two proteins among all the MS-Samples included in the experiment. The Shared evidence value L is defined as follows:

$$L(A, B) = \frac{\left[\sum_{MS-Sample} \text{\# Shared peptides in A and B in MS-Sample}\right]}{\left[\sum_{MS-Sample} \text{\# peptides in either A or B in MS-Sample}\right]}$$

Notice that:

$$0 \leq L(A, B) \leq 1$$

and that

$L(A, B) = 0$

if and only if A & B do not share any peptide.

$L(A, B) = 1$

if and only if A & B have exact matching peptides in all MS-samples.

Level of clustering λ

Let's fix a number λ in the range

$0 < \lambda \leq 1$

and then consider A and B similar at level λ if

$L(A, B) \geq \lambda$.

For any λ, we may cluster at level λ by joining into a cluster all proteins which are similar at level λ, or may be joined by a chain of proteins that are pairwise similar at level λ.

Currently Scaffold Quant offers

- Perfect Clustering, which corresponds to protein grouping

  $\lambda = 1$

- Moderate Clustering

  $\lambda = 1/3$

- any evidence clustering

  $\lambda = 0$

Appendix D.Distance Based Clustering

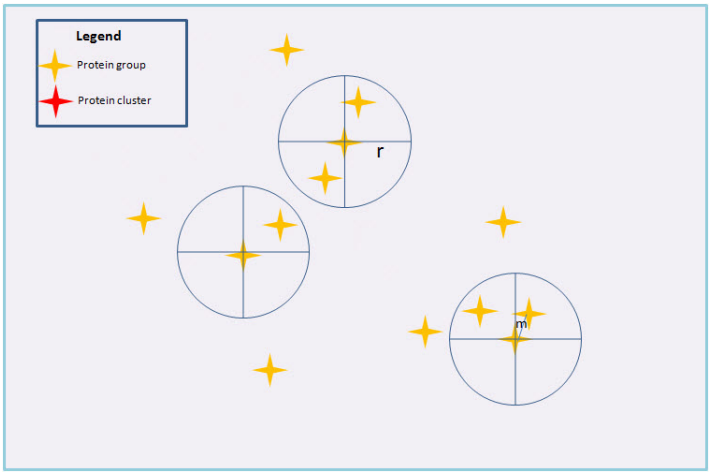
This appendix contains a section that describes the general algorithm, Distance Based Clustering Algorithm and a section that points out the specifics of the implementation for GO Term clustering, GO Term Clustering distance metric.

Distance Based Clustering Algorithm

The algorithm relies on there being the notion of a distance *d(A,B)* between every pair *A* and *B* where *A, B* can be either protein groups or clusters and each of them is characterized by a series of numbers represented included in a multiple dimensional vector. The main difference between Expression Profile Clustering and GO Term Clustering is given by their distinct definition of distance.

The algorithm works recursively, at each stage there is a pool of proteins/clusters to draw from, and some of these are selected and joined together to form a new cluster. At this point the old clusters and protein groups are removed from the pool, the new cluster added in their place and a new minimum distance is also calculated within the new pool.

Figure 5: Initial pool of protein groups



Description of the distance based clustering algorithm:

1. Forming the initial pool. This is composed of protein groups present in the Samples Table.
2. Defining a maximum radius r for cluster formation based on the following formula:

$$r = m + (D - m) \cdot (p + 0.05)$$

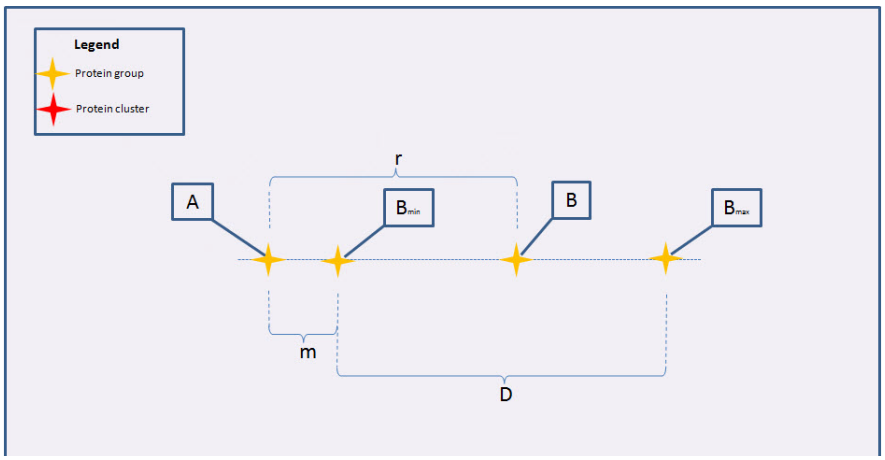
where

- m = the current calculated minimum distance of any two members of the pool,
- D = the maximum possible distance (depends on the distance metric)

- $p = m/D$ = a measure of the percent of the maximum possible distance.

Note that the value 0.05 added to p helps avoiding the formation of clusters of protein groups or clusters characterized by identical vectors.

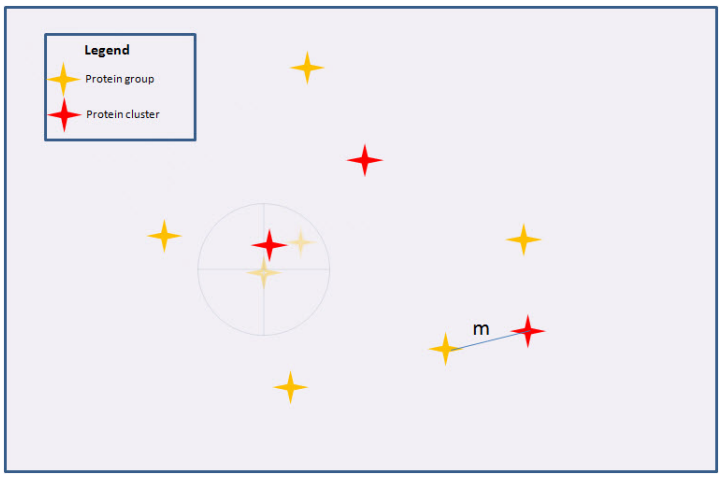
Figure 6: Definition of the maximum radius r for cluster creation.



3. Selecting at random one of the members achieving the minimum distance to some other member, and joining all the members of the pool that are within radius r of this member into a new cluster.

4. Creating a new cluster with these members, and removing them from the pool, replacing with the new cluster

Figure 7: Formation of clusters



5. Recalculating the new minimum distance m and checking if it exceeds 10% of the maximum possible distance D. If so, stop. Otherwise, go back to Step 2.

GO Term Clustering distance metric

The GO Term Clustering uses the Distance Based Clustering as described in Distance Based Clustering Algorithm and the following distance metric.

Each protein group or cluster, when the GO annotations have been applied, is characterized by a vector of dimensions n dependent on the number of GO terms columns shown in the Samples Table. This number of columns is determined by the selection of displayed GO terms performed in the The Displayed GO Terms Tab. Each component of the vector will record if a particular GO term is present or not for a protein group or cluster.

Let's A be a protein or a cluster shown in the samples table containing *n* GO terms columns.

The i_{th} component of the GO vector $G(A)_i$, where i=1,2,...n, is given by:

$$\begin{cases} G(A)_i = 1 & \text{if A has the i-th GO term} \\ G(A)_i = 0 & \text{if A does not have the i-th GO term} \end{cases}$$

For example G(A) = (0,1,0,1,1,1,0,1,1,0,0,0).

The distance is then defined for any pair of protein groups/clusters A, B, by the following equation:

$$d_{euc}(A,B)^2 = (G(A)_1 - G(B)_1)^2 + ... + (G(A)_n - G(B)_n)^2$$

Since

$$0 \le (G(A)_i - G(B)_i)^2 \le 1$$

Notice that the maximum distance in is D = n.

Since

$$0 \le (G(A)_i - G(B)_i)^2 \le 1$$

Notice that the maximum distance $d_{eucl}(AB)^2 = n$.

Appendix E.Weighted spectrum counts

When forming proteins, peptides belonging to a protein might be associated only with a single protein, referred to as exclusive unique peptides, or shared with other proteins. Peptides and spectra belonging to a protein may be counted differently depending on how the shared peptides are treated.

In Scaffold Quant, shared peptides are apportioned to the different proteins to which they belong according to a criterion that looks at the abundance of exclusive unique peptides in each of the proteins that share the peptide. In each protein, a weight proportional to the number of exclusive unique peptides associated with that protein is assigned to each shared peptide, see examples shown in Figure 8 and Figure 9.

Figure 8: Shared peptides between two proteins

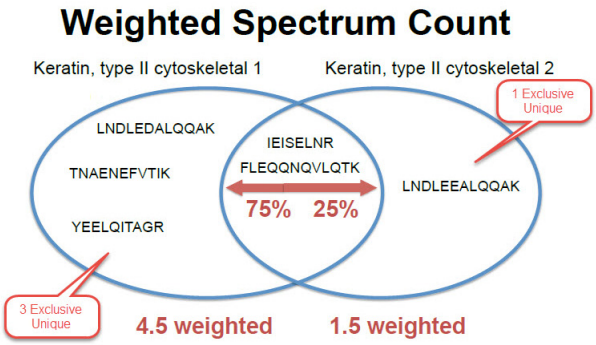
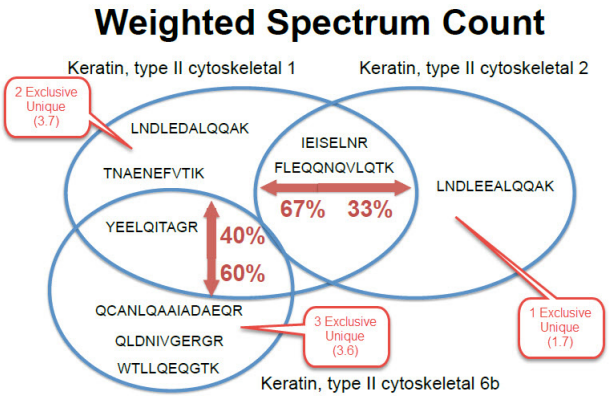


Figure 9: Shared peptides among three proteins



A peptide might be associated with one or more spectra; the number of unique spectra for a shared peptide is then apportioned to the proteins that share it according to its weight as defined above.

For example, in Figure 8 protein “Keratin type II cytoskeletal 1” has three exclusive unique peptides and two shared peptides with Keratin type II cytoskeletal 2. The weighted count is computed as follows:

Keratin type II cytoskeletal 1 weighted spectrum count= 3+2*0.75= 4.5

Keratin type II cytoskeletal 2 weighted spectrum count= 1+2*0.25= 1.5

Appendix F.Terminology

Blocking

When groups of experimental units are similar, it is often a good idea to gather them together into blocks. By blocking the variability attributable to the differences between the blocks is isolated so that the differences caused by the treatments appear clearer.

Contingency table

In statistics, a contingency table (also referred to as cross tabulation or cross tab) is a type of table in a matrix format that displays the (multivariate) Frequency Table or distribution of the variables.

Dendrogram

Or tree diagram, is a common method of graphically displaying the output of hierarchical clustering.

Exclusive

Associated with a single protein group

Exclusive Unique Peptide Count

Number of different amino acid sequences that are associated only with this protein

Exclusive Unique Spectrum Count

Number of unique spectra associated only with this protein

Exclusive Spectrum Count

Number of spectra associated only with this protein

Frequency Table

In statistics, a frequency table is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample. Bivariate joint frequency distributions are often presented as (two-way) Contingency tables.

Protein Hypothesis

The protein hypothesis within Scaffold Quant is a combination of different pieces of information extracted from the input files that define the presence of a protein within a valid MS Sample.

Total

Associated with a protein group, whether or not it is shared with other protein groups

Total Spectrum Count

Total number of spectra associated with this protein including those shared with other proteins

Total Unique Peptide Count

Number of different amino acid sequences that are associated with this protein including those shared with

other proteins

Total Unique Spectrum Count

Number of unique spectra associated with this protein including those shared with other proteins

Treatment

The process, intervention or other controlled circumstance applied to randomly assigned experimental units. Treatments are the different levels of a single factor or are made up of combinations of levels of two or more factors.

Unique peptides

Peptides with different amino acid sequences, regardless of any modifications

Unique spectra

Spectra that differ in amino acid sequence, charge state or modifications

Weighted spectrum count

See Weighted spectrum count.

Appendix G.Heat map clustering

The goal of reordering the columns and rows of a data matrix is to place protein and samples with similar characteristics close to each other. Generally the reordering in heat maps is typically done using an agglomerative hierarchical clustering algorithm that groups similar data contained in a matrix or table. The clustering information is then displayed using a dendrogram.

An agglomerative hierarchical clustering algorithm on n objects begins by considering each object to be a cluster of its on containing 1 object. At each step, the two closest groups are merged together until n objects are in a single group. In the case of a data matrix an object is typically a multidimensional vector whose components are given by the data listed in a row or a column.

There are a number of possible algorithms used to create agglomerative hierarchical clusters. Their differences mainly pertain to the definition of closeness or similarity between two objects or clusters before they are merged and to the agglomeration process by which clusters are merged into larger clusters.

Similarity or closeness is typically represented by the measurement of a distance *d(A,B)* between every pair *A* and *B* of objects. Typically *A, B* are multidimensional vectors that contain a series of numbers belonging to any two rows or columns depending on the group that is being clustered, i.e. either among the columns or rows of the data matrix.

Measuring the distance between clusters that have to be agglomerated into larger clusters is more complicated than measuring distances between single vectors. Different algorithms take different approaches in defining which is the link between clusters being considered as a measure of closeness when performing agglomeration.

Scaffold Quant uses an agglomerative hierarchical algorithm called **Single-Linkage clustering** with a Euclidean distance metric. The distance metric is applied to the coordinate-wise rank vectors of the rows or columns of the data matrix containing values that depend on the selected display type in the Samples View. The ranking is done over the whole ensemble of data included in the data matrix.

This distance metric will tend to associate measurements that rank at close levels.

$$d_{Euclidean}(r_A,r_B) = \sqrt{(r_{A1} - r_{B1})^2 + \ldots + (r_{An} - r_{Bn})^2}$$

Where:

- *A* and *B* are vectors whose coordinates characterize two rows or two columns of the data matrix at a selected Display Type and summarization level, see Summarization Bar; and r_A and r_B are coordinate-wise vectors of A and B. The components of the vector are given by the displayed values shown in the Samples Table at specific summarization levels, filtering and thresholding conditions. When considering the clustering of rows, the summarization level determines how many values for a selected statistics are shown in a row and consequently defines the dimensions of *A* and *B*. When considering the clustering of the columns in the data matrix the dimension of the vectors that represent the columns is defined by filtering and thresholding applied to the Samples table.

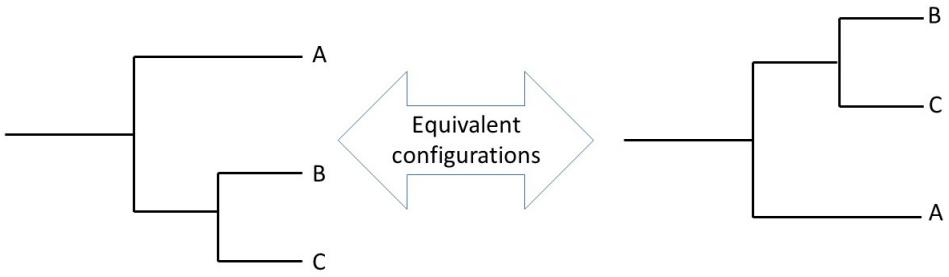
In single-linkage, clustering agglomeration is made based on a single element pair, namely those two elements (one in each cluster) that are closest to each other. The shortest of these links that remains at any

step causes the merging of the two clusters whose elements are involved. This method is also known as nearest neighbor clustering.

In Scaffold Quant clustering agglomeration is also based on a single pair element but the element selected for each cluster to be used to evaluate the shortest metric distance is the multidimensional vector that represents the center of mass among the group of vectors or points belonging to a cluster.

A common method used to graphically displaying the output of hierarchical clustering is to draw a dendrogram of the linkages among different clusters. At the bottom of the graph, each line corresponds to each object (cluster of size 1). When two clusters are merged, a line is drawn connecting the two clusters at a height corresponding to how similar the clusters are. The order of the objects is chosen to ensure that at the point where two clusters are merged, no other clusters are between them, but this ordering is not unique. When two clusters are merged, the choice of which of them is on the left or on the right side is arbitrary, this feature is called binary switching .

Binary switching example



The Heat map in Scaffold Quant

Scaffold Quant constructs a heat map using information and data available in the and displays it in the .

Whatever thresholding and filtering are applied to the determine the number of rows and columns considered for the data matrix used to create the Heat map shown in .However the rows listed include only groups of proteins even if any type of clustering might have been applied to the Samples table. Each column contains data from any MS sample or selected level of summarization.

In Scaffold Quant the quantitative values used for performing the agglomeration is determined by the selected Display type. This means that every Display type will show a different ordering of the Heat map.

The result of the clustering is visualized as a dendrogram, which shows the sequence of cluster nodes and the distance at which each node is created.

Appendix H.Techniques to Control the Family-wise Error Rate

Currently Scaffold Quant supports control of the family-wise error rate, FWER, using Holm's step-down procedure and Hochberg's step-up procedure.

The way Scaffold Quant develops the two procedures is described in the following publication: Y. Huang *et al.* Biometrika (2007), 94,4,pp.965–975.

The two methods make the same type of comparisons, but Holm starts at the smallest p-value and works down the list until one fails the bound, while Hochberg starts at the largest p-value and works up the list until one passes the bound (and then declares that everything below that passes. Hence the Holm bound is in general more conservative than the Hochberg.

For example, let's suppose we have (m=5) proteins A, B, C, D, E with p-values 0.030, 0.014, 0.013, 0.060, and 0.009 respectively, and want to reject the null hypothesis at alpha = 0.05. Let's sort the p-values and make the following table:

k	p-value	$\alpha/(m+1-k)$	p-value < $\alpha/(m+1-k)$
1	0.009	0.01	yes
2	0.013	0.0125	no
3	0.014	0.0167	yes
4	0.030	0.025	no
5	0.060	0.05	no

- The Holm step-down procedure would start at k=1 and reject $H_0(1)$ but it would stop at k=2 since the p-value is larger than the bound.
- The Hochberg step-up procedure would start at k=5, go to k=4, go to k=3, see that the bound passes and stop, accepting $H_0(1)$, $H_0(2)$, and $H_0(3)$.

Instead of simply saying whether null hypotheses are rejected or not, we report the lowest $\alpha/(m+1-k)$ bound value. When p-values are lower than the reported bound the Null hypothesis can be rejected.

This means that in the example described above, the bound for Holm would be 0.0125, and for Hochberg 0.025 and the bound reported would be 0.0125.

The reason we are reporting both methods is that technically the Hochberg procedure should only be used if the hypothesis tests are independent (which they are certainly not for Fisher's Exact Test, and not usually really for the other tests as well).

Appendix I.Using Principal Component Analysis in Scaffold Quant

Using Principal Component Analysisin Scaffold Quant

Principal Component Analysis is a tool for identifying the underlying sources of variation in a data set. PCA looks for patterns of expression among the proteins that can be used to group samples in meaningful ways. When used in combination with the flexible summarization offered in Scaffold Quant, this provides a powerful tool for exploring the biological meaning of quantitative differences observed in an experiment.

An Example

This example was performed in Scaffold perSPECtives. It uses demo file spectral_counting.sfdb. The data used in this example comes from a study to measure the effects of thermal processing on allergens in English walnuts[1] and was obtained from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org)[2] via the PRIDE partner repository, with the dataset identifier PXD000907.

To begin, we create Categories corresponding to the variables in the experiment and apply Attributes to the samples to represent the experimental design.

This study used a block design, in which four replicate samples were divided and subsamples of each underwent several treatments. The proteins from each were then extracted in two different ways.

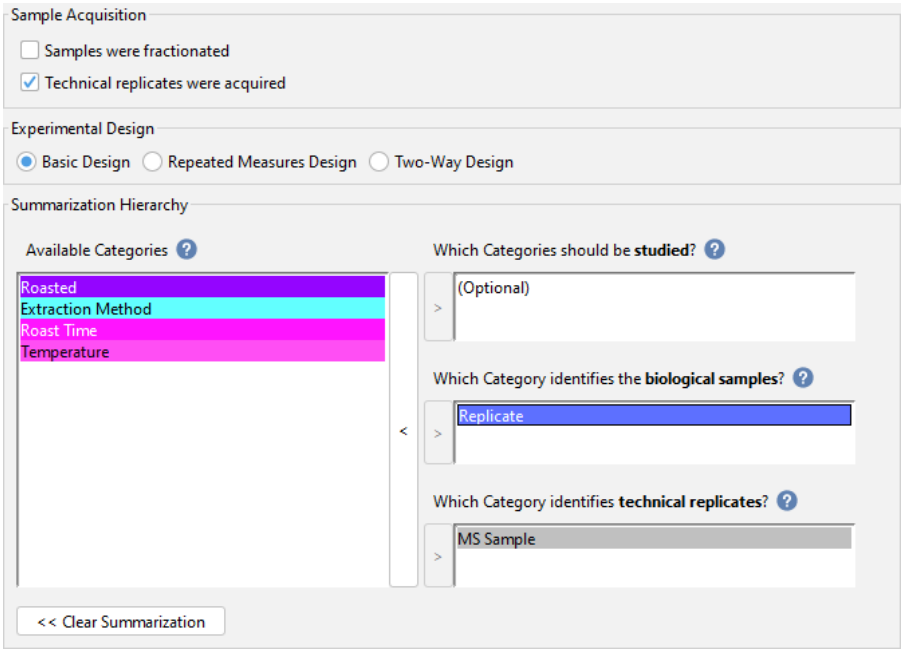
In Scaffold Quant, we create a Category for the Replicate Group Number, and one for each of the variables studied. These are Protein Extraction Method, Roasting Time, Roasting Temperature. We create the appropriate Attributes to represent the different values each of these variables may take and assign the values to the samples in the Organize View.

Figure 10: Samples with Attributes assigned

We might initially set the Summarization Hierarchy to specify our technical and biological replicates:

1. Downs ML, Baumert JL, Taylor SL, Mills EN. Mass spectrometric analysis of allergens in roasted walnuts. J Proteomics. 2016 May 2. pii: S1874-3919(16)30177-4 PubMed: 27150359.
2. Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H. 2016 update of the PRIDE database and related tools. Nucleic Acids Res. 2016 Jan 1;44(D1): D447-D456. PubMed PMID:26527722.

Figure 11: The Initial Summarization Hierarchy



The Samples View shows spectral counts for 53 proteins in the various MS Samples, but it is difficult to discern any patterns at this point:

Figure 12: The Samples View Initially

We must apply the treatment-related Categories to make this display meaningful, but there are several different treatments and we do not yet know which are important and how they interact. PCA can guide us in making these determinations.

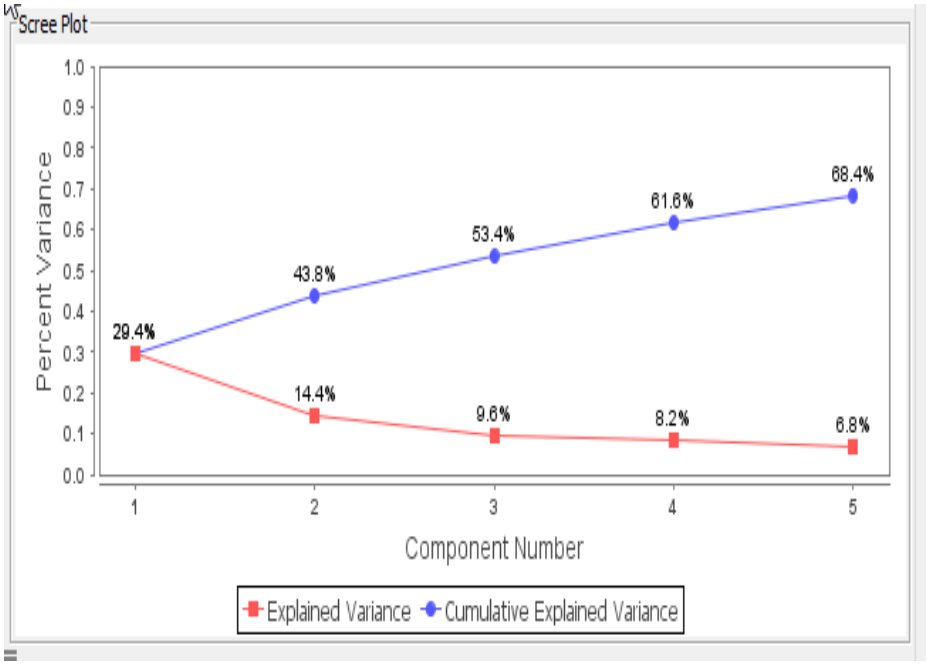
PCA analyzes data in order to find patterns in the expression levels of the various proteins that differentiate samples or groups of samples. It constructs a weighting function that, when applied to the quantitative values of the proteins, results in the greatest separation of the samples, or, put another way, explains most of the variation between samples. This is Principal Component 1. The algorithm then continues to find other independent functions that also separate the samples in different ways, although they may account for somewhat less of the variation. These become Principal Component 2, 3, etc.

The Principal Component Analysis tab in the Visualize View provides several plots to help us interpret the results of PCA.

Scree Plot

The Scree Plot indicates the percentage of variation in the data explained by each Principal Component. This may help in determining which and how many of the factors in the study need to be considered.

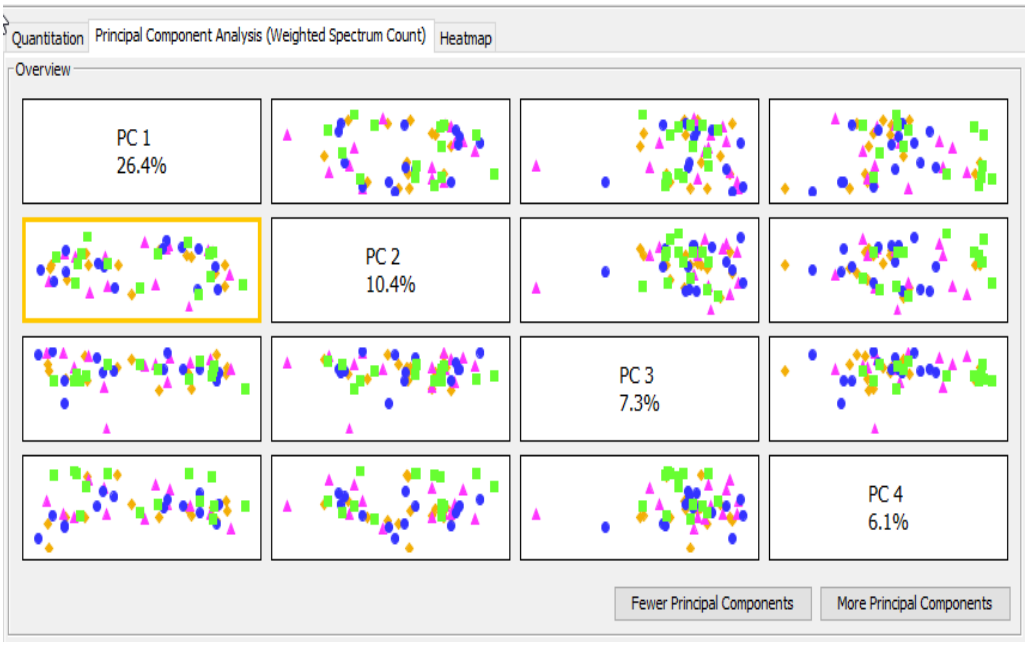
Figure 13: The Scree Plot



Overview

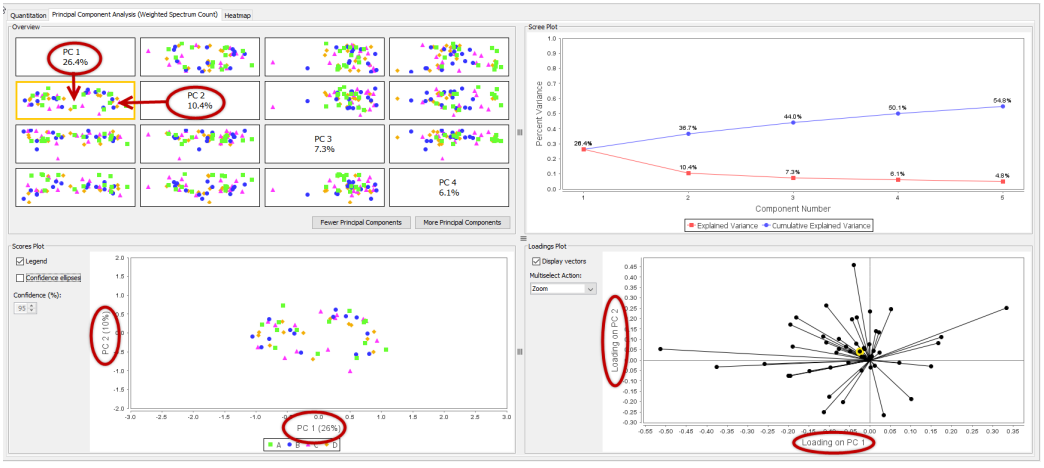
The Overview is a series of graphs where one Principal Component is plotted against another. The points in these graphs represent samples, and the X and Y coordinates are the values computed from the corresponding Principal Component functions. We can see that the samples tend to cluster in different ways depending on the Principal Components applied.

Figure 14: The Overview Plot



Clicking on a graph in the Overview selects the combination of Principal Components for display in greater detail in the Loadings and Scores Plots below.

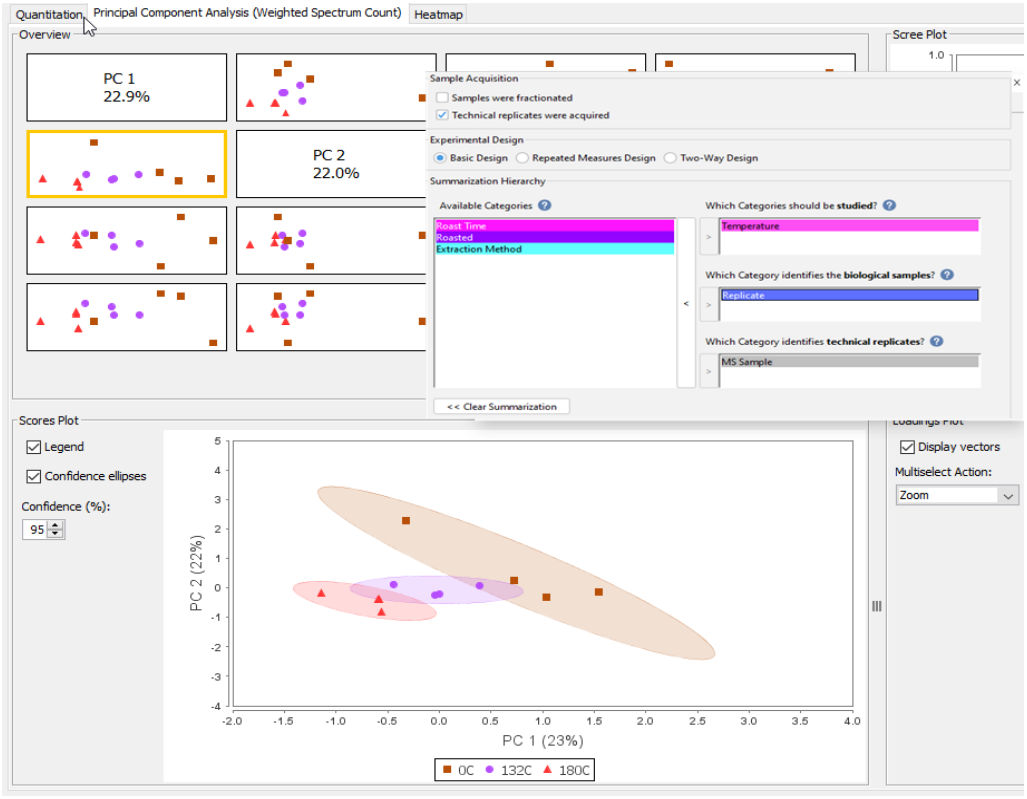
Figure 15: Selecting Principal Components with the Overview



In the plot of PC2 vs. PC1, we can see that there appears to be clustering; to determine whether one or more of the treatments applied are responsible for the variation in the data, we will try applying different Categories.

First, we try Temperature:

Figure 16: Exploring the relationship of Temperature and PC1 and PC2

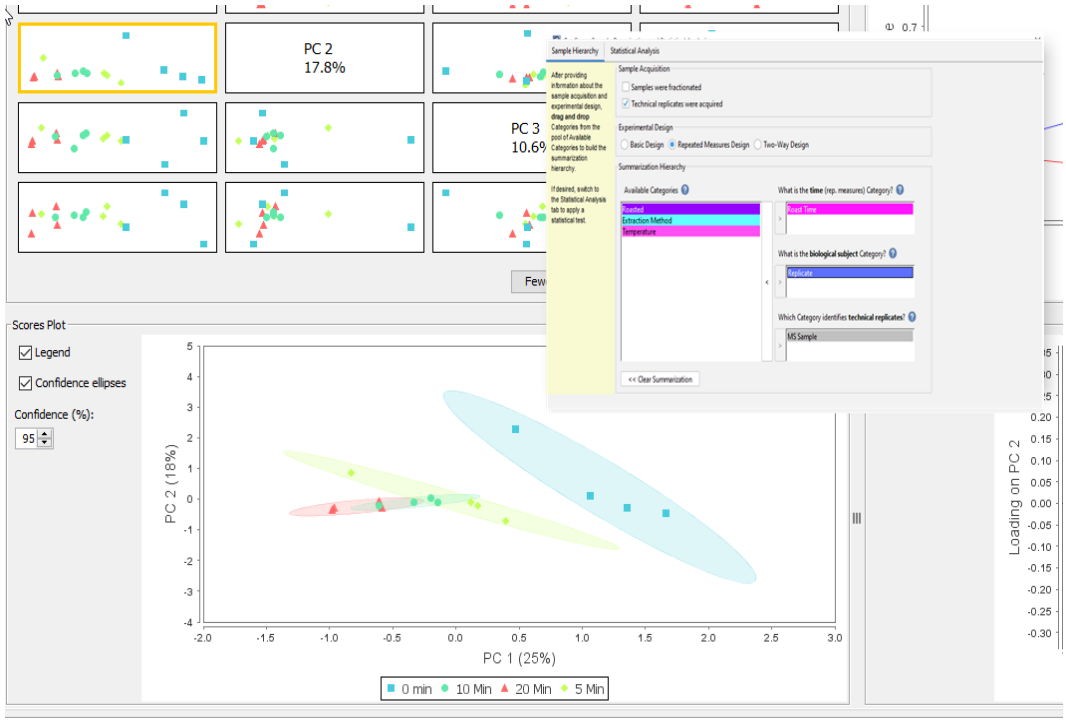


Here we see clustering in the Scores Plot. Roasting Temperature correlates with PC1, as the samples subjected to higher temperatures appear to the left in the chart (lower PC1 values), those roasted at lower temperatures are in the middle, and unroasted samples are on the right (higher PC1 values).

The temperature also appears to be contributing to some degree to PC2, as the samples roasted at the highest temperature appear slightly lower in the plot (lower PC2 values).

Looking at Roasting Time, we see:

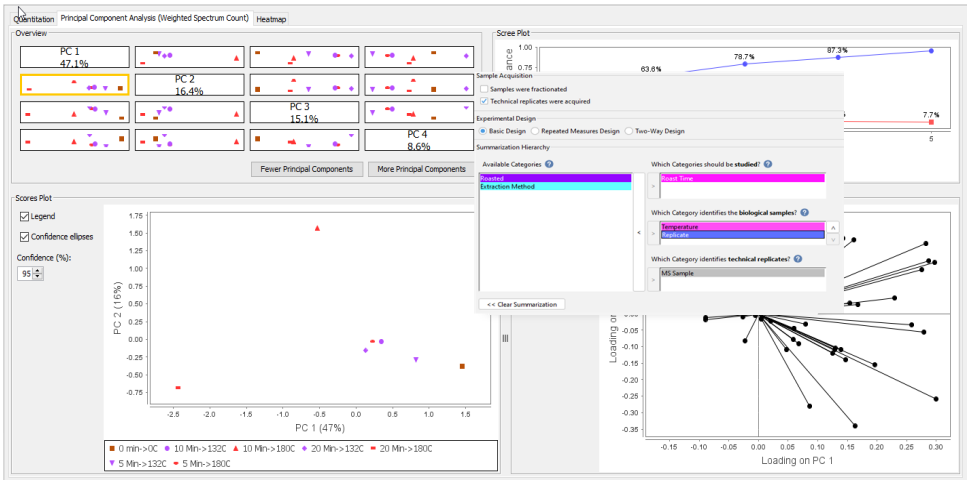
Figure 17: Exploring the Relationship of Roasting Temperature with PC1 and PC2



Once again, we see a similar pattern, with unroasted samples to the right, and the samples roasted for the longest period to the left, but there is more overlap between the different time groups. It appears that Roasting Time is correlated with PC1, but that the protein changes occur at various time points in different samples. This is probably because of interactions between roasting time and roasting temperature.

If we examine these two variables together, we see that the roasting for 10-20 minutes at 132C is similar to roasting for 5 minutes at 180C.

Figure 18: Relationship between Time and Temperature

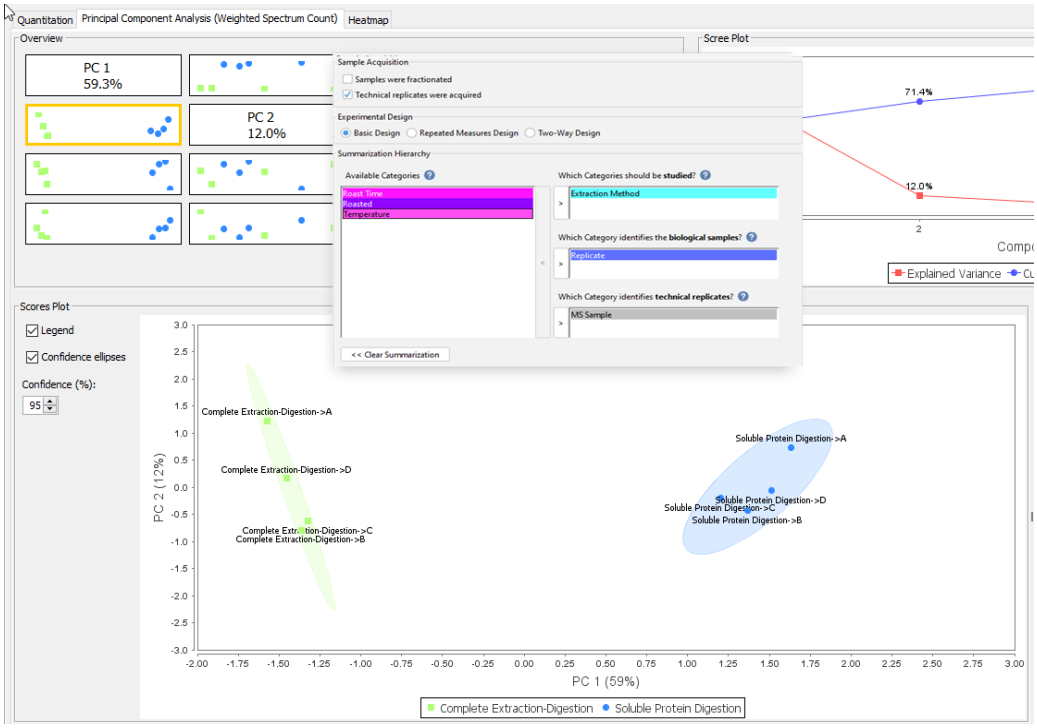


It appears, then, that the variation is explained by how thoroughly the nuts are roasted, which is governed

by a combination of time and temperature. Since temperature gave clearer results, we will use temperature as the measure of degree of roasting. Another alternative would be to create a new attribute that captures the combination of time and temperature.

As can be seen below, Extraction Method produces the clearest clustering of all in PC2 vs PC1:

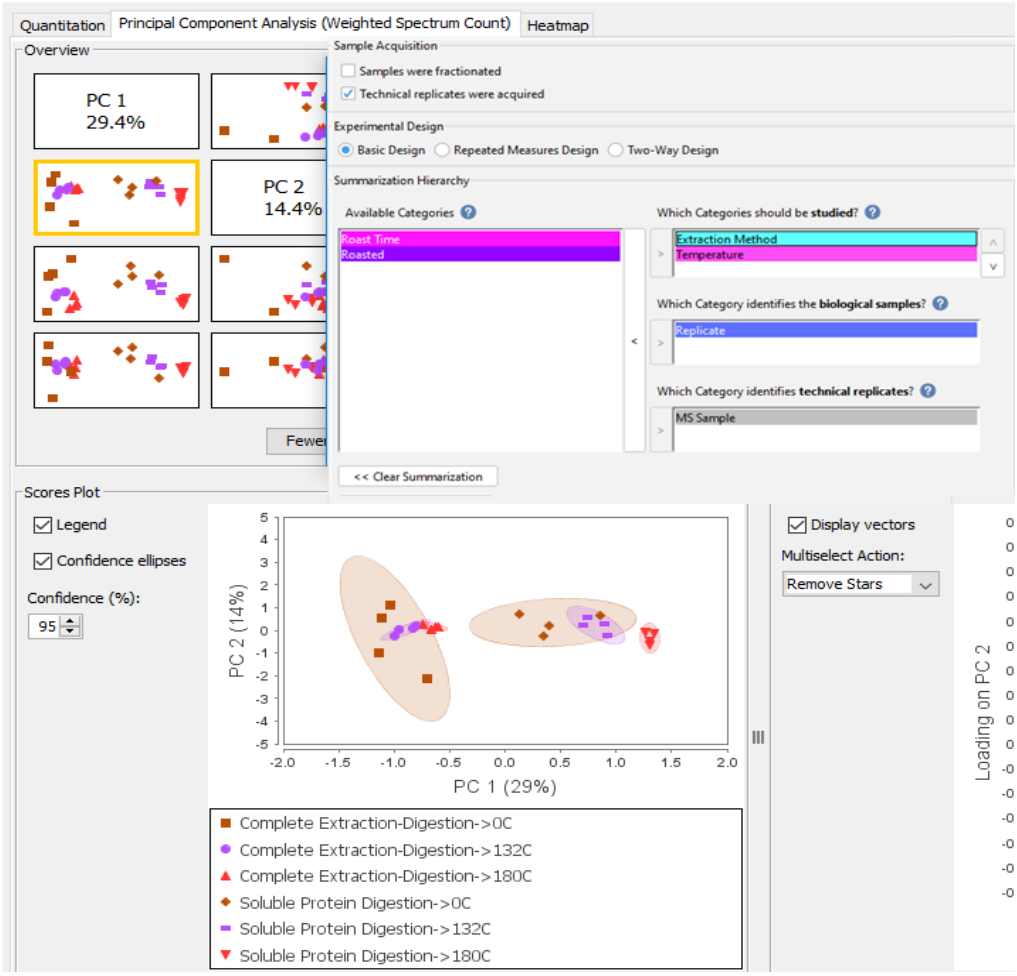
Figure 19: Exploring the Relationship between Extraction Method and PC1 and PC2



By examination of the labels, we can see that PC2 is probably based on differences among Replicates, since the Replicates appear in the same order in each extraction method and separate along the PC2 axis.

Exploring combinations of the factors produces some very clear clustering:

Figure 20: Roasting Temperature and Extraction Method



Here, the samples cluster by a combination of Extraction Method and Roasting Temperature. PC1 appears to represent a combination of Extraction Method and thoroughness of roasting.

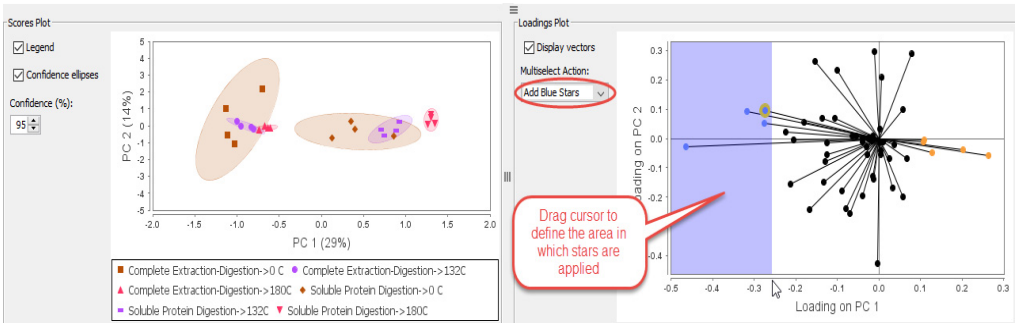
Once we have established which treatments have a significant effect on protein content and levels, we may wish to determine which specific s are most affected by them. This can help in answering questions such as which pathways are implicated in a disorder, which s are affected by a treatment, or which s might be useful in developing assays. for a certain condition. To move from samples to s, we examine the Loadings Plot.

Loadings Plot

In the Loadings Plot, each point represents a . The coordinates of each point are a measure of the contributions of that to each of the components in the plot. For example, if the plot displays PC1 on the x-axis and PC2 on the y-axis, points far to the left and right represent s that contribute strongly to Principal Component 1. s near the top and bottom contribute strongly to PC2. As a result, corresponding locations in the Scores and Loading plots are related.

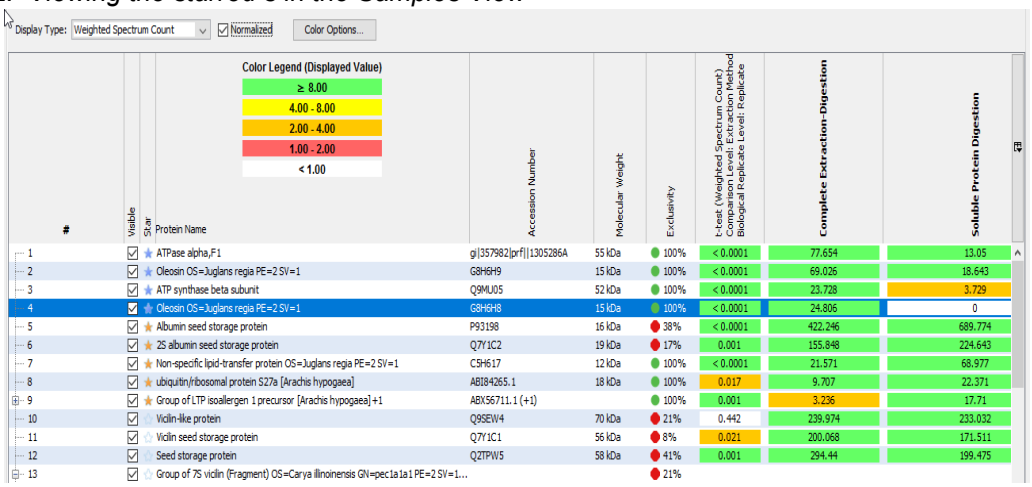
We can mark s through the Scores plot that may prove useful in identifying samples that are particularly effective at differentiating samples based on certain criteria. For example, we place orange stars on the s to the right in the Loadings Plot, and blue stars on s to the left:

Figure 21: Starring s of Interest through the Loadings Plot



In the Samples View, after summarizing by extraction method and applying a statistical test, we can see that indeed the orange-starred s are significantly higher in the soluble extraction while the blue-starred s are significantly higher in the complete extraction.

Figure 22: Viewing the starred s in the Samples View



#	Visible	Star	Protein Name	Accession Number	Molecular Weight	Exclusivity	t-test (Weighted Spectrum Count) Comparison Level: Extraction Method Biological Replicate Level: Replicate	Complete Extraction-Digestion	Soluble Protein Digestion
1	☑	★	ATPase alpha,F1	gi|357982|prf||1305286A	55 kDa	100%	< 0.0001	77.654	13.05
2	☑	★	Oleosin OS=Juglans regia PE=2 SV=1	G8H6H9	15 kDa	100%	< 0.0001	69.026	18.643
3	☑	★	ATP synthase beta subunit	Q9MU05	52 kDa	100%	< 0.0001	23.728	3.729
4	☑	★	Oleosin OS=Juglans regia PE=2 SV=1	G8H6H8	15 kDa	100%	< 0.0001	24.806	0
5	☑	★	Albumin seed storage protein	P93198	16 kDa	38%	< 0.0001	422.246	689.774
6	☑	★	2S albumin seed storage protein	Q7Y1C2	19 kDa	17%	0.001	155.848	224.643
7	☑	★	Non-specific lipid-transfer protein OS=Juglans regia PE=2 SV=1	C5H617	12 kDa	100%	< 0.0001	21.571	68.977
8	☑	★	ubiquitin/ribosomal protein S27a [Arachis hypogaea]	ABI84265.1	18 kDa	100%	0.017	9.707	22.371
9	☑	★	Group of LTP isoallergen 1 precursor [Arachis hypogaea] +1	ABX56711.1 (+1)		100%	0.001	3.236	17.71
10	☑	☆	Vicilin-like protein	Q9SEW4	70 kDa	21%	0.442	239.974	233.032
11	☑	☆	Vicilin seed storage protein	Q7Y1C1	56 kDa	8%	0.021	200.068	171.511
12	☑	☆	Seed storage protein	Q2TPW5	58 kDa	41%	0.001	294.44	199.475
13	☑	☆	Group of 7S vicilin (Fragment) OS=Carya illinoinensis GN=pec1a1a1 PE=2 SV=1...			21%			

In summary, by combining the insights gained through PCA analysis with flexible summarization and statistical analysis, we can gain insight into the biologically significant patterns in the data.

Appendix J.How PCA is Performed in Scaffold Quant

In Scaffold Quant, the variables we consider are the (thresholded and filtered) s that are currently viewable in the Samples View. These s' intensities are measured across the samples at the Biological Replicate Level. We can consider this as a collection of vectors

$$\begin{aligned} \vec{I}_1 &= (I_{11}, I_{12}, I_{13}, \ldots, I_{1m}) \\ \vec{I}_2 &= (I_{21}, I_{22}, I_{23}, \ldots, I_{2m}) \\ &\vdots \\ \vec{I}_n &= (I_{n1}, I_{n2}, I_{n3}, \ldots, I_{nm}) \end{aligned}$$

where there are n samples and m s.

Intensity data is generally roughly log normal, that is, after applying a log transformation it becomes roughly normally distributed. There is a large wrinkle introduced with this idea of applying a logarithm, however, namely, how to deal with missing values.

In order to mitigate this problem, we have opted to apply a generalized logarithm (glog) instead of a regular logarithm. We use a generalized logarithm very similar to that used by Durbin[3] which is also used by MetaboAnalyst[4].This allows us to impute missing values as intensity I=0, and still apply the transformation. Explicitly, the transformation is:

$$\mathrm{glog}(I) = \log\left(\frac{I + \sqrt{I^2 + 1}}{2}\right).$$

Note that when I is large, glog(I)≈log(I), while for I near 0, glog(I) is perfectly well defined and approximately linear.

After applying glog to all intensities,

$$a_{ij} = \mathrm{glog}(I_{ij}),$$

3. Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics. 2002;18(Suppl. 1):S105–S110.
4. Xia, J., Sinelnikov, I., Han, B., and Wishart, D.S. (2015) MetaboAnalyst 3.0 - making metabolomics more meaningful . Nucl. Acids Res. 43, W251-257.

we form the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & & a_{nm} \end{bmatrix}.$$

The rows of A correspond to the samples S1, S2, ..., Sn, while the columns correspond to the s prot1, prot2,..., protm.

For spectral counts, the same transformation is applied, but with Count substituted for I.

Now, since we are interested in the variance of this “cloud” of vectors, it makes sense to center them by subtracting out the column means. This moves the “cloud” so that it is around the origin. Call this centered matrix X.

At this point in PCA, one must make a choice between using the covariance or the correlation matrix. In the second case, one would further scale each column of the centered matrix by the standard deviation of that column. This scaling is a good choice for those whose variables are not comparable to each other, being measured on different scales, it puts everyone on equal footing. However, in this case the variables, being s measured in the same way on the same machine, etc. are comparable in scale to each other so we opt to use the covariance matrix, that is:

$$\Sigma = \frac{1}{n-1} X^T X.$$

The entries in the matrix Σ measure the covariance of the variables (s).

Now, since Σ is a real symmetric matrix, it can always be diagonalized :

$$\Sigma = V D V^T$$

where

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & & \lambda_m \end{bmatrix}$$

is a diagonal matrix consisting of the eigenvalues of Σ arranged so that λ1≥λ2≥λ3≥⋯≥λm≥0,and V is an m×m matrix whose ith column, vi, is an eigenvector corresponding to λi. (That means: Σ·vi=λivi.) It turns out that these eigenvectors v1,v2,...,vm are exactly the principal basis vectors we are seeking, and satisfy the desired bullet points.

A.Interpretation

Each principal component points in turn at the direction of greatest remaining variation. Moreover, the eigenvalues measure how much variation is accounted for by each principal component.

Percent explained variance

The percentage of variance explained by the ith principal component is given by the formula:

$$\%\ \text{explained variance} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_m}.$$

Interpretation of scores

How does dimension reduction work? Recall that each sample has a vector of its values across the s. We can project this vector onto the space spanned by, say, the first two principal components. This will give us a 2-dimensional understanding of how the samples differ. The plot of these 2-dimensional projections is called the 2D Scores Plot.

Interpretation of loadings

The dual question is how do the principal basis vectors correspond to the s? Take the first two principal basis vectors:

$$\vec{v_1} = (v_{11}, v_{12}, \ldots, v_{1m})$$
$$\vec{v_2} = (v_{21}, v_{22}, \ldots, v_{2m})$$

The coordinates of these vectors are called the loadings[5] of the s on to the principal components. Each v1j is a measure of how much the jth contributes to the first principal component (while v2j measures how

5. Actually `loading' is a loaded term in the literature; it sometimes means the coordinates of a scaled version of the basis vector. This is more often done when using the correlation matrix instead of the covariance matrix.

much it contributes to the second principal component). Note these are unit length, so

$$v_{i1}^2 + v_{i2}^2 + \cdots + v_{im}^2 = 1.$$

plotting the points (v1j,v2j) for j=1,2,...,m gives the 2D loadings plot. Each point corresponds to a . If a 's point is close to (1,0) or (−1,0) on the loadings plot, it means that this is "mostly responsible" for the first principal component hence it must explain a great deal of the variation among the samples. If it is close to (0,1) or (0,-1) it means that this is "mostly responsible" for the second principal component.

Example

Suppose we have the s prot1, prot2, prot3, and prot4 and samples S1,S2, S3, S4, and S5, and that the following table shows the glogged intensities of the s in the samples:

	prot1	prot2	prot3	prot4
S_1	6	7	10	3
S_2	4	6	8	4
S_3	5	5	6	5
S_4	3	4	4	6
S_5	7	3	3	7

This table basically shows the matrix A. Note that prot3 behaves a lot like prot2, and that prot4 also has a similar expression profile to prot2 except reversed.

This sort of observation, though tricky to see here, will become exceedingly clear after PCA decomposition.

Already the trend is a bit more clear after we compute the means of 5, 5, 6, and 5 respectively and subtract these from the columns to get the matrix X:

$$X = \begin{bmatrix} 1 & 2 & 4 & -2 \\ -1 & 1 & 2 & -1 \\ 0 & 0 & 0 & 0 \\ -2 & -1 & -2 & 1 \\ 2 & -2 & -4 & 2 \end{bmatrix}.$$

The covariance matrix is:

$$\Sigma = \frac{1}{4}\begin{bmatrix} 10 & -1 & -2 & 1 \\ -1 & 10 & 20 & -10 \\ -2 & 20 & 40 & -20 \\ 1 & -10 & -20 & 10 \end{bmatrix} \qquad (2)$$

We can see that this matrix shows that s 2, 3, and 4 are highly correlated (large values off the diagonal except in the first row/column), while 1 is not correlated with the others. We can diagonalize - (we will skip the details of how), to figure out that the principal basis vectors in this case are:

$$\vec{v_1} = \begin{pmatrix} 0.04 \\ -0.41 \\ -0.82 \\ 0.41 \end{pmatrix} \text{ and } \vec{v_2} = \begin{pmatrix} 0.99 \\ 0.02 \\ 0.04 \\ -0.02 \end{pmatrix},$$

$$\lambda_1 = 15.0$$

$$\lambda_2 = 2.5$$

(The third and fourth eigenvalues are both 0.)

Let us interpret these results. The first principal basis vector shows the linear relationship between s 2, 3, and 4. In particular, the component for the 3rd is twice that of the 2nd and 4th, and going in the same direction as the 2nd. The second principal basis vector shows that all of the remaining variation is basically occurring with 1.

The percentage of variance explained by the first principal component is

$$\frac{15.0}{15.0 + 2.5 + 0 + 0} = 85.9$$

In this case, the second principal component explains the remaining 13:1% of the variation.

Interface in Scaffold Quant

Users will find the Principal Component Analysis tab in the Visualize View. The tab shows four components.

Overview Chart

The Overview Chart allows an initial view into the first 3, 4, or 5 principal components. The squares along the diagonal denote the principal components (PCs) and show the percent explained variance for each. Off the diagonal, each square is a 2D scores plot whose axes are determined by the PC for the corresponding row and column. For details on interpreting scores plots, see section 4.3 below.

The Overview Chart allows the user to select the axes for the scores and loadings chart. Simply mouse-over the square corresponding to the desired axes and click to select those axes for the other charts in the PCA view.

Scree Plot

The Scree Plot gives a graphical display of the percent explained variance by the first 5 principal components. The lower curve shows the percent explained by each individual principal component, while the upper curve show the cumulative percent explained variance.

Scores Plot

The scores plot shows the scores: the projections of the original vectors onto the space spanned by the selected principal components. The samples, taken from the Biological Replicate Level, are denoted as dots which are colored according the attribute to which they correspond in the Comparison Level.

The scores plot also shows the 95%-confidence ellipses for each attribute. (Actually the p%-confidence ellipses where p can be specified by the user.) These ellipses show the region where 95% of the data points will lie assuming their distributions are independent and normally distributed in the dimensions being plotted. These ellipses can be used to see if attributes separate well in the currently examined dimensions.

Loadings Plot

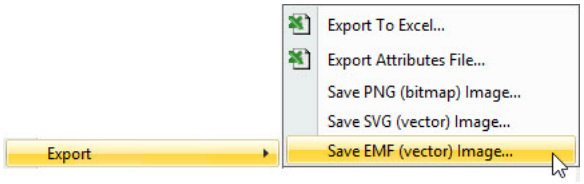
The 2D Loadings Plot shows the loadings as described above. The plot is interactive. In addition to allowing zooming, one can also use the plot to select the current (with a single click), or select a and switch to the s View with a double click.

Appendix K.Description of Mouse Right Click Context Menu Commands

- **Select All** - Select all rows in the tree table.
- **Stars** - Provides a mechanism to allow flagging of selected samples with colored stars
  - Add Orange - Adds an orange star to selected row(s). If a sample already has a blue star, it becomes a blue and orange star.
  - Add Blue - Adds a blue star to selected row(s). If a sample already has an orange star, it becomes a blue and orange star.
  - Remove Star - Removes any colored star from selected row(s).
  - Remove All Stars - Removes colored stars from all rows in the table.
- **Show/Hide** - Show or hide rows of the table.
  - Show - Set the Visible box of selected row(s) to True (checked). To use this function, select the Show Hidden box in the Filters bar at the top of the window, then select one or more rows with unchecked Visible boxes before clicking Show.
  - Hide - Set the Visible box of selected row(s) to False (unchecked). If the Show Hidden box in the Filters bar is unchecked, the row(s) will disappear from the table.
  - Hide Others - Set the Visible box of the rows **not selected** to False (unchecked). All rows except those selected will disappear from the table if the Show Hidden box is not checked.
  - Show All - Sets the Visible box of all rows to True (checked) and makes them visible in the table, regardless of the state of the Show Hidden box.
  - Show Decoys - Make all Decoys visible in the table.
  - Hide Decoys - Set Visible boxes of all decoy rows in the table to False (unchecked).
- **Clusters** - Expand or collapse clusters of proteins.
  - Expand All - Expand all clusters in the table
  - Collapse All - Collapse all clusters in the table.
  - Cluster Value Suppression - Determine whether to show quantitative values on cluster lines.
    - Show values for all clusters - Display quantitative values on all cluster header rows.
    - Only show values for collapsed clusters - Only display quantitative values for collapsed cluster header rows.
    - Do not show values for clusters - Do not display quantitative values in any cluster header rows.

- **Copy >** - Provides a number of option for copying data from a table

  - *Copy Image* - Copies the image of the current table.
  - *Copy Selected Cell* - Copies data contained in the selected cell of the current table.
  - *Copy Selected Row* - Copies data contained in the selected row of the current table.
  - *Copy All Data* - Copies data contained in the current table.
- **Export** -Provides access to a couple of exports and three different ways to export an image of the table.

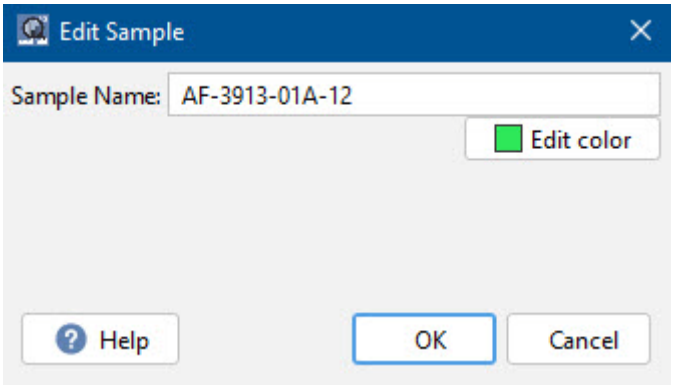


  - *Export To Excel* - Generates a tab delimited text file of the currently selected table. The file can be opened and viewed in Excel.
  - *Export Attribute File* - Generates a tab delimited text file of the meta-data attributes assigned to each MS sample in the current experiment. The file can be opened and viewed in Excel.
  - *Save PNG (Bitmap) Image* - Saves a PNG image of the selected table.
  - *Save SVG (Vector) Image* - Saves a PNG image of the selected table.
  - *Save EMF (Vector) Image* - Saves a EMF image of the selected table.
- **Find** - Opens the Find dialog.
- **Print** - Prints an image of the current table.

Organize View

- **Edit sample...** - Opens a dialog that allows editing of the names of the selected sample or the color assigned to that sample. If more than one sample is selected, the edit applies only to the first selected sample.

.



- **Category names** - One row for each Category that has been defined. Selecting a Category opens a dialog showing all attributes defined for that Category. Selecting an attribute assigns that attribute value to the selected sample(s).
- **Expand All** - Expands the display of the Samples tree to show all samples.
- **Collapse All -** Collapses the Samples tree to show only the categories.
- **Edit Selected Sample Names**... - opens a dialog box similar to the Bulk Edit Sample Names dialog, but changes made through this dialog will apply only to the currently selected sample(s).
- **Copy, Export, Print, Find** - as described in The Samples View.

Index

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W