Scaffold DIA User Guide
Version 3.4.0

Release Information

The following release information applies to this version of the Scaffold DIA User's Guide. This document is applicable for Scaffold DIA, Release 3.3.1 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 PTM, Scaffold DIA, Scaffold DDA,* Scaffold DDA 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)
https://www.proteomesoftware.com

Contents

Preface.... 4
Chapter 1: Getting Started with Scaffold DIA.... 6
Chapter 2: Preparing data for Scaffold DIA.... 18
Chapter 3: Loading data into Scaffold DIA.... 27
Chapter 4: Scaffold DIA Main Window.... 49
Chapter 5: The Samples View.... 78
Chapter 6: The Organize View.... 93
Chapter 7: The Proteins View.... 117
Chapter 8: The Visualize View.... 131
Chapter 9: The Analysis View.... 146
Chapter 10: The Publish View.... 148
Chapter 11: Quantitative Methods and Tests.... 153
Chapter 12: Protein Grouping and Clustering.... 163
Chapter 13: Reports.... 165
Appendices.... 168
- Appendix A: Structure of Scaffold DIA files (*.SDIA).... 169
- Appendix B: Computation of FDR in Scaffold DIA.... 170
- Appendix C: Summarization: Rolling up Values.... 172
- Appendix D: Missing Values.... 174
- Appendix E: Shared Evidence Clustering.... 176
- Appendix F: Heat map clustering.... 177
- Appendix G: Techniques to Control the Family-wise Error Rate.... 179
- Appendix H: Using Principal Component Analysis in Scaffold DIA.... 180
- Appendix I: How PCA is Performed in Scaffold DIA.... 189
- Appendix J: Description of Mouse Right Click Context Menu Commands.... 195

Index.... 198

Preface

Welcome to the Scaffold DIA User's Guide. Its purpose is to answer users' questions and guide them through the procedures necessary for using Scaffold DIA efficiently and effectively.

Using the manual

A Table of Contents and an Index are provided in this manual for the user's convenience. This Preface also provides a brief discussion of each chapter to further assist users in locating needed information.

Special information about the manual

This User's Guide has a dual-purpose design. It can be distributed electronically and printed on an as-needed basis, or it can be viewed on-line in its fully interactive capacity. If users print the document, for best results it is recommended that they print it on a duplex printer; however, single-sided printing is also possible. When the document is viewed on-line, a standard set of bookmarks appears in a frame on the left side of the document window for navigation through the manual. For better viewing, users can decrease the size of the bookmark frame and use the magnification box to adjust the display according to their viewing preferences.

If users do print the document using a single-sided printer, they might see a single blank page at the end of some chapters. This blank page has been added solely to ensure that the next chapter begins on an odd-numbered page. This blank page in no way indicates that the book is missing information.

Conventions used in the manual

The *User's Guide uses the following conventions:*

- Information that can vary in a command—variable information—is indicated by alphanumeric characters enclosed in angle brackets; for example, <Protein Name>.
- A new term, or term that must be emphasized for clarity of procedures, is *italicized*.
- Page numbering is "on-line friendly." Pages are numbered from 1 to x, *starting with the cover* and ending on the last page of the index.
- This manual is intended for both print and on-line viewing.
- If information appears in blue, it is a hyperlink. Table of Contents and Index entries are also hyperlinks. Click the hyperlink to advance to the referenced information.

Assumptions in the manual

The Scaffold DIA User's Guide assumes that*:*

- The user is familiar with Windows operating systems, and basic Windows navigational elements, content formatting and layout tools.
- The user has the appropriate licensing to run Scaffold DIA

Chapter 1: Getting Started with Scaffold DIA

System Requirements

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

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

Installing Scaffold DIA

Scaffold DIA runs on Windows, MAC or Linux systems., but because it uses ProteoWizard's MSConvert program to read raw files, it can only process vendor format raw files on Windows. MzML files may be loaded or the Scaffold DIA Viewer may be used to process already loaded data on any of the operating systems. Follow these instructions to install the application on your system:

Request an evaluation by filling out the form found at www.proteomesoftware.com/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 DIA installation Setup Wizard

3. In order to process raw data, Scaffold DIA requires a working installation of MSConvert, which is an application included in the open source cross-platform tool kit **Pr**oteoWizard[1]. During the initial installation of Scaffold DIA, the installer will prompt the user to download and install ProteoWizard. If you already have MSConvert installed on your system and the version number is greater than or equal to the indicated version, you may skip this step and proceed to step 5.
4. If a ProteoWizard installation is not yet available on your computer or the proper version is not installed, click the button labeled “Download ProteoWizard”. A web browser will open; select the option, “I agree to the licensing terms -download ProteoWizard” and install ProteoWizard.Once you have completed the ProteoWizard installation, return to the Scaffold DIA dialog box to finish the installation of Scaffold DIA.
5. Click “Next”. If Scaffold DIA has located an acceptable version of MSConvert, its location will be displayed. If the box is empty, use the Browse button to locate and select the MSConvert.exe file on your system; generally, this will be located in a subfolder of “C:\Program Files\ProteoWizard”. Click “Next”.
7. The installer will then provide you an opportunity to allocate memory to Scaffold DIA. We recommend that you set the Maximum Memory to approximately 80% of the amount of physical RAM on your system. Click “Next”.
8. 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 SDIA files and Scaffold DIA, and the following screen allows creation of desktop icons. Clicking “Next” begins the installation.
9. Finally, Scaffold DIA allows you to select the option to have the program open at the closing of the wizard. Click “Finish”.

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

After Scaffold DIA 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 DIA, as does, for Windows computers, selecting the option from the Start menu (**Start > All Programs > Scaffold DIA > Scaffold DIA**)Licensing

Scaffold DIA Licensing

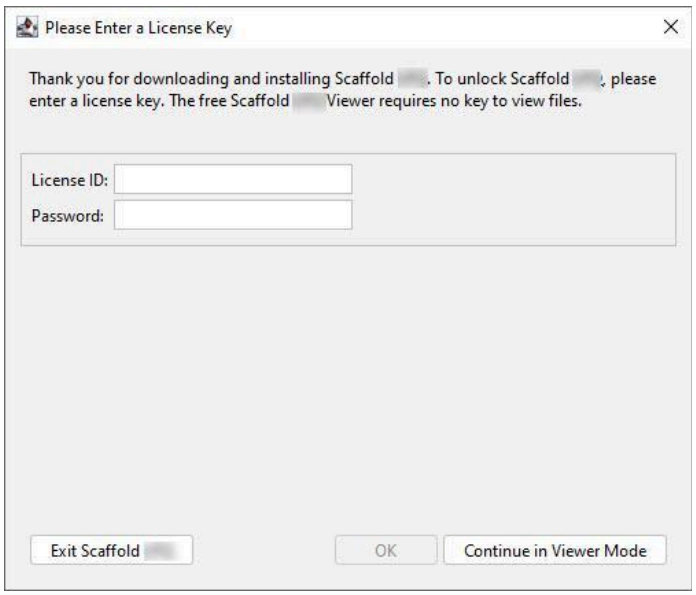
The first time Scaffold DIA opens after installation, the Enter License Key dialog box opens.

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

---

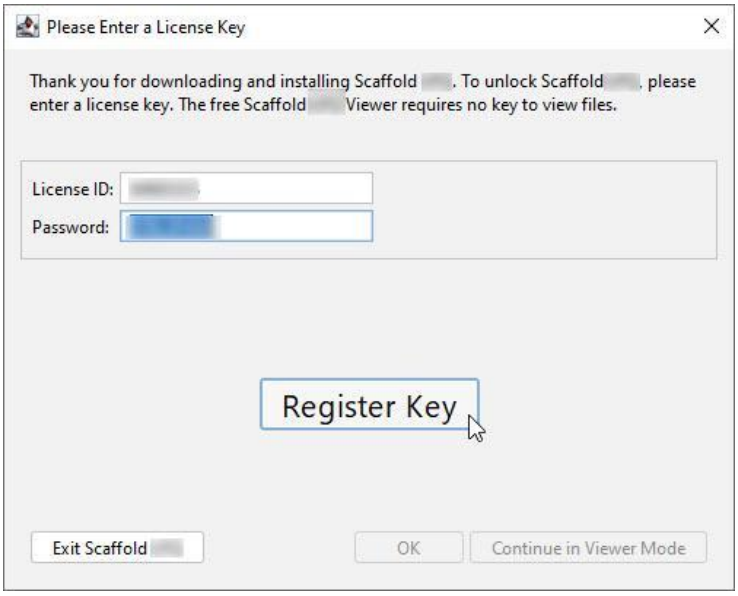
[1]Chambers, M. et *al*. Nature Biotechnology, 30, 918–920 (2012) doi:10.1038/nbt.2377

Figure 1-2: Scaffold License Key messages



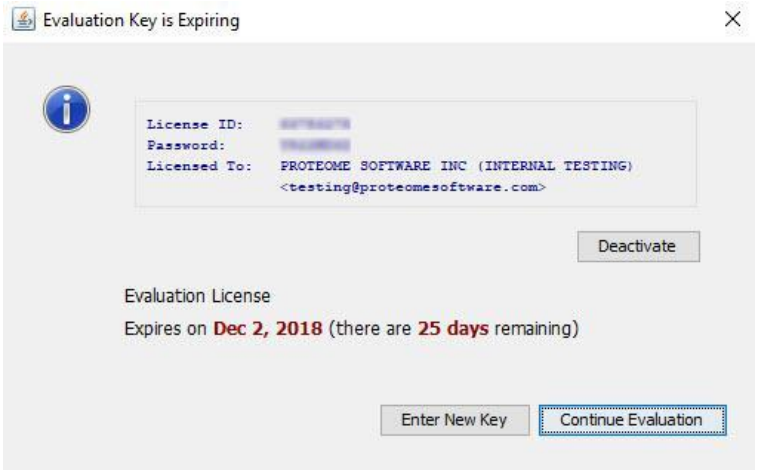
Evaluation key - An Evaluation key is valid for a limited period. A free evaluation key for Scaffold DIA may be obtained through www.proteomesoftware.com. An evaluation key may be used on two computers. Once the key and password have been copied and pasted into the license key dialog box, a message will appear below it, displaying confirmation of the key registration. Pressing OK starts the application.

Figure 1-3: Evaluation License key



Every time Scaffold DIA 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 1-4: Message appearing when launching an evaluation copy of Scaffold DIA

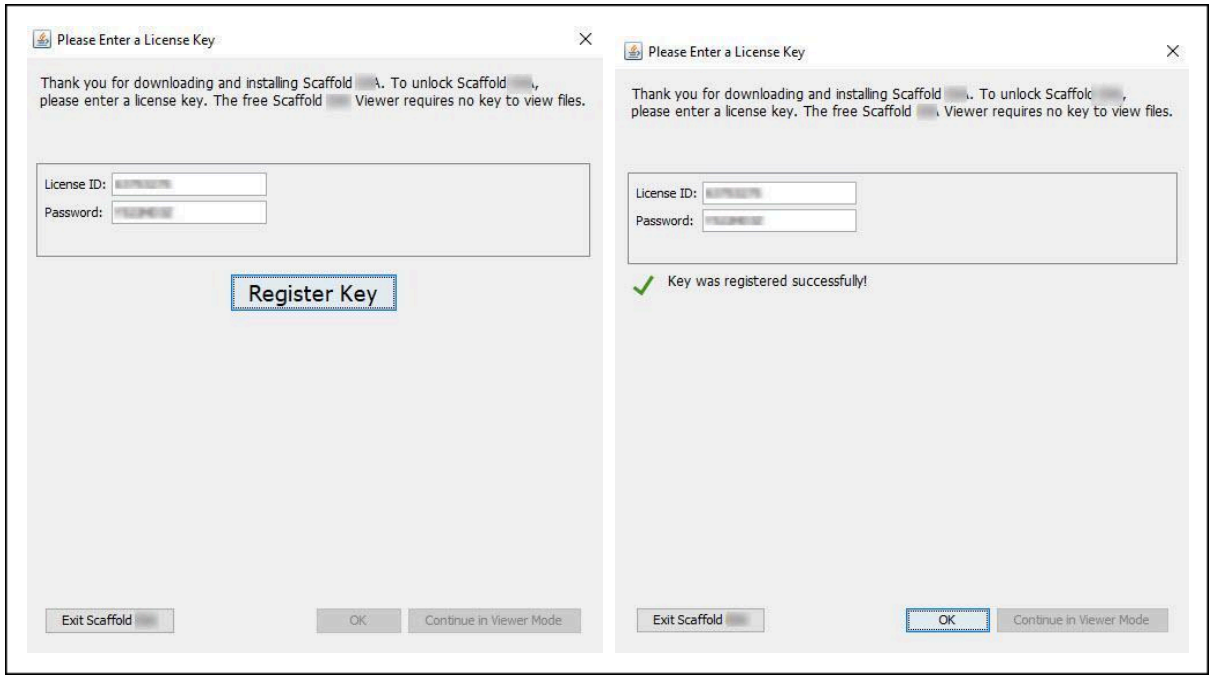


Time-Based License key—a Time-Based License key allows the user to access all features of the software permanently. It only allows upgrades within a certain time limit, however. The time tracks the length of the support contract. Once expired, Scaffold DIA will continue to work beyond the expiration 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 DIA installation to a different computer, see Moving Scaffold DIA to another computer for instructions to transfer the key at no charge.

Figure 1-5: 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 DIA Welcome message opens.

If the user is using an evaluation copy of Scaffold DIA, 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 DIA Welcome message opens.

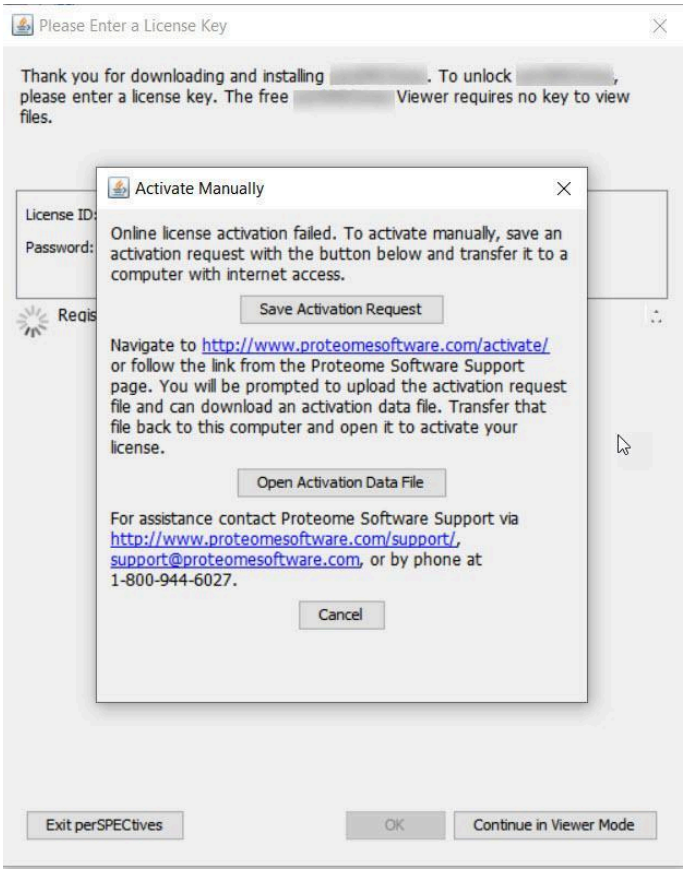
From this window, the user may create a new experiment, open an existing experiment (*.SDIA file), or work with the demonstration data that is provided in the Scaffold DIA installation.

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.

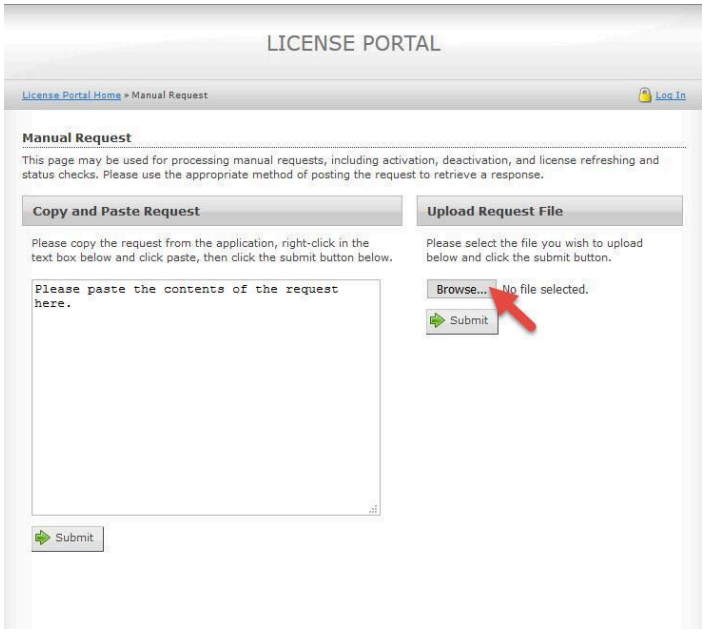
Figure 1-6: Manual or offline activation dialog



To activate Scaffold DIA without an internet connection:

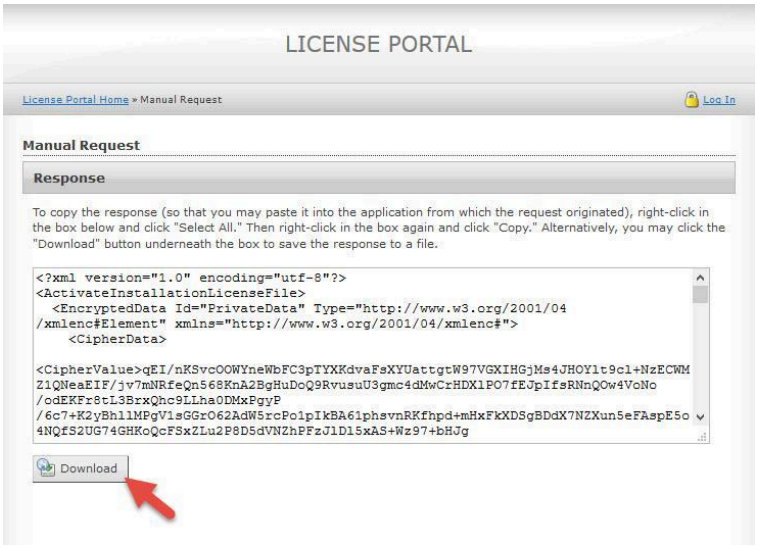
1. First, use the Save Activation Request button to create an activation request file.
2. Transfer this file to a computer with internet access (e.g. using a USB drive).
3. On the connected computer, navigate to https://www.proteomesoftware.com/activate/**.** This link is also accessible from the Proteome Software Support page *(*https://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 1-7 below).

Figure 1-7: 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 1-8 below).

Figure 1-8: License Portal Response to Activation Request

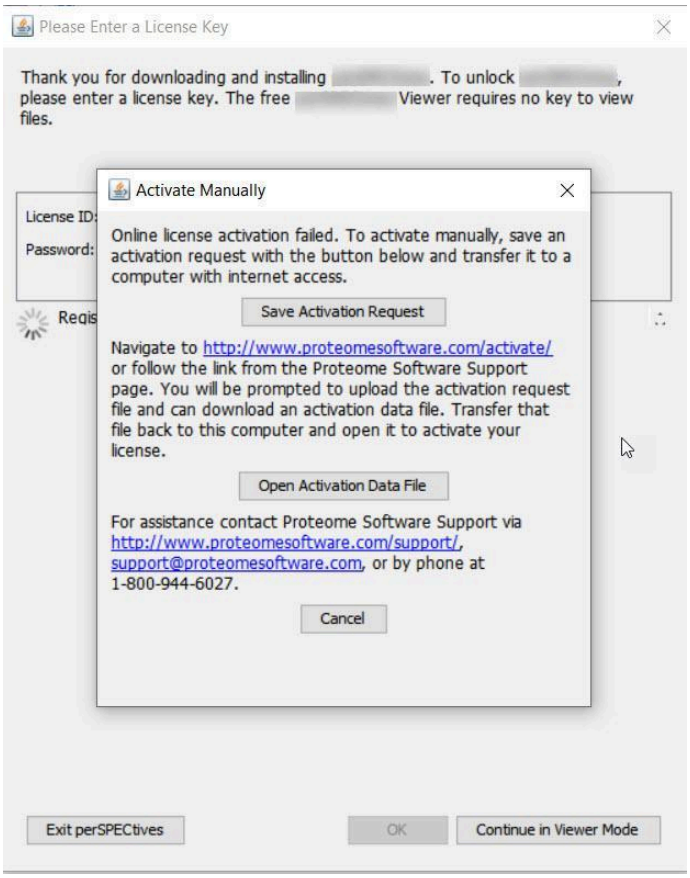


6. Click the Download button to save the response to a file named response.xml, which will be downloaded to the default download location.

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

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

Figure 1-9: Select Activation File returned by the License Portal



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

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

Time based license key renewal

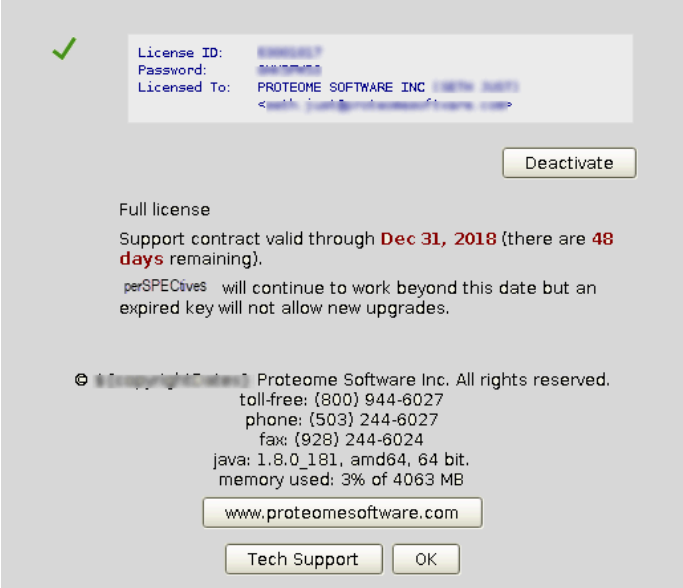
Time based license keys have time limits connected to the term of the user's support contract. When the support contract expires, Scaffold DIA continues to work but upgrades are not allowed until the contract is renewed. The status of the Scaffold DIA license key may be checked by selecting **Help >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

If the contract has expired and the user wishes to upgrade Scaffold DIA, 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 DIA** from the **Help** menu displays a screen that supplies information about the licensing status of the software.

Figure 1-10: About Scaffold DIA dialog



Moving Scaffold DIA to a different computer

Each permanent Scaffold DIA 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 DIA dialog.
2. Select Help>>Update License Key and click the Deactivate button.

To reinstall Scaffold DIA:

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 DIA.

Scaffold DIA Viewer

A free Scaffold DIA Viewer may be downloaded from www.proteomesoftware.com. The Viewer can open and display any *.SDIA file created by Scaffold DIA, and allows users to distribute Scaffold DIA 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 DIA 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 DIA, 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 DIA may be run on a computer at one time. Additional instances will function as Viewers

Scaffold DIA highlights

Scaffold DIA is an identification and quantification program for data independent acquisition proteomics. As with all of our software, users can expect a simple, easy to use interface, numerous options for collaboration such as reports and sharable files and robust statistical analysis.

Scaffold DIA is designed to help researchers identify, organize, summarize, refine, quantify and visualize results of their data independent acquisition proteomics experiments. This program supports complex experiments by combining sample attribute details (metadata) with Mass Spectrometry results.

Identify

Scaffold DIA can identify proteins from DIA proteomics data by comparing against peptides in a previously created Scaffold DIA, Skyline or Scaffold library, or it can perform an in-silico digestion of a protein FASTA database and match against the resulting peptides. The user-friendly Workflow Dialog helps users specify the details of an experiment, and the program does the rest. Two distinct workflows allow the user to analyze experiments of various degrees of complexity. Scaffold DIA handles protein inference, grouping and clustering, and provides peptide and protein-level FDR thresholding.

Organize

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 DIA, the user can create categorical variables or "Attribute Groups" and annotate samples using the "Attributes" in these groups. Attributes may be added through a graphical interface or loaded from a file. This provides the foundation for evaluation of proteomics experiments from many viewpoints.

Summarize

Once Attributes are applied, Scaffold DIA offers a great deal of flexibility to allow easy viewing of similarities and differences across sample groups. Flexible summarization allows the user to select Attribute Groups 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.
- Filter by FDR, peptide sequence, taxonomy, etc.

Quantify

Quantification is based on carefully selected representative ion fragments for each peptide. Normalization is performed across samples; peptide intensities are summed to calculate the protein intensities, and are also properly combined according to the organizational hierarchy specified. This allows the user to:

- Evaluate various types of experiments, including basic designs, repeated measures and two-way analyses, assessing significance with a variety of statistical tests.
- Validate peptides used for quantification.

Visualize

All proteins are easily visible, even those with equivalent peptide evidence. 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 DIA helps display patterns of expression across many samples with various attributes to provide new insight into an experiment.

Referencing Scaffold DIA Results

Users are free to copy, modify, and distribute the following example when citing Scaffold DIA in their publications and reports:

Scaffold DIA (Proteome Software, Portland, Oregon, USA) was used to identify and quantify proteins in Data Independent Acquisition proteomics results. Peptide false discovery rate was controlled by Percolator (Käll (2007)). Quantification was performed by EncyclopeDIA[2] (https://doi.org/10.1038/s41467-018-07454-w).

[2]Searle, B.C., Pino, L.K., Egertson, J.D. et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. Nat Commun 9, 5128 (2018).

Chapter 2: Preparing data for Scaffold DIA

Collecting Data for DIA Experiments

For information about how to collect data for analysis in Scaffold DIA, please see http://*www.proteomesoftware.com/documentation/scaffold_dia_ms_methods/*

DIA vs. DDA

DDA workflows are complex, especially when the goal is maximal coverage and reproducibility. Replication is necessary to ensure a specific peptide can be identified and quantified with confidence. DIA Analysis provides more complete coverage and more confident quantification with fewer replicates than DDA.

Figure 2-1: Collecting Data for DDA Analysis

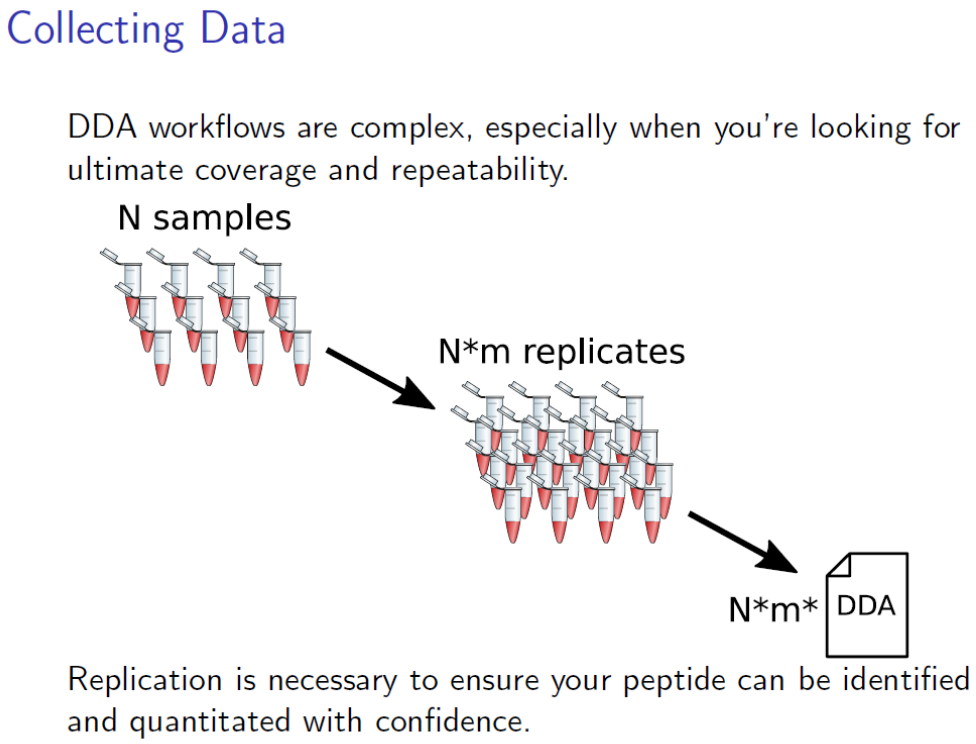
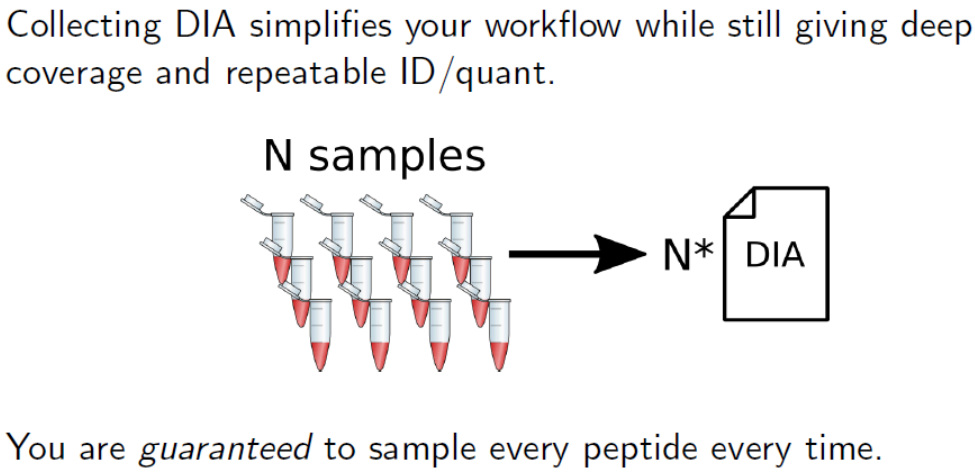


Figure 2-2: Collecting data for DIA Analysis



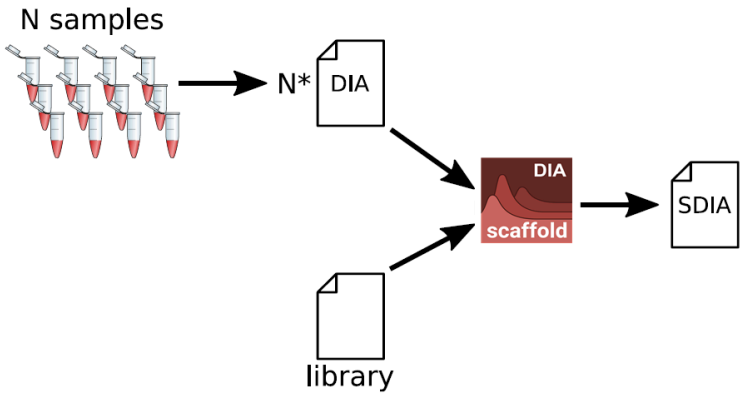
Scaffold DIA Workflows

Experimental Data Search Workflow

In simplest terms, Scaffold DIA identifies and quantifies proteins by comparing fragment patterns in the DIA data to patterns in a spectral library.

Figure 2-3: The Basic Scaffold DIA Workflow

Scaffold DIA offers one-step analysis of DIA experiments using library-based identification algorithms.

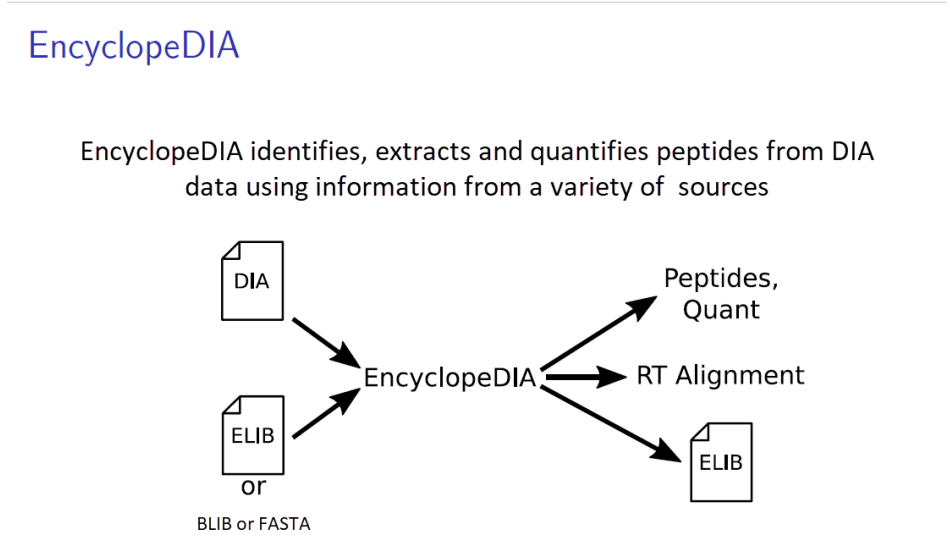


The program's flexibility comes from the fact that the reference libraries to be searched may be created in a number of ways. Any combination of the following may be used:

- **Prosit Library** - Scaffold DIA supports searching against spectral libraries created with Prosit[3]. Prosit creates high quality spectral libraries from FASTA databases. Searching these libraries produces more identifications with lower FDR than direct FASTA searches. Proteome Software has created a number of Prosit libraries formatted for use in Scaffold DIA for various species. Proteome Software will also create and share additional libraries at users' requests; please click the link in the Download dialog in the Library Manager or contact *support@*proteomesoftware.com for assistance. Note that creation of custom libraries may take up to 48 hours. Users may also create their own Prosit libraries using the online Prosit server (https://www.proteomicsdb.org/prosit/). Scaffold DIA will accept these Prosit libraries directly if they are named using the file extensions .csv or .spectronaut.
- **Scaffold or Skyline BLIB** - A library may be exported from Skyline or Scaffold as a BLIB file. It is also possible to export an iRTDB file from the same Skyline analysis and to use this to guide retention time alignment in Scaffold DIA.
- **Spectral Library** in one of the following formats: **DLIB**, **TraML**, **MSP, SPTXT**.
- **FASTA file** - Scaffold DIA can use EncyclopeDIA to create a theoretical library from an in-silico digest of a protein FASTA database. This process is slower and more memory-intensive than searching against a spectral or chromatogram library, and should only be used for processing a small number of files with relatively few modifications. A FASTA search may be used to create a chromatogram library which is then used in a subsequent search of the full experiment.

Figure 2-4: Search against an existing library

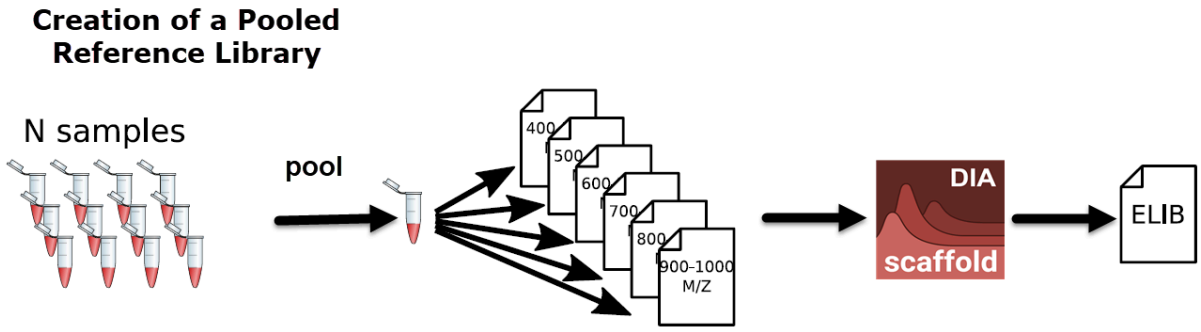
[3]Gessulat, Schmidt et al. 2019" DOI 10.1038/s41592-019-0426-7



- **Scaffold DIA ELIB** - A chromatogram library, designated as an ELIB file, may be created by Scaffold DIA by searching DIA data against either a Prosit library, a BLIB or other spectral library or a FASTA database.

  One especially rigorous workflow is to create a pooled reference sample with aliquots from all of the biological samples and to use gas-phase fractionation then create narrow-window DIA files from the fractions.

Figure 2-5: Creating a chromatogram library

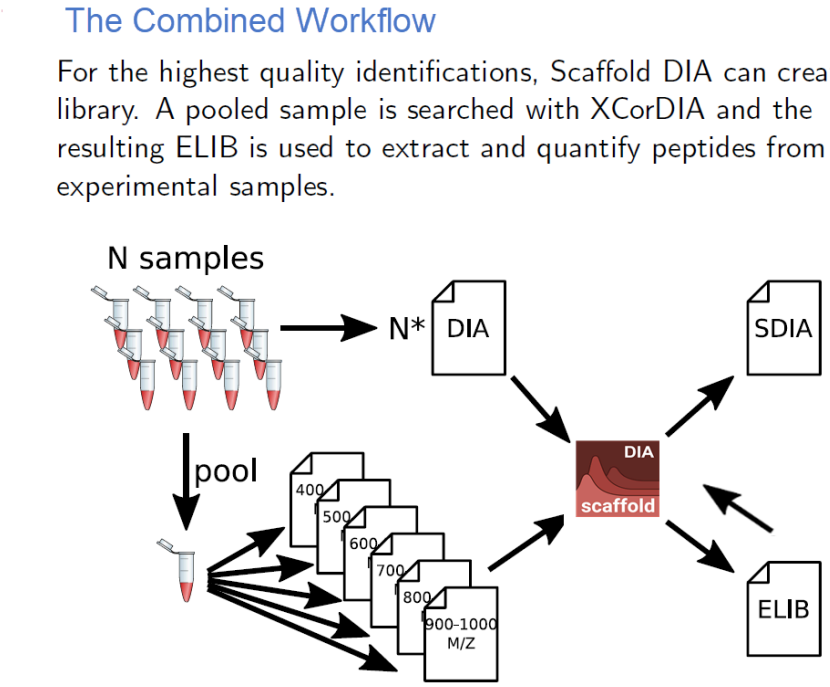


The reference samples may be searched against a Prosit library, a BLIB obtained from Skyline or Scaffold, a FASTA database, or against an ELIB created in a previous run of Scaffold DIA. The resulting ELIB is then used as the reference library for subsequent analysis of the individual biological samples.

Combined Reference and Data Search Workflow

Scaffold DIA also allows the user to create and search the reference library in a single run. In this case, two searches are performed sequentially. First, the reference samples are searched against a specified library or FASTA file to create an ELIB and then the biological samples are searched against this ELIB.

Figure 2-6: The combined workflow using a pooled reference



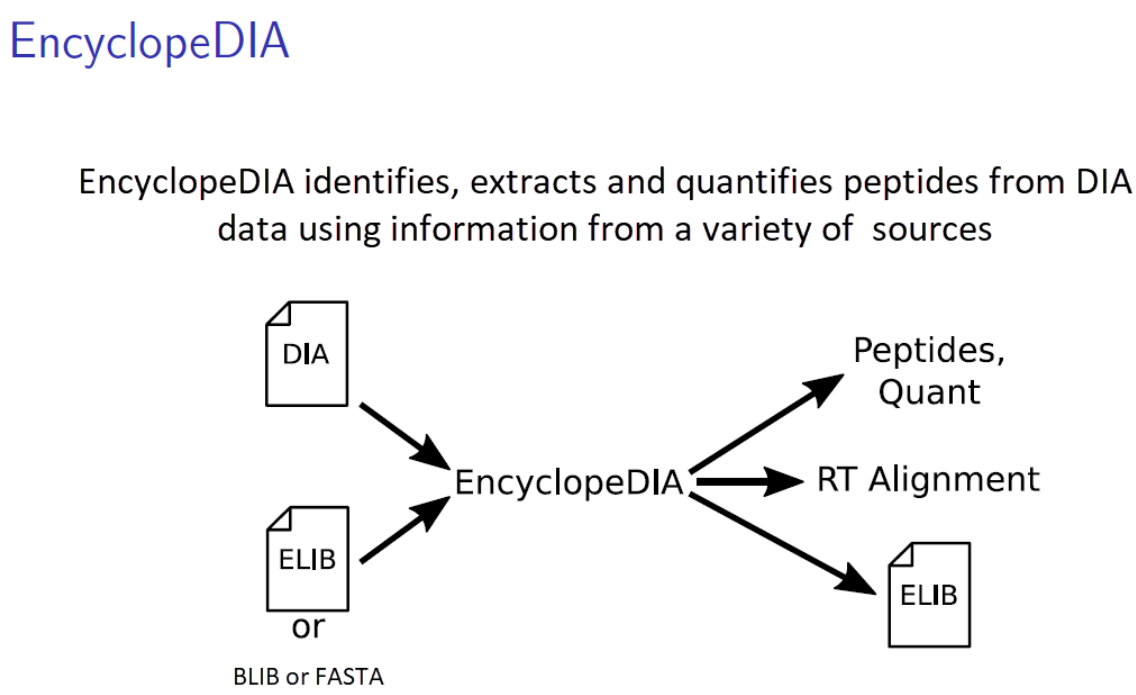
The Workflow Dialog is designed to support all of these options. The interface adjusts to prompt for the proper input parameters based on experimental design options specified by the user.

For detailed instructions about setting up each type of experiment, see The Workflow Dialog.

How Scaffold DIA Analyzes an Experiment

Scaffold DIA identifies proteins in samples processed by Data Independent Acquisition (DIA) mass spectrometry; performs relative quantification in order to illuminate differences in expression between samples or groups of samples and provides statistical tools, figures and links to external data sources to assist in assessing the biological significance of these differences. For its core functions, Scaffold DIA uses EncyclopeDIA[4], a DIA search tool developed by the MacCoss laboratory at the University of Washington.

Figure 2-7: Flow of Data in Scaffold DIA



Raw File Processing

Scaffold DIA uses the MSConvert program supplied with ProteoWizard[5] to process raw DIA files. The

Scaffold DIA installer prompts the user to download and install ProteoWizard, and once it is installed, all interactions with it are handled transparently by Scaffold DIA. MSConvert reformats the raw files to mzML format, performing demultiplexing of windows if necessary. The mzML file is temporarily written to the specified Processing Directory (Processing Directory) and then deleted once its data has been extracted and stored in a SQL database for further processing.

[4]https://bitbucket.org/searleb/encyclopedia

[5]Chambers, M. et *al*. Nature Biotechnology, 30, 918–920 (2012) doi:10.1038/nbt.2377

Peptide Identification in Scaffold DIA

Scaffold DIA uses EncyclopeDIA to identify and quantify peptides. EncyclopeDIA identifies peptides by comparing the ion patterns in each precursor isolation window of the DIA data file to a reference library created from a set of peptides which might be present in the sample. These potential peptides may be drawn from a variety of sources:

Potential Peptide Sources

Skyline or Scaffold library - If a previous analysis of the samples (or of a reference sample) has been performed, a BLIB file may be exported from Scaffold or from Skyline. This provides a set of highly confident peptide identifications on which to base the DIA analysis. If the BLIB results from processing multiple samples together in Skyline, the user may also export an iRTDB file and provide that to Scaffold DIA for internal RT alignment.

No special peptides are necessary to create an iRTDB.

Prosit library - Prosit creates high quality spectral libraries from FASTA databases, and searching against these libraries produces more identifications with lower FDR than FASTA searches. Prosit libraries for a number of species are available for download through the Scaffold DIA library manager, and Proteome Software will create and share additional libraries at users' requests. The user may request creation of a custom Prosit library via a link in the Download Dialog of the library manager. Scaffold DIA will also accept libraries created by users with the online Prosit server, if they are named with the file extensions .csv or .spectronaut (https://www.proteomicsdb.org/prosit/).

Spectral library in one of the following formats: **DLIB**, **TraML**, **MSP, SPTXT**.

FASTA – An in-silico digest of the protein sequences in the FASTA file is performed, and theoretical peptides are constructed, based on the amino acid sequences and the specified modifications. Searching against a FASTA provides the possibility of achieving full coverage of the proteome, but it is much slower than searching against a previously created library.

DIA Chromatogram library – Scaffold DIA offers the option to create a deep-coverage reference library entirely from DIA data. We call this a "chromatogram library." This option offers maximal coverage of the proteome with less costly and time-consuming data acquisition and processing. A DIA Chromatogram library may be created and used in a subsequent search within a single Scaffold DIA run (see Combined Reference and Data Search Workflow), or may be created independently and saved for future use.

Decoy Peptides

Within EncyclopeDIA, a decoy sequence is generated for each library entry by reversing the amino acid sequence with the exception of the digestion enzyme-specific termini. Peaks from the library spectrum are shifted to the expected m/z values of the corresponding fragment ions in the decoy, e.g. the m/z of the b3 ion of the library spectrum is shifted to the expected m/z of the b3 ion of the decoy sequence, while retaining its intensity, delta mass, etc. If the library was collected in profile mode, all peaks within the fragment delta mass tolerance are reproduced in this way.

Peptide Scoring in EncyclopeDIA

The primary peptide score computed by EncyclopeDIA assesses only the peaks in the precursor isolation window that match peaks in the library spectrum. No penalty is assigned for ions that do not match, as co-eluting peptides are common in DIA analyses. Each isolation window is scored against every library

spectrum with an appropriate precursor mass, allowing identification of multiple peptides from a single window.

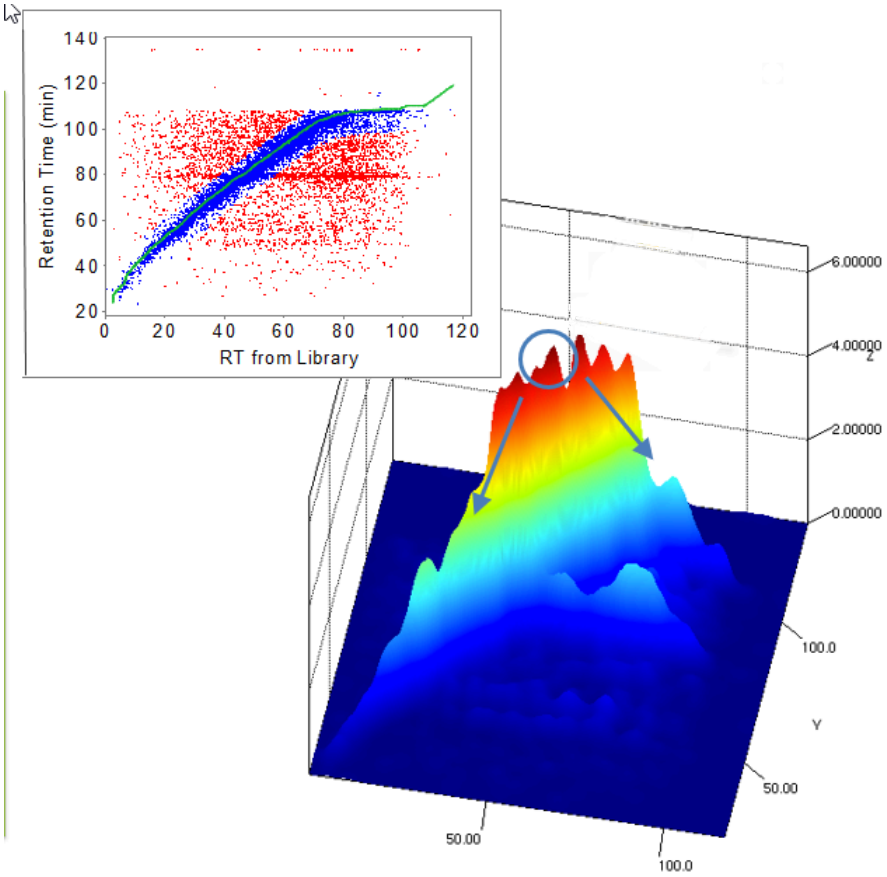
The primary score is similar to the HyperScore from X! Tandem, comprised of a weighted dot product of the intensities of the fragments in the acquired spectrum and the library spectrum multiplied by the factorial of the number of matching ions. Other scores, such as delta CN, fragment ion accuracy, precursor mass accuracy, and retention time accuracy are also calculated.

Retention Time Alignment

RT Alignment of Individual Samples

Conceptually, EncyclopeDIA places all identified peptides on a plot with the library RT on one axis and the experimental RT on the other. Kernel density estimates of the coordinates are constructed, and the resulting distribution is placed onto a 1000 X 1000 point grid. Using the grid points, the program traces the optimal fit using a ridge walking algorithm, starting at the point with the maximal intensity estimate, and walking the highest trajectory following the majority of peptides.

Figure 2-8: The Retention Time Alignment Algorithm in EncyclopeDIA [6]



Experiment-wide RT Alignment

For each peptide that is identified by the experiment-wide Percolator analysis, the best-scoring match is chosen from among the matches for that peptide in any of the individual samples. The sample that contains the largest number of these best-scoring peptide matches is selected as the reference sample, and the other samples are aligned to it. For peptides in each sample with RT values that are deemed outliers, an attempt is made to reassess the matches. Finally, peptides in each sample are assigned RT values corresponding to the RT in the reference sample, as are peptides that were identified globally but not in a specific sample.

Peptide Quantification

Quantification in Scaffold DIA is based entirely on the intensities of the fragment ions in the chromatograms. Precursor intensity is not considered.

[6]Searle, The Nuts and Bolts of EncyclopeDIA, Proteome Software Users Group, ASMS Conference 2017

Chapter 3: Loading data into Scaffold DIA

Scaffold DIA imports and analyzes data from various types of Mass Spectrometers. Data loading is conveniently performed through a dialog that guides the user in setting all of the parameters needed for data analysis and protein identification.

Raw files supported by Scaffold DIA

Scaffold DIA reads and processes raw DIA MS data in a variety of formats.

Scaffold DIA supports various vendors' data file formats and open file formats. Figure 3-1 shows the list of supported raw formats.

Figure 3-1: Scaffold DIA supported raw formats

Vendors' Formats	
Company	Format
AB/SCIEX	*.wiff
Thermo	*.raw
Agilent	.d directory
Open Formats	
	Format
HUPO Proteomics Standards Initiative mzML	*.mzML

Although only these file formats have been tested, Scaffold DIA should analyze windowed DIA data in any format supported by ProteoWizard's MSConvert program. Please contact Proteome Software support for further information about use with other file types.

Notes:

Many AB/SCIEX instruments produce a .WIFF.SCAN file for each .WIFF file. In this case, the user will only specify the loading of the .WIFF files, but Scaffold DIA needs both, and requires that each .WIFF.SCAN file be located in the same directory as its corresponding .WIFF file.

Thermo .RAW files are supported.

For Agilent instruments, select the entire *.d directory. Only fixed window sizes are supported, and it may be necessary to specify the window size in the loading dialog.

Vendor's proprietary format raw data files can be processed only on Windows operating systems.

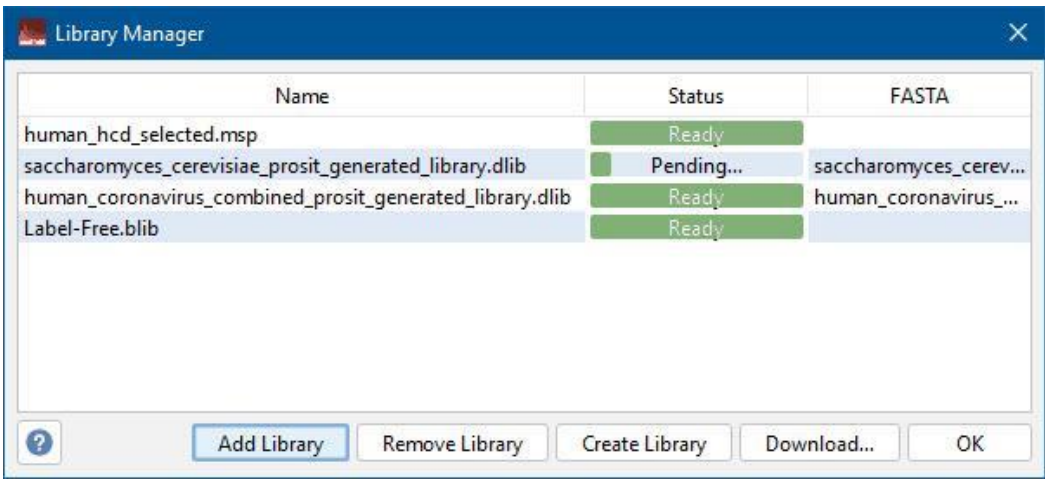
Command Line Interface

Scaffold DIA offers a command-line interface to enable use in server or automated environments. It requires only a workflow file and raw data, and will produce a ready-to-use .SDIA file. To use the CLI, run the ScaffoldDIABatch.exe executable (ScaffoldDIABatch on Mac/Unix) in the program's installation folder. Use the --help option to see a list of available options for the CLI. For further information, contact Proteome Software support.

The Library Manager

In order to perform an analysis, Scaffold DIA needs libraries to search. The Library Manager provides a convenient means to download, create, maintain and select libraries. It is especially useful for curating a list of libraries that will be used often in creating chromatogram libraries or processing small numbers of samples. The library manager also provides the ability to select multiple libraries and to simplify FASTA selection.

Figure 3-2: The Library Manager

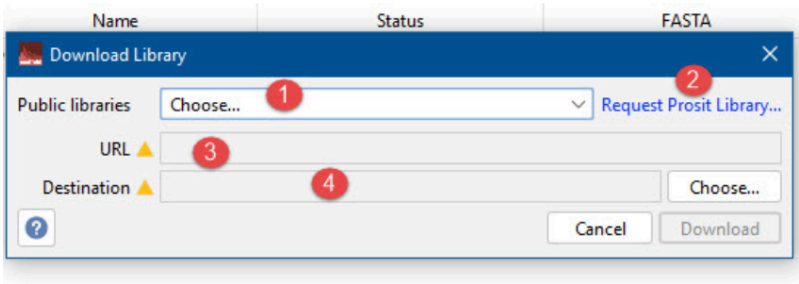


It can be launched either by selecting **Open Library Manager...** from the **File** menu. clicking on the Library Manager icon in the toolbar or through the Workflow dialog by selecting Library Manager from the dropdown that appears when the user clicks **Choose...** in the Reference File selection box.

The first time the Library Manager is opened, the list will be empty. Libraries may be added in a number of ways:

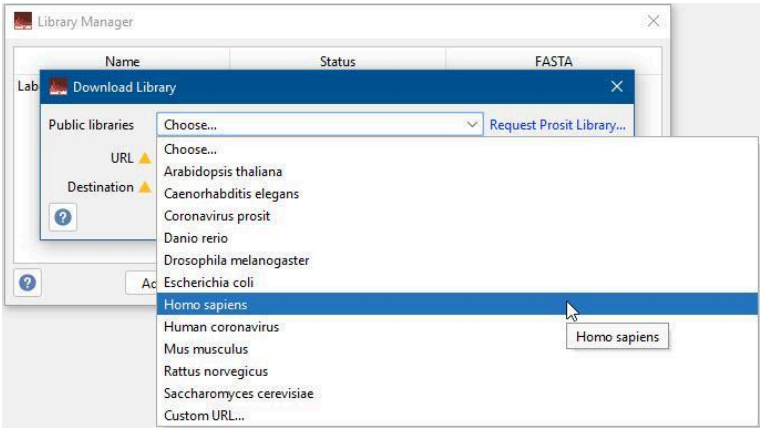
- **Add Library** - clicking this button opens a file chooser to allow selection of a library file. Files of any supported library type (except FASTA) may be added to the Library Manager. After the library has been added, right-clicking on it will allow the user to associate a FASTA to be used with the library during searches or open the folder containing the file.
- **Create Library** - Launches the dialog to create a New Chromatogram Library. When the library has been created, it is automatically added to the Library Manager, and the user may associate a FASTA file to it.
- **Download -** This button opens a dialog to allow the user to download prepared libraries from Proteome Software's collection of Prosit libraries, or from a different website via the Custom URL option.

Figure 3-3: The Download Library Dialog



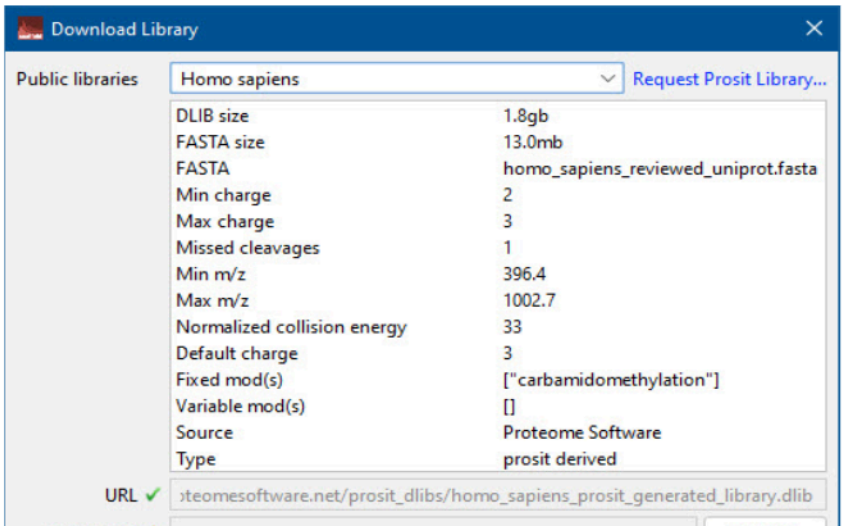
1. Clicking **Choose...** displays a list of Prosit libraries available from the Proteome Software Prosit library repository. To download a library from another public site, select Custom URL.. from the list.

Figure 3-4: The list of publicly available Prosit libraries



When a library is selected for download, information pertaining to the library is displayed. It is important to review this information to ensure that the collision energy, modifications and other parameters used in creation of the library are appropriate for the user's experiment.

Figure 3-5: Library information



2. If an appropriate Prosit library is not found in the download list, the user may submit a FASTA file and request that Proteome Software creates a custom Prosit library. Clicking the **Request Prosit Library...** link opens a request form. Please allow up to 48 hours for creation of the library.
3. To download a library from another source, select Custom URL... from the dropdown and paste the URL of the library into the URL field.
4. When a library is downloaded, it will be stored on the user's computer in the location specified in the Destination field. The location will be tracked in the Library Manager. The file location appears in a tooltip, and can be accessed by right-clicking on the library and selecting **Show in File Browser**. If the file is moved or deleted, the library entry in the table will show an error condition.When a library is downloaded from the Proteome Software Prosit repository, its corresponding FASTA is also downloaded to the same location, and the FASTA is automatically associated with the library.

Other Actions available through the Library Manager include:

- **Remove Library -** Selecting a library in the table and clicking Remove Library removes the library from the table.
- **Associate a FASTA to a library** - Right-clicking on a library in the table brings up a menu that provides an option to associate a FASTA database with the library. Once the FASTA is associated with the library, it will automatically be selected for use whenever the library is selected.
- **Reorder Libraries in the Library Manager table** - The columns may be sorted by clicking on the column headers. When column sorting is not in effect, libraries may be arranged by dragging and dropping. Note that the library must be selected with a click and then clicked again to activate the dragging operation.

Selecting Reference Libraries

- **Selecting a library from the Library Manager** - Click on the library to select it and click OK.
- **Selecting a Pending Library** - While a library (and its associated FASTA) is downloading, its status is shown as "Pending." The library may be selected while it is pending. The download status will be

shown in the Workflow dialog. The dialog remains active and other parameters may be specified, but the Load Data button will not be active until the download completes.

- **Selecting Multiple Libraries** - Click multiple libraries while holding the CTRL key, or select a group of libraries adjacent in the table while holding the SHIFT key, then click OK. The libraries will be combined. Libraries of different types may be selected together. All of the modifications contained in any of the libraries will appear in the Modification list, but note that each modification will only be recognized in peptide matches obtained from the libraries which contained that modification. If the selected libraries are associated with different FASTA files, a combined FASTA will be created and the user will be prompted to select the location to which it should be written.
- **Selecting a Library File** - It is also possible to select a library directly using a file chooser. In some cases, the user may not wish to add a library to the manager, and instead to select the file directly. For example, it may not be useful to put chromatogram libraries created for specific experiments into the Library Manager. An option is offered, however, to have library files selected in this way added automatically to the Library Manager. To activate this option, select Edit>>Preferences>>User Interface and in the Library Manager section, check the box labeled **Automatically add selected files to Library Manager**.

Creating a New Experiment

To create a new experiment the user may either:

- select ***File > New***
- click the **New** icon located in the toolbar below the main menu in the Scaffold DIA window

Figure 3-6: New Experiment icon

- click on the **New** button appearing in the “Welcome to Scaffold DIA” dialog when the program is launched.

When a new experiment is created, the Workflow Dialog appears to allow the user to enter the parameters for the protein search and to select the raw data files to be analyzed.

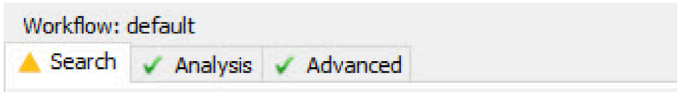
Workflow Dialog

The Workflow dialog allows the user to specify the type of experiment to be analyzed and to provide the necessary program parameters and files.

Warning Icon

Default values are provided for many of the parameters, but some require user input. If a tab is missing required information, a small warning icon appears near the tab name. A yellow triangle alerts the user that some of the required options have not been specified. When all of the required options have been defined, a green check replaces the triangle. Similar icons are used to indicate specific parameters which require user specification. Note that it is important to check the other parameters even if a default value is provided or the parameter is not required.

Figure 3-7: Tab warning icons



Workflow Tabs

Three Workflow tabs are provided for specification of different types of parameters:

- The Search Tab - options on this tab define the type of analysis to be performed, and the search parameters to be used. Input files are also specified through this tab. Parameters set in this tab are used during the analysis and may not be changed once the data has been loaded.
- The Analysis Tab - these options define how the results will be displayed when the analysis is complete. These include protein-level thresholds and clustering options, which may be adjusted in the display without reanalyzing the data.
- The Advanced Tab - this tab contains parameters that are used internally during analysis and that rarely need to be changed. Options selected in this tab are saved as defaults for future analyses.

In most cases, the user only needs to complete the Search tab. Options on the Analysis tab may be adjusted after data is loaded, and options on the Advanced tab generally may be set once and need not be changed.

- **Workflow Buttons** - The buttons located at the bottom of the Workflow Dialog allow the user to either save the current selection of options and parameters listed in the pane to a named workflow file or to load options from a previously saved workflow file.
  - *Load Workflow From File*- This selection opens a file browser to locate the WORKFLOW file to be loaded.
  - *Save Workflow* - After all parameters and options have been properly defined, the user has the option to save the information to a WORKFLOW file. Clicking the button opens a file browser so the user can assign a name and save the WORKFLOW file to a convenient directory. File selections are not saved in the WORKFLOW file.

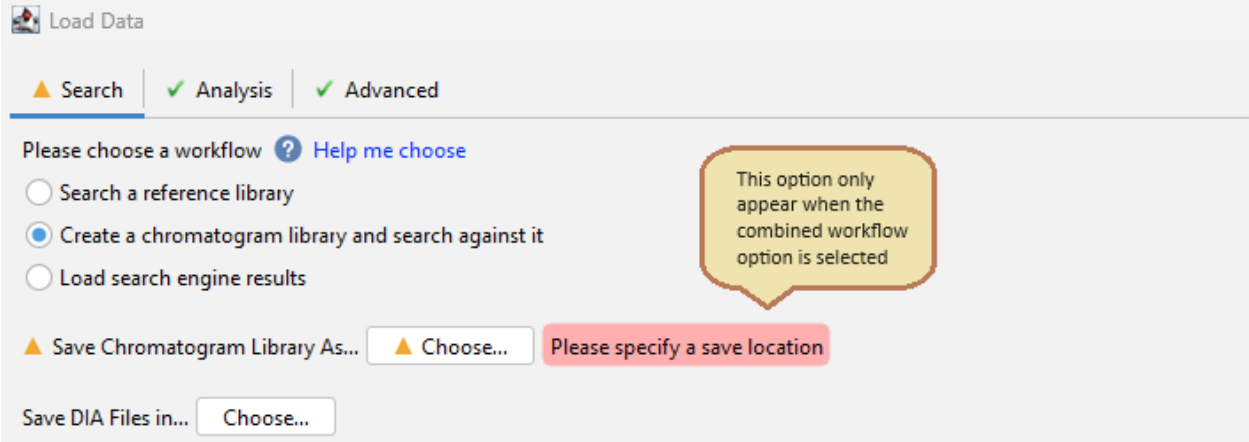
A message appears at the upper left corner indicating the name of the currently loaded workflow.

Depending on the amount of raw data submitted and the type of analysis, the loading and analysis phase might take considerable time. Once processing is complete, the Samples View will open to show the list of identified protein groups meeting the specified FDR thresholds. Note that if the Load Data operation is canceled while ProteoWizard's MSConvert is running, it may be necessary to terminate the MSConvert process through the Task Manager.

The Search Tab

This tab allows the user to specify the type of experiment to be analyzed and then adjusts to prompt for the necessary parameters and files.

Figure 3-8: Specifying the type of analysis



The first option which must be selected is the type of analysis to be performed. When a different analysis type is selected, the remaining content of the Search Tab adjusts to provide the necessary parameters for the selected option.

There are threeoptions:

- **Search a reference library** - DIA data files are searched against one or more existing reference libraries. Reference libraries may be BLIB files exported from Skyline or Scaffold, or may be ELIB files created previously by Scaffold DIA. For more information, see New Chromatogram Library.

  A particularly effective option is to search against a Prosit[7] library. Prosit is a deep learning algorithm developed by the Wilhelm and Kuster group to predict fragmentation and retention times for peptides. Proteome Software offers Prosit libraries for a number of organisms for download from https://support.proteomesoftware.com/hc/en-us/articles/360035151172-Prosit-Derived-Spectral-Libraries-for-Scaffold-DIA-Searches. If a library is needed for an organism that is not found in the dropdown list, Proteome Software can create one based on a user-provided organism-specific FASTA. Searching a Prosit library can provide extensive coverage but can be slow, so for large experiments it may be better to use a Prosit search to create a chromatogram library.

  Finally, it is possible to conduct a search against theoretical spectra obtained from an in-silico digest of sequences in a FASTA database. This option is slow and should only be used for processing a very small number of samples.

  Searching multiple libraries - Using the library manager, it is possible to select more than one library for a search. This allows combining libraries for different species, adding a contaminants library, or combining different types of libraries.

- **Create a chromatogram library and search against it** - this type of analysis allows the user to create a chromatogram library using (pooled) reference samples and search experimental DIA

[7]"Gessulat, Schmidt et al. 2019" DOI 10.1038/s41592-019-0426-7.

samples against this library in a single run. In this type of analysis, two searches are performed. First, a specified set of DIA reference files are searched against an existing library or a FASTA database to produce an ELIB file. Then, another set of DIA data files are searched against this ELIB. For details, see Combined Reference and Data Search Workflow.

When this option is selected, two sets of Search Parameters are required, so the user must complete both the **Set Reference Search Parameters** section and the **Set Data Search Parameters** section.

- **Load search engine results -** load results from supported search engines: Spectronaut, DIA-NN or PaSER. See Load Search Engine Results.

New Chromatogram Library

It is also possible to create a reference library from reference samples and save it for future analyses without running a data search. This is accomplished through the **File>New Chromatogram Library...** menu option or through The Library Manager.

For advice on selecting the appropriate search type option for a specific experiment, see How to choose the appropriate search type option.

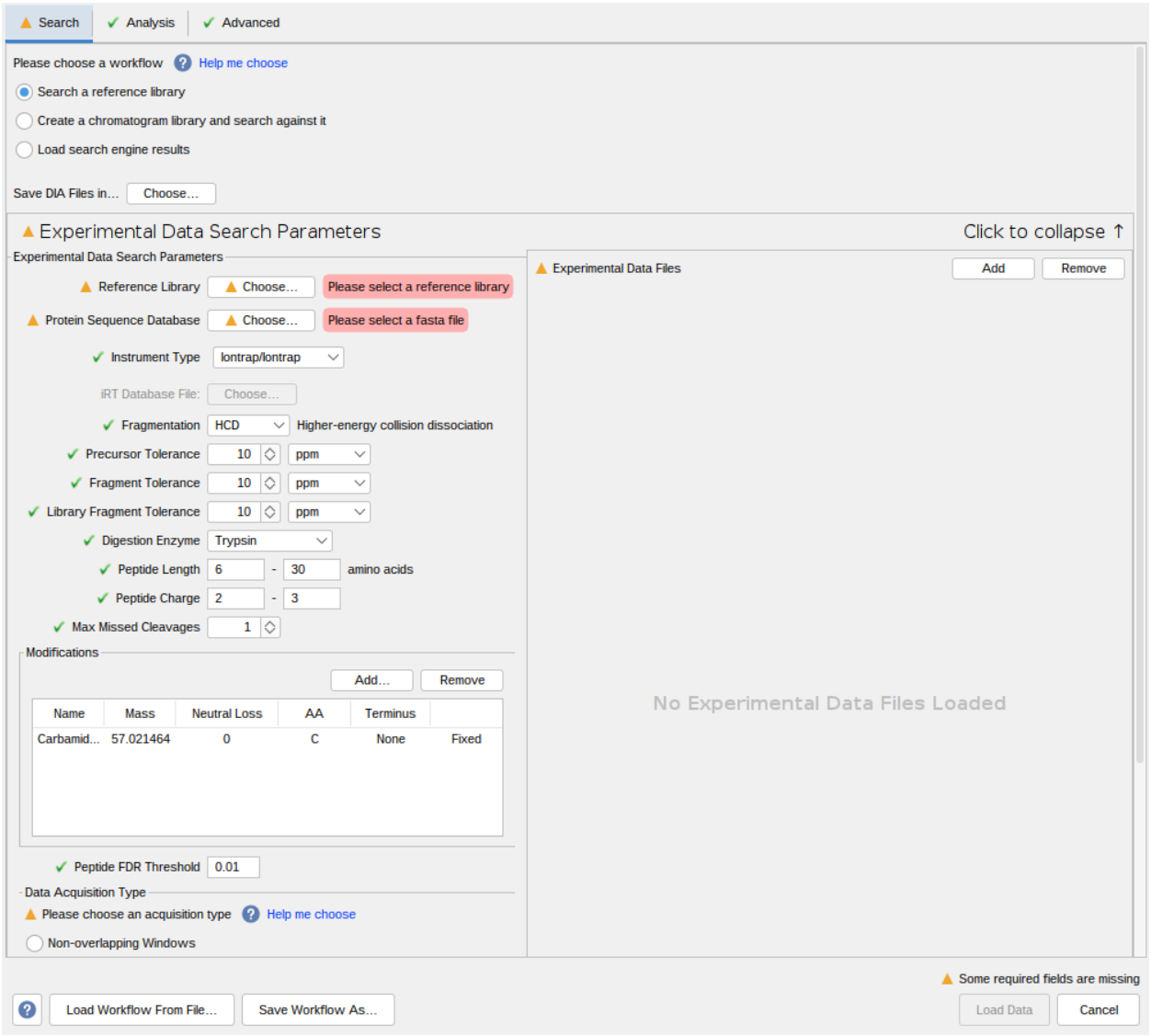
Saving Intermediate Files

Save Chromatogram Library As...Save Chromatogram Library As... - This required parameter only appears when the user has selected “Create a chromatogram library and search against it” or File>New Chromatogram Library... The file chooser allows the user to specify the file name and location to be used in storing the reference library file (ELIB) that will be created by the analysis. This library may be reused in subsequent searches.

Save DIA Files in...Save DIA Files in...- This is an optional parameter. When Scaffold DIA processes raw data files, it creates an intermediate file with extension .DIA. By default, these files are written to a temporary directory and deleted after the data is loaded, but if the user selects a location with this control, Scaffold DIA will write the .DIA files to this location and will not delete them. This allows the user to reprocess the data without reconverting the raw files.

Search a reference library

Figure 3-9: Setting Parameters for a Search of a Prosit or Existing Library



Experimental Data Search Parameters

- **Reference Library** - select the appropriate radio button, then click the Choose... button to choose one or more existing libraries (BLIB), (MSP), (TraML), (SPTXT) or (ELIB) through the library manager or select a library or a FASTA file using the Choose Library file... option (see Experimental Data Search Workflow) to be used in the reference search. Note that certain SPTXT files may not be supported by Scaffold DIA.
- **Protein Sequence Database** - The FASTA database which will provide protein sequences for linking identified peptides to proteins. Libraries in the Library Manager may have FASTA files associated

with them. When a library with an associated FASTA is selected, that FASTA is chosen as the Protein Sequence Database. If a FASTA database has been specified above, the same FASTA file will be used.

- **Instrument Type** - Choose an instrument type to automatically set parameters for EncyclopeDIA. Suggested tolerances are also updated with this choice.
- **iRT Database File** - Optional file obtained from Skyline to anchor retention time alignment. This option is only available when performing a library search with a BLIB file. An iRTDB file should be selected using the "Choose..." button.
- **Fragmentation - CID/HCD/ETD** - determines which ions will be used in identification and quantification. HCD uses y ions only for identification when performing a FASTA search if Trypsin is the specified digestion enzyme or b and y otherwise, but uses both b and y ions for quantification; CID uses b and y ions for both identification and quantification; and ETD uses c and z ions for both.
- **Precursor Tolerance** - Selects the m/z tolerance and the units in which the precursor tolerance is expressed.
- **Fragment Tolerance** - Selects the m/z tolerance used in comparing the fragment ions to the searched library, along with the units in which this tolerance is expressed. Note that the fragment tolerance for DIA analysis may be smaller than for DDA analysis.
- **Library Fragment Tolerance** - Active only when a library search is selected. Allows the user to select a different fragment tolerance for the search that created the library.
- **Digestion Enzyme** - Enzyme used to digest proteins. If the correct enzyme is not listed in the dropdown, it is possible to read in a custom enzyme specification from a workflow file. For assistance, please contact Proteome Software Support. If Non-specific is selected, the maximum number of missed cleavages should be set to one less than the maximum peptide length.
- **Peptide Length** - Specifies the size of peptides to be considered.
- **Peptide Charge** - Range of charge states to be considered in the search.
- **Max Missed Cleavages** - The maximum number of missed cleavages (based on the specified digestion enzyme) that is permitted in identified peptides.
- **Modifications** - Modifications that will be considered in the search.

  In a FASTA search, modifications must be selected from the **Add Modifications** dropdown, and may be removed from the list by selecting them and clicking "Remove Selected". Only the modifications selected here will be considered. Additional modifications may be added to the dropdown from the included Unimod file by clicking the Unimod button.

  Custom modifications may be added by selecting **Add**, then clicking the **Create...** button at the bottom of the resulting dialog. Remember to "Add" the custom modification you have just created.

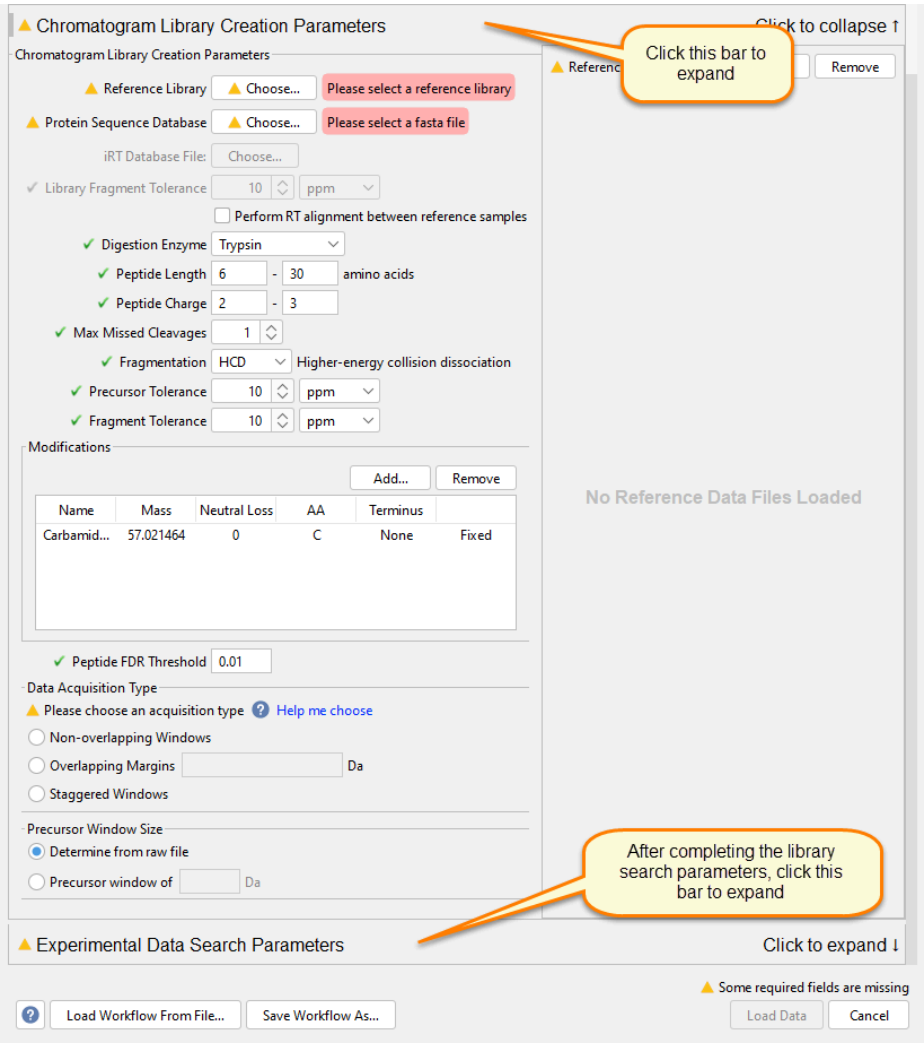
  When searching against a peptide or chromatogram library, any modifications in identified library spectra are retained in the search. Scaffold DIA makes an effort to extract the modifications in the library and to display them in the Modifications pane. The user may wish to use the Edit button to edit the names, neutral losses, terminal statuses and fixed/variable modification designations so that the modifications may be properly annotated in the resulting Scaffold DIAfile. It is not possible to add or remove modifications during a library search.

Note that currently Scaffold DIA will only display one modification per amino acid from a library in the Modifications pane. If the library contains spectra with different modifications on the same amino acid, these modifications will all be used in the search and may appear with generic names in the search results.

- **Peptide FDR Threshold** - The FDR threshold specified to Percolator. **The Peptide FDR may not be changed after the search**.
- **Data Acquisition Type** - Specifies whether the data was captured in **Non-overlapping Windows,** windows with **Overlapping Margins** or **Staggered Windows** (see How to Choose the Appropriate Data Acquisition Mode
- **Experimental Data Samples**- This pane to the right allows the user to select the data files to be searched. See Files supported by Scaffold DIA for a comprehensive list of supported file formats.

Create a chromatogram library and search it

Figure 3-10: Setting Parameters for Reference Library Creation and Search



Reference Library Creation Parameters

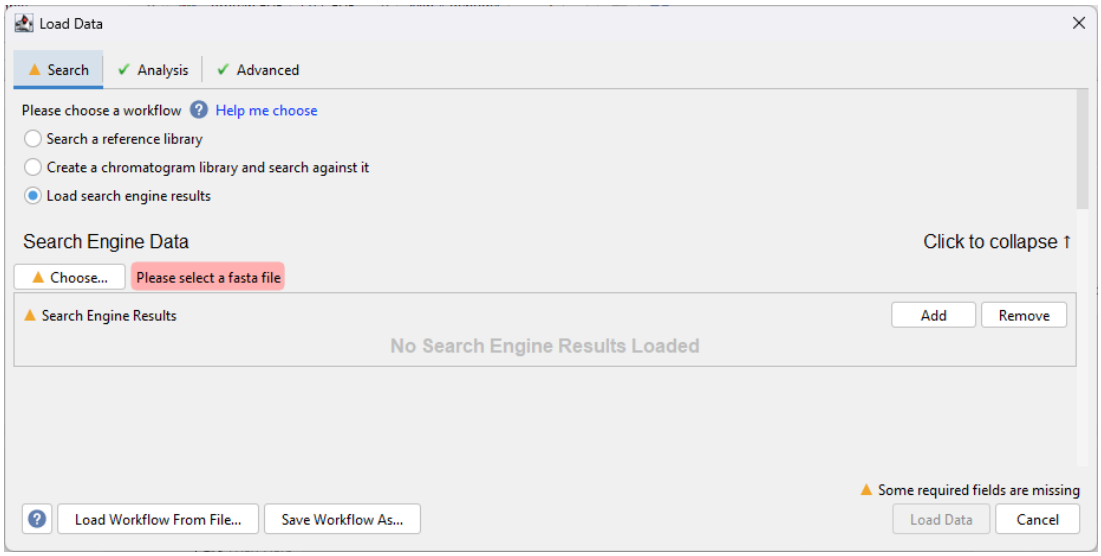
- **Reference Library** - select the appropriate radio button, then click Choose... to select an existing spectral (BLIB) or chromatogram (ELIB) library from the Library Manager or a library or FASTA file using the File Browser (see Scaffold DIA Workflows).
- **iRT Database File** - Optional file obtained from Skyline to anchor retention time alignment. This option is only available when performing a library search. An iRTDB file should be selected using the "Choose..." button.
- **Library Fragment Tolerance** - Active only when a library search is selected. Allows the user to select a different fragment tolerance for the search that created the library.
- **Perform RT alignment between reference samples** - this box should be checked if the reference library is being constructed from replicates, which should be aligned, but not if it is being constructed from fractions.
- **Peptide Length** - Specifies the size of peptides to be considered.
- **Peptide Charge** - Range of charge states to be considered in the search.
- **Max Missed Cleavages** - The maximum number of missed cleavages (based on the specified digestion enzyme) that is permitted in identified peptides.
- **Fragmentation - CID/HCD/ETD** - determines which ions will be used in identification and quantification.
- **Precursor Tolerance** - Selects the m/z tolerance and the units in which the precursor tolerance is expressed.
- **Fragment Tolerance** - Selects the m/z tolerance used in comparing the fragment ions to the searched library, along with the units in which this tolerance is expressed. Note that the fragment tolerance for DIA analysis may be smaller than for DDA analysis.
- **Modifications** - Modifications that will be considered in the search. Modifications must be selected from the **Add Modifications** dropdown, and may be removed from the list by selecting them and clicking "Remove Selected". In a FASTA search, only the modifications selected here will be considered. When searching against a spectral or chromatogram library, any modifications in identified library spectra are retained, however **Skyline exports only the nominal mass of modifications in its BLIB exports, so all modifications should be specified with accurate mass, even if they have been specified in the Skyline analysis.**

  Custom modifications may be added by selecting **Add**, then clicking the **Create...** button at the bottom of the resulting dialog. Remember to "Add" the custom modification you have just created.
- **Peptide FDR Threshold** - The FDR threshold specified to Percolator. **The Peptide FDR may not be changed after the search**.
- **Data Acquisition Type** - Specifies whether the data was captured in **Non-overlapping Windows,** windows with **Overlapping Margins** or **Staggered Windows** (see How to Choose the Appropriate Data Acquisition Mode.

- **Reference Data Samples** - This pane to the right allows the user to select the data files to be searched in order to create the reference chromatogram library. See Combined Reference and Data Search Workflow.

Load Search Engine Results

Figure 3.10.2 Loading search engine results



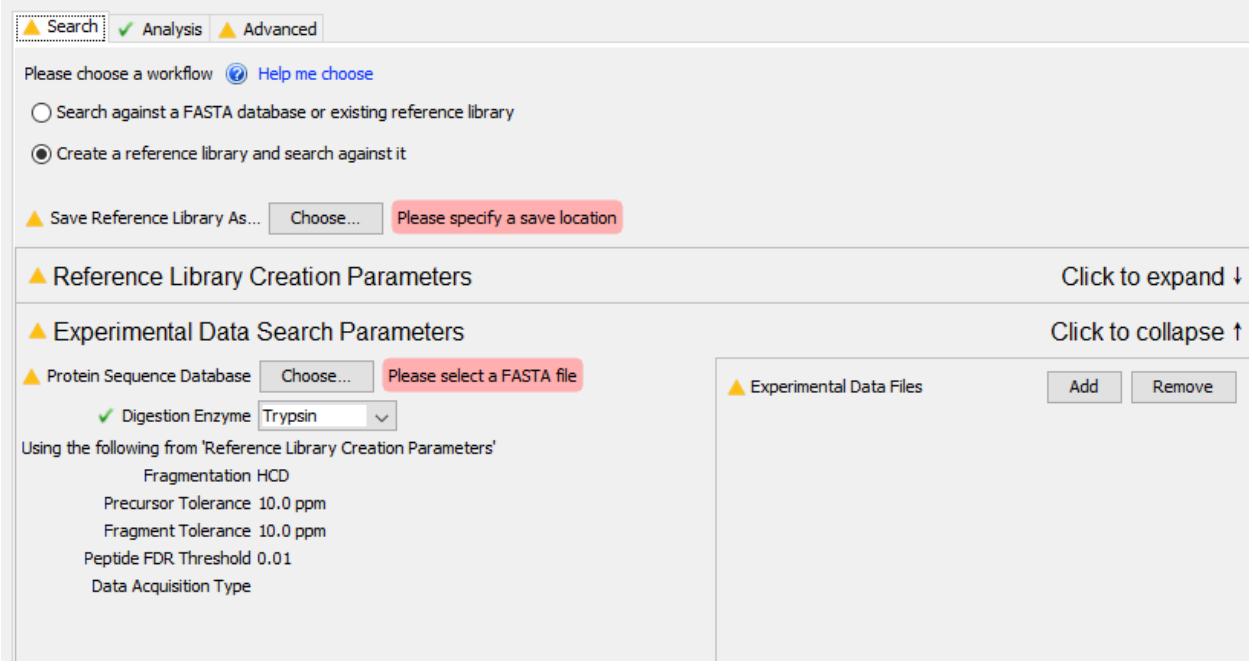
FASTA File - Select a FASTA file containing the desired protein names and sequences

Search Engine Results - Selected results from Spectronaut, DIA-NN, or PaSER that have been exported in as tab-separated (*tsv*) or comma-separated (*csv*) results. The following table lists the required and recommended columns to include in your export

	Required Columns	**Recommended Columns**
Spectronaut	PG.ProteinAccessions R.Label EG.ModifiedSequence EG.TotalQuantity FG.Charge	FG.PrecMz EG.IsDecoy EG.PEP EG.ApexRT F.MeasuredMz F.PeakHeight (or F.PeakArea) F.FrgIon (or F.FrgType + F.FrgNum) F.ExcludedFromQuantification
DIA-NN	Protein.Ids Run File.Name Modified.Sequence Precursor.Charge Precursor.Translated (or Precursor.Normalized or Precursor.Quantity)	CScore Decoy.CScore PEP RT RT.Start RT.Stop Fragment.Info Fragment.Quant.Corrected
PaSER	*same as DIA-NN*	*same as DIA-NN*

The Search Tab

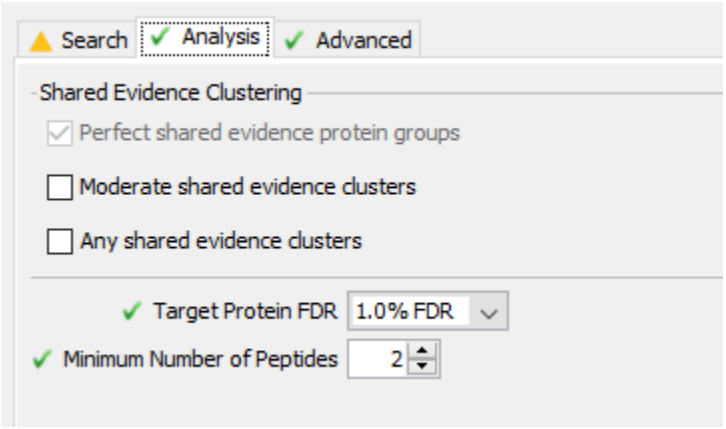
Figure 3-11: Data Search Parameters in a combined Reference and Data Search Analysis



- **Protein Sequence Database** - The FASTA database which will provide protein sequences for display of identified proteins. This is not adjustable, as the FASTA database that was specified in the creation of the library is always used.
- **Digestion Enzyme** - Enzyme used to digest proteins. If the correct enzyme is not listed in the dropdown, it is possible to read in a custom enzyme specification from a workflow file. For assistance, please contact Proteome Software Support. If Non-specific is selected, the maximum number of missed cleavages should be set to one less than the maximum peptide length.
- The remaining parameters are provided for informational purposes only, but must agree with the parameters set for the reference search.

The Analysis Tab

Figure 3-12: The Analysis tab



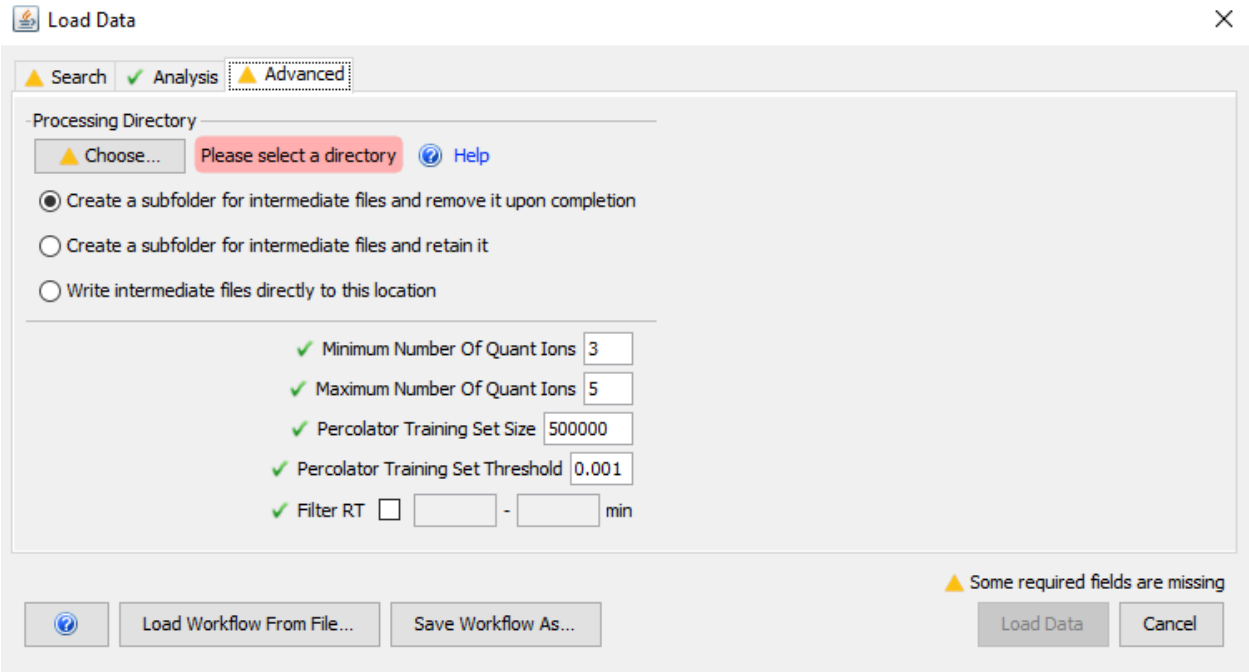
This tab allows the user to set properties of the experiment as it will be displayed when the search and analysis have been completed. These properties may be changed without reloading or researching the data.

- **Clustering** - Shared Evidence Clustering - more than one level may be selected. Protein clusters are displayed in a hierarchical manner in the Samples View table.
  - **Perfect Shared Evidence Clustering** is always performed. Proteins identified by identical peptide evidence are grouped together. This level of clustering may not be turned off.
  - **Moderate Shared Evidence** - clusters proteins which share a significant portion of their peptide evidence (see Appendix E: Shared Evidence Clustering Algorithm).
  - **Any Shared Evidence** - clusters proteins which share even a single peptide.
- **Target Protein FDR**- sets the protein FDR threshold to be applied when results are initially displayed. This threshold may be adjusted through the Protein FDR Threshold control.
- **Minimum Number of Peptides** - Sets the initial threshold for display of results. Only proteins identified by at least this number of peptides will be displayed. This threshold may be adjusted through the Minimum Number of Peptides Threshold control.

The Advanced Tab

This tab contains settings that are fundamental to the analysis and generally are not changed between experiments. The tab is marked as needing attention the first time the program is run, but once the Processing Directory and Processing Directory Policy parameters have been set, the new values are treated as default values and the user is not required to set them in subsequent runs.

Figure 3-13: The Advanced Tab



Processing Directory - A required parameter that specifies the directory to which Scaffold DIA (and EncyclopeDIA) will write its intermediate results files. Scaffold DIA writes many intermediate files during the analysis, some of which may be quite large, so it is important to choose a location with sufficient free storage space to contain these files.

Processing Directory Policy - Three options for dealing with intermediate processing files are offered. The default option is for the program to create a new subfolder in the specified processing directory to hold the intermediate files for a specific experiment and then to delete this subfolder when processing is completed. The second option also creates a subfolder for each experiment, but does not delete the subfolder after processing. This allows the user to retain the intermediate files, perhaps for debugging a problem or viewing EncyclopeDIA results directly. The third option is to write the intermediate files directly to the location specified, with no subfolder created. The files are not deleted on completion.

- **Minimum Number of Quant Ions** - The minimum number of quantitative fragment ions that a peptide must contain in order to be used for quantification. Note that it is not necessary that all of these fragments have non-zero intensity in every sample, so the number of quantitative fragments may vary in the Peptide Match table in the Proteins View. In library creation, this parameter determines the minimum number of ions meeting the quality threshold an entry must have in order for the entry to be saved into the library.

- **Maximum Number of Quant Ions** - The maximum number of fragment ions that will be used in quantifying a peptide.

Percolator Settings - These are settings that Percolator uses to construct the set of peptide-spectrum matches used for training its model. The default settings are

- **Percolator Training Set Size** - defaults to 500,000
- **Training Set FDR Threshold** - defaults to 0.001.

These default values have been experimentally determined by Proteome Software to give good results in a variety of experiments, but may be adjusted by the user if necessary for unusual data sets.

Filter Retention Time Range - This allows the user to restrict the retention time range to be analyzed by the program. **RT Filtering can only be applied when converting raw data files to .dia format.** The resulting .dia file will contain only the scans collected between the minimum and maximum retention times specified. If a user attempts to load previously created .dia files with RT Filtering selected, an error message is displayed and processing is not allowed to continue.

Workflow Files

The buttons located below the “Search Setup” pane allow the user to either save the current selection of options and parameters listed in the pane to a named workflow file or to retrieve them from an existing one.

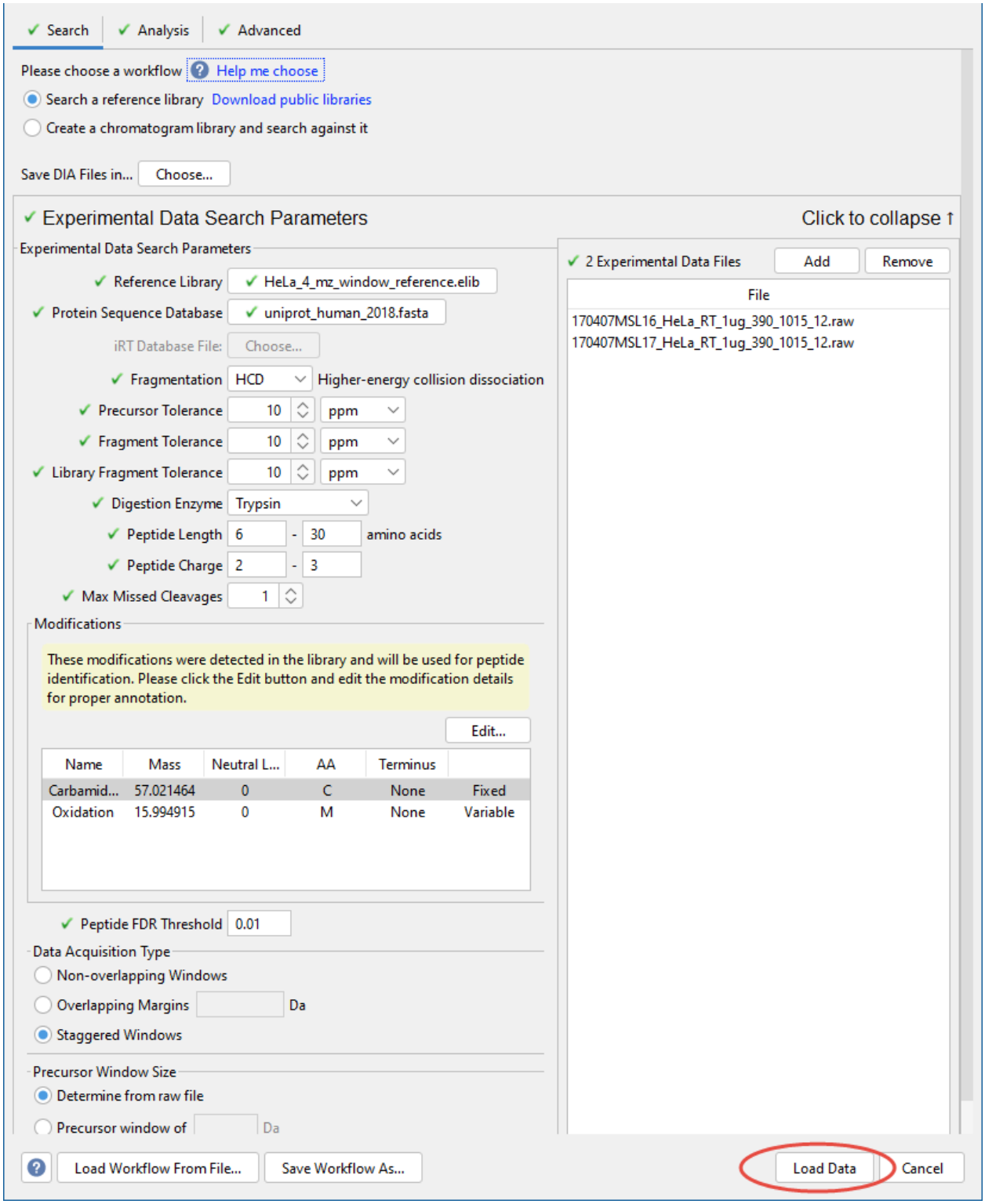
- *Load Workflow From File...*- This selection opens a file browser to locate the WORKFLOW file to be loaded.
- *Save Workflow* - After all parameters and options have been properly defined and when each tab shows a green check, the user has the option to save the information to a WORKFLOW file. Clicking the button opens a file browser so the user can assign a name and save the WORKFLOW file to a convenient directory.

A message appears at the upper left corner indicating the currently loaded workflow.

Beginning the Analysis

When all of the tabs are properly completed, the “Load Data” button located on the lower right corner of the dialog becomes available for use. Clicking this button begins loading and analysis of the data.

Figure 3-14: Starting the search



Resource-Intensive Search Warning Message

If the user has selected a workflow that will require a FASTA search that will be particularly resource-intensive, a warning message is displayed. In a DIA search, because it is impossible to use precursor mass as a screening tool, every possible peptide match must be scored. As a result, searching against a large FASTA, or with a number of variable modifications, or with many samples can cause the search space to expand dramatically.

It is recommended that the user use an appropriate FASTA (e.g. limited by taxonomy) and limit the number of variable modifications searched. When analyzing a large number of samples, it is also better to use a DDA-based library or to create a chromatogram library and to search the full data set against that.

File type created by Scaffold DIA

Scaffold DIA creates its own file type called SDIA, which stands for Scaffold DIA File. It is a lightweight, high performance SQL database file. Scaffold DIA provides the user the opportunity to query the experiment file using Structured Query Language (SQL) and to save these queries for future use. This direct access to the data structure gives the user of Scaffold DIA a unique capability to manipulate and analyze their data.

SDIA files may be opened with the free **Scaffold DIA Viewer**, allowing users to easily share their results with clients or collaborators. The Viewer retains all functionality of the fully licensed program except the ability to load and process new data.

How to choose the appropriate search type option

Two options are offered in the search tab:

Choose “Search a reference library” if:

- You have DIA data files for analysis, and want to search them using a Prosit library or a library you have exported from Scaffold or Skyline or a FASTA file. FASTA searches should be limited to only a few samples.
- You should also use this option if you have previously created a chromatogram library with Scaffold DIA and want to use that library for your search.

Choose “Create a chromatogram library and search against it” if:

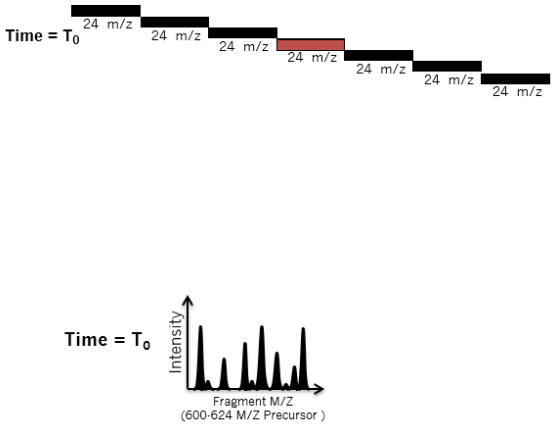
- You have collected two sets of DIA files: one set of reference data files for constructing a search library and another set which you wish to analyze by searching against this library. For more information about this workflow, see Combined Reference and Data Search Workflow. This workflow allows you to construct the library and use it in a search in a single run.
- Note that if you choose this option you may save the reference library for use in subsequent searches if you later collect more data samples or wish to analyze your data samples with different parameters.
- If you prefer to create a reference library but not perform an experimental search at the same time, use the New Chromatogram Library option in the File menu.

How to Choose the Appropriate Data Acquisition Mode

DIA data can be collected in various acquisition modes. Set the Acquisition Mode to match the window pattern in your data files. The three modes supported by Scaffold DIA are:

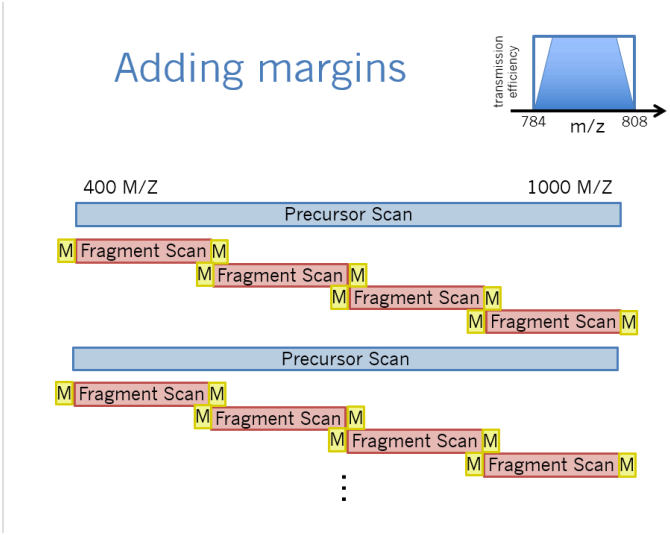
- Non-overlapping Windows - in this mode, each window is distinct.

Figure 3-15: Non-overlapping windows[8]



- Overlapping Margins - in this mode, a small margin is added to each window, overlapping adjacent windows slightly to offset edge effects which may cause poor data acquisition at the edges of windows. When this option is selected, the user must specify the size of the margins in Da.

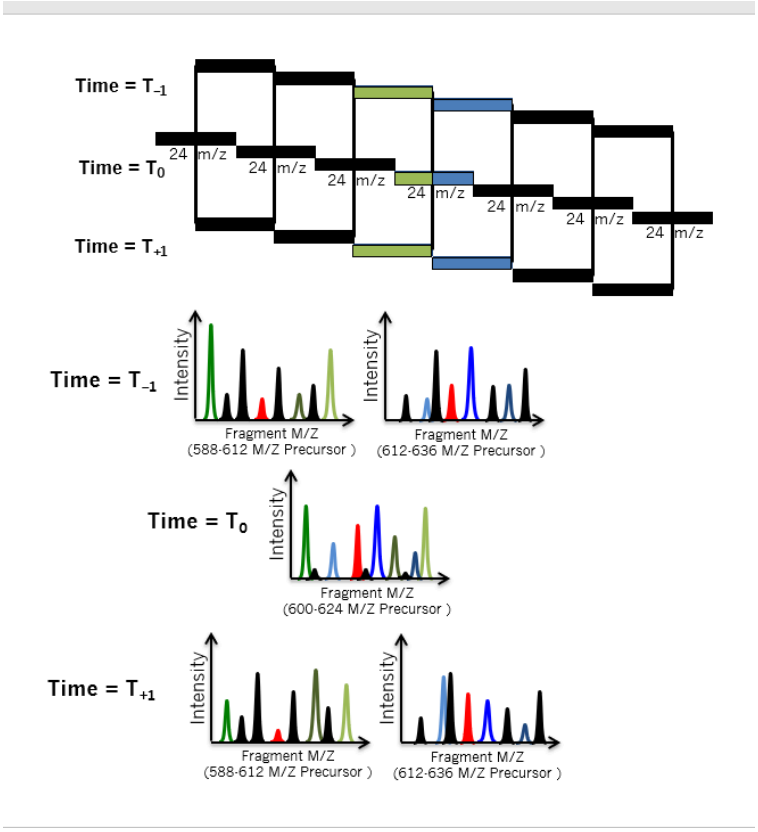
Figure 3-16: Windows with added margins



[8]Searle,B. Introduction to Data Independent Acquisition, ASMS 2017

- Staggered Windows - in this mode, windows overlap, with the window center offset in each duty cycle. When this mode is specified, Scaffold DIA instructs MSConvert to demultiplex the windows during raw file conversion.

Figure 3-17: Staggered Windows



Chapter 4: Scaffold DIA Main Window

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

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

- The Title bar
- The Main Menu CommandsMain Menu Commands
- The Tool-bar
- The Display Type Bar
- The Navigation Pane
- The Summarization Bar
- The Display Pane
- The FDR Dashboard

Title bar

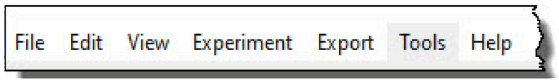
Figure 4-1: Title bar

The title bar at the top of the Scaffold DIA window always displays “Scaffold DIA”. 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 “- Scaffold DIA_Experiment” is appended. When a file is saved with a different name, the default name is replaced by the new file name. When a .SDIA file is opened, the title bar displays the file name.

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

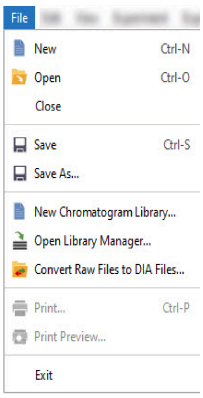
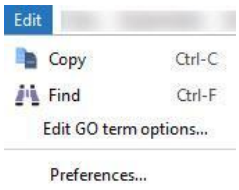
Main menu commands

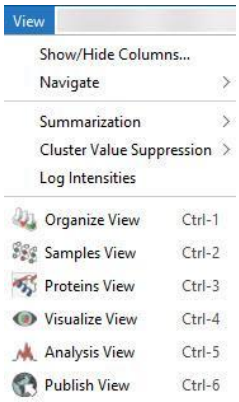
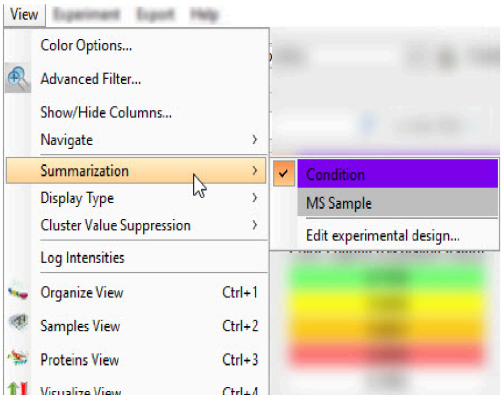
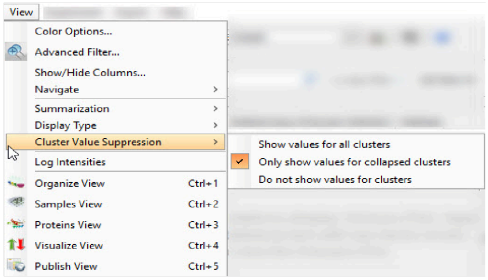
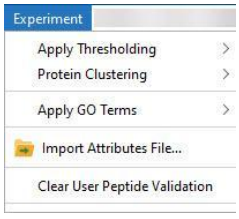
Figure 4-2: Main Menu

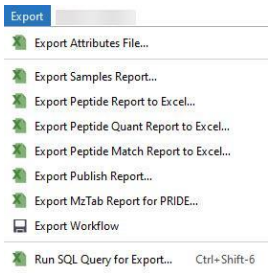
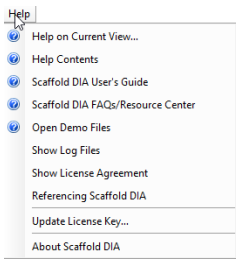


The Scaffold DIA main menu is organized in a standard Windows menu format with commands grouped into menus (File, Edit, View, Experiment, Export and Help) across the menu bar.

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 DIA. See Loading data into Scaffold. • **Open**—Opens a saved Scaffold DIA experiment file, *.SDIA, 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. • **New Chromatogram Library**—Opens a dialog to allow creation of a reference chromatogram library. • **Open Library Manager—**Opens the Library Manager to allow curation of library files, download of libraries, and association with FASTA files. • **Convert Raw Files to DIA Files**—Opens a dialog to allow conversion of Raw files to DIA files for later processing. • **Print**—Prints the current view. • **Print Preview**—Previews the current view with the option of printing the document. • **Exit**—Closes the Scaffold DIA 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**—Import GO information and select GO Terms to display • **Preferences**—Adjust program settings

Menu	Menu Commands
View 	• **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  • **Cluster Value Suppression**—Helps the user select the clustering values that need to be visible, see Cluster Value Suppression  • **Log Intensities** — Toggles the Log Intensities checkbox. Determines whether intensity values are shown as base 10 logarithms.
Experiment 	• **Apply Thresholding**— Applying Quant Thresholds • **Protein Clustering**—See Protein Grouping and Clustering • **Apply Go Terms**—Applies imported Gene Ontology Annotations to the Samples Table. To import GO databases see Edit GO Term Options. To select a specific taxonomy, see Apply GO terms • **Import Attributes File**— See Import Attributes File... • **Clear User Peptide Validation...**—This option reestablishes the initial selection of peptides at load. For more info see The Visualization Pane

Menu	Menu Commands
Export 	• **Export:** • **Attributes file...**—Generates a tab-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 tab-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 tab-delimited text file of the Peptide table for all proteins appearing in the Samples View, that can be opened and viewed in Excel. • **Peptide Quant Report to Excel…**—Generates a tab-delimited text file displaying quantitative information across samples or categories at the peptide level. • **Peptide Match Report**—Generates a tab-delimited text file containing information about the features that matched each peptide. • **Publish Report to Excel...**—Generates a tab-delimited text file of the information about the current experiment provided in the Publish View, that can be opened and viewed in Excel. • **Export MzTab Report for PRIDE…**—Generates an mzTab format file representing the results of the analysis. This file may be submitted to PRIDE for publication. • **Export Workflow**... — Exports a workflow file capturing the parameters used in the current experiment. • **Run SQL query for Export...**—Opens the SQL dialog box see SQL Export tab.
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 DIA User's Guide** —Opens the current Scaffold DIA User's Guide. • **Scaffold DIA FAQs/Resource Center**—Opens the user's default web browser to the Homepage of the Proteome Software's resource center. • **Open Demo Files**—Opens the folder where Scaffold DIA demo files are stored. The user can choose any of the pre-loaded files to test Scaffold DIA capabilities. • **Show Log Files**—Opens the folder containing Scaffold DIA error log and output_log files • **Show License Agreement**—Display the software's EULA • **Referencing Scaffold DIA**—Opens the Online Help that contains a sample of how to reference Scaffold DIA when publishing data analyzed with this application. • **How to Purchase** — (in Viewer mode) Opens the user's default web browser to the Purchase page of the Proteome Software web page: https://www.proteomesoftware.com/products/purchase. • **Update License Key...—**(in licensed mode) Opens a dialog where the user can paste a purchased license key that will allow full use of the application.

Menu	Menu Commands
	For more info see Scaffold Variants Licensing. When a full licensed application is in use this option is not visible. • **About Scaffold DIA**—Provides the release information for the current version of Scaffold DIA, license information, contact information for Proteome Software, Inc. It also reports information about the system where Scaffold DIA is installed, the amount of memory available to the software and the percentage of memory used by the application.

Edit

Some commands appearing under the Edit Menu:

Preferences

The main menu option: **Edit > Preferences**, opens the **Preferences** dialog which contains the following tabs:

- System
- User Interface
- Colors
- ProteoWizard
- Heatmap

System

This tab provides a number of options related to system settings:

- Memory Usage - allows the user to specify the amount of RAM to be allocated to Scaffold DIA. The more RAM allocated the better the program will perform. It is recommended that this be set to approximately 75% - 80% of the physical RAM available on the system.

- *The new memory setting will take effect only after the application has been closed and restarted.*
- *MAC OS, once Scaffold DIA Viewer is installed, does not allow the memory allocation to be reset unless the user is an administrator. A non-administrator user can update the memory only when reinstalling the software.*

- Number of Processors - allows selection of the maximum number of processors available to Scaffold DIA for threading computations. The default value is the maximum number of processors available on the system where the application is installed.

- Internet Settings - determines whether the program is permitted to connect to the internet. If this is allowed, an option is provided to specify an HTTP proxy server. Proxy servers may be used by an organization's IT departments to filter communications to and from the Internet. If a proxy is in use, the user needs to set the Proxy Server Name and Port Number.

  To determine whether there is a need to use proxy server settings, check how your web browser is connected to the internet.

User Interface

This tab offers options related to the function of the Scaffold DIA graphical user interface.

- Search Fields - a checkbox allows the user to specify whether or not search fields should be interpreted as regular expressions.
- Prompts - a button allows users to re-enable any messages that have been disabled by checking a box labeled “Do not show this message again” in any dialog.
- Views - Chooses the View that will open by default when files are loaded or when a Scaffold DIA file is opened.
- Identifiers - Selects the preferred format for displaying protein identifiers in the Proteins View and the units in which Retention Time should be displayed (seconds or minutes).
- Library Manager - Automatically add selected files to the Library Manager - When this box is checked, library files selected through the “Choose Library file...” option are added to the Library Manager table.

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.

ProteoWizard

ProteoWizard location - Allows the user to select the msconvert.exe file to be used for conversion of raw files to MZML format. MSConvert is a program in the ProteoWizard suite, and the executable file may be found in the folder into which ProteoWizard was installed. Generally, this will be in a folder named C:\ProteoWizard followed by the version number.

Heatmap

This tab contains a checkbox that specifies that the heatmap should be built without clustering of the columns. When this is checked, the columns will be the same as the columns in the Samples View. The rows, representing the proteins, are clustered as usual (see Heat map clustering).

View

Some items appearing in the View menu:

- **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
- **Cluster Value Suppression** - see Cluster Value Suppression
- **Log Intensities** - when this item is checked, intensity values in the Samples View are displayed as base 10 logarithms. When it is not checked the intensity values are not shown in log form. This selection coordinates with the Log Intensity checkbox.

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 Ontology Annotations

Many genes and proteins are annotated in a public data repository by Gene Ontology or GO terms or annotations. Scaffold DIA can annotate protein using the GO terms found in the GOA database downloaded from the EMBL-EBI website (http://www.ebi.ac.uk/GOA/) or 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.

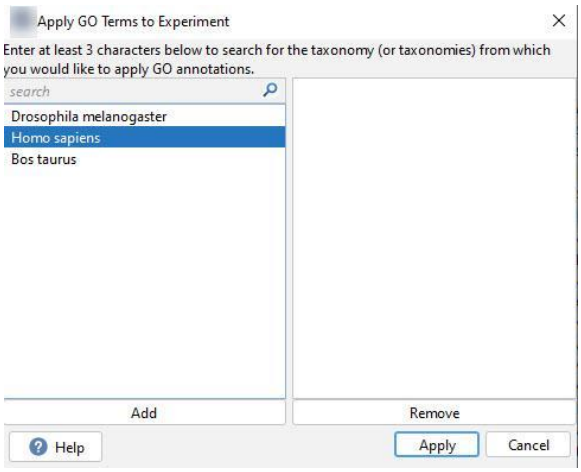
Scaffold DIA shows the high-level GO annotations as colored dots in the appropriate column added to the Samples table. Depending on the type of evidence used to define a GO term the colored dots will appear either as open or closed circles, see GO Terms - Open vs Closed Circles.

Apply GO terms

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

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 are selected, click Apply

Figure 4-4: Applying 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 DIA 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 opening the Tables Column Control and unselecting the GO Terms entry.

GO Terms - Open vs Closed Circles

Scaffold DIA 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 4-1 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*.

Table 4-1: GO Evidence Codes

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
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

The user should 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.

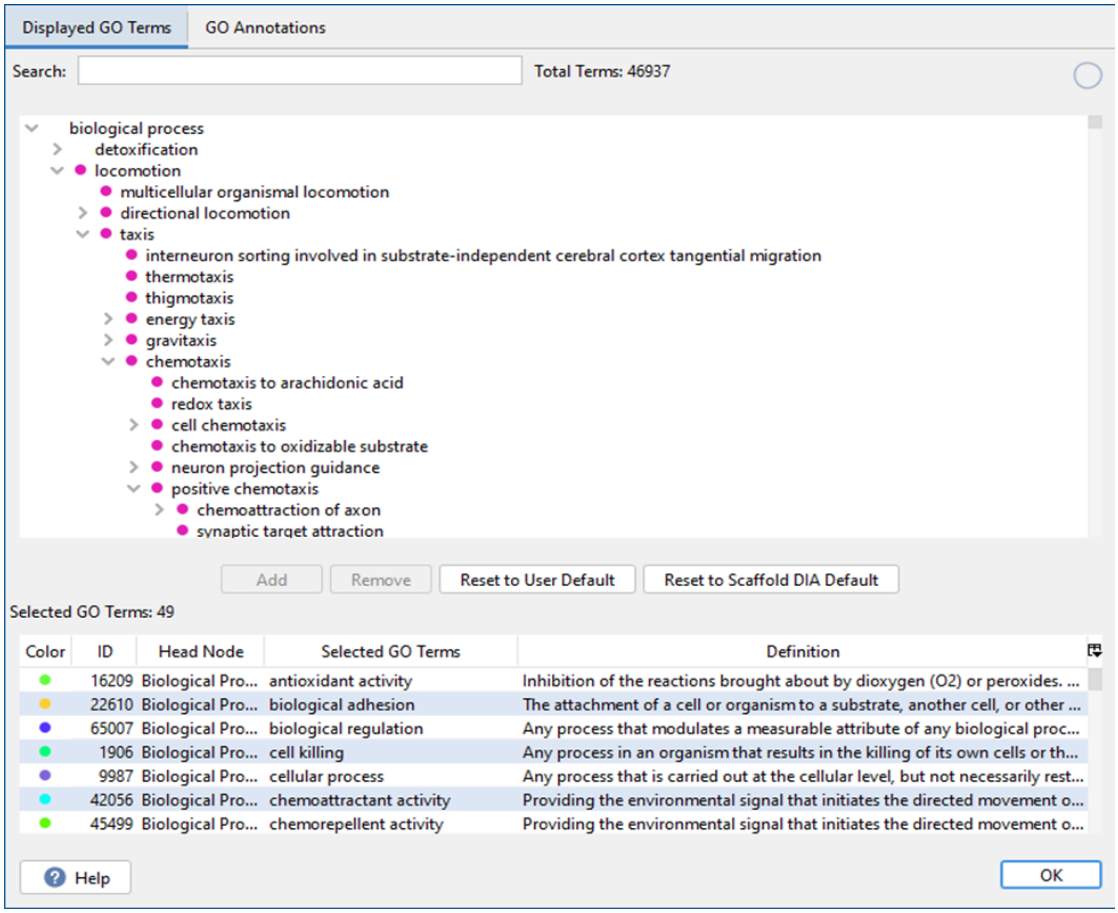
Edit GO Term Options

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

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 DIA as extra columns in the Samples Table.

Figure 4-5: Displayed GO Terms



The Tab window is divided into the following sections:

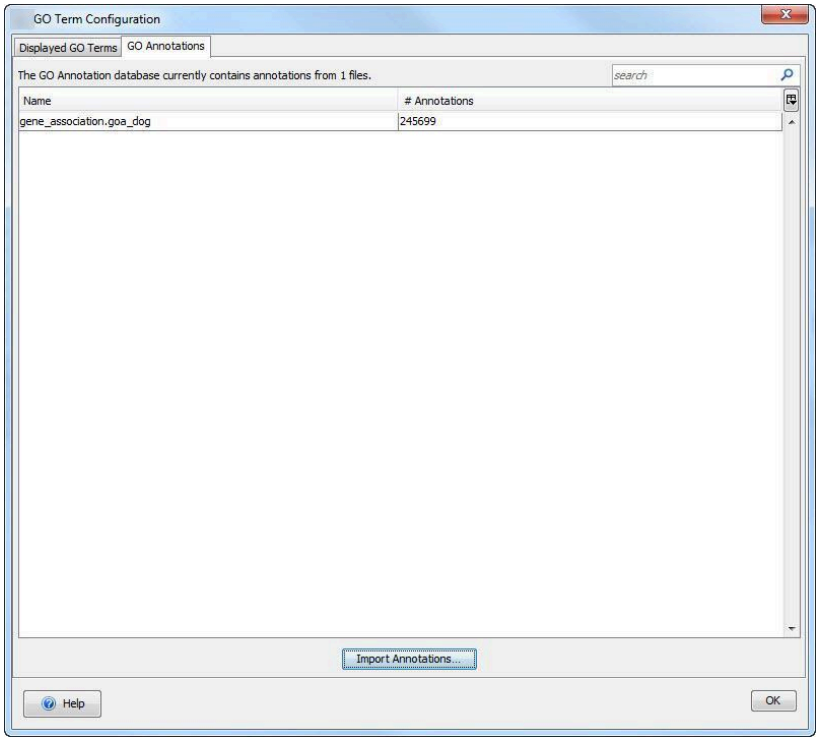
- Search Field - Searches terms available in the GO terms database loaded in Scaffold DIA
- GO Tree list - Hierarchical list of all the terms present in the loaded GO database
- Add and Remove GO terms - Provides tools for creating the custom Display list
- Display List - List of GO terms, selected by 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:

- If the Display List is not empty select all the rows and press delete.
- Search and select any GO term of interest from the loaded GO database either by typing a name in the Search Field or by selecting a row in the GO Tree List.
- 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.
- To remove terms from the Display List, select a term or group of terms to be discarded then click Remove.
- To save the current selections as User Defaults check the box Save displayed GO terms as user default.
- When a Scaffold DIA experiment is saved, the displayed GO terms are saved in the *SDIA file.
- When a new file is created, or when Scaffold DIA 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 DIA Default button.

The GO Annotations Tab

Figure 4-6: Go annotations tab

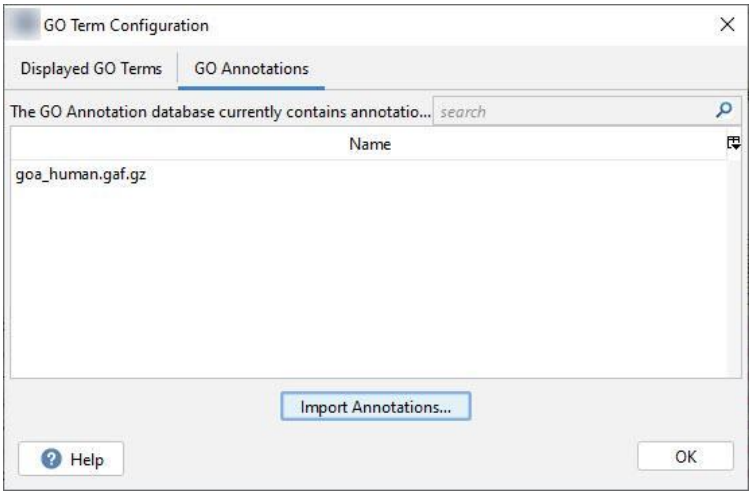


- Import annotations

This button opens a dialog from which the user can import GO databases in Scaffold DIA.

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

Figure 4-7: Add GO Annotations 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 DIA 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/.
- 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 DIA. 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 DIA 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 DIA .

Tool-bar

Figure 4-8: Scaffold DIA toolbar

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

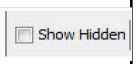
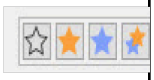
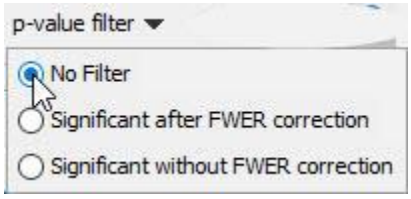
Icon	Function
	New—Starts a new experiment and opens a file browser to allow the selection of files to load in Scaffold DIA, see Loading data into Scaffold DIA.
	Open—Opens a saved Scaffold DIA experiment file, *.SDIA through a file browser.
	Save—Standard Windows behavior.
	Library Manager - Opens the Library Manager dialog.
	Print—Prints the current view.
	Print Preview—Previews current view with the option to print the document.
	Copy—For each view copied to the clipboard the first table appears 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 The Configure Sample Organization and Statistical Analysis Dialog.
	Excel—Exports the information that is contained in the current view to a tab-delimited text file that can be opened and viewed in Excel.

Icon	Function
	**Help**—Opens the Scaffold DIA Online Help.

Filtering control bar

The Scaffold DIA Filtering 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 4-8: Scaffold DIA Filtering control barl

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 Significance Level tab. If Family-wise error correction has not been applied, the "Significant after FWER correction option is grayed out.  **Note:** This filter appears in the filtering control bar only after a Quantitative test has been applied, see Quantitative Testing.

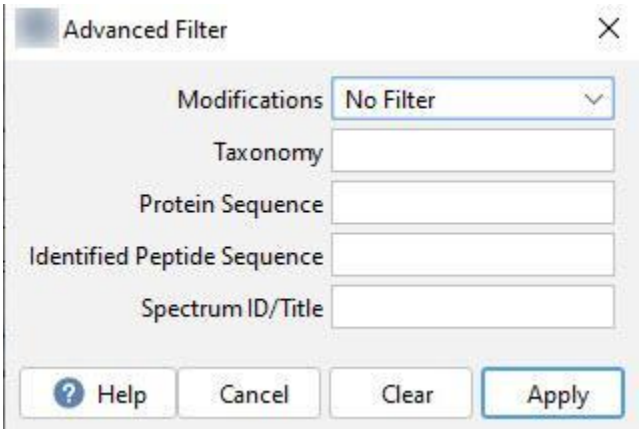
GO Term	**GO Filter**—Filters the protein list to show only those proteins matching the selected GO Term. **Note:** This filter appears in the filtering control bar only after GO Terms have been applied, see Apply GO terms. 
	**Advanced Filter** icon— Clicking this icon opens a dialog box where the advanced searches can be set up, see Advanced Filter.

The Filtering control bar includes the following functions:

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 DIA. 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 4-10: Advanced Filter dialog

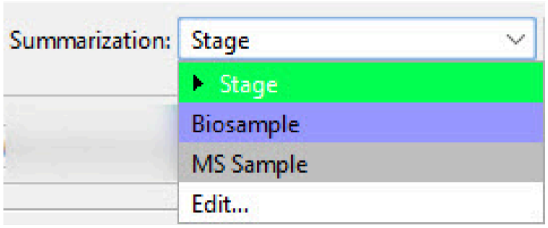


If more than one filter is selected at the same time, Scaffold DIA "AND"s the chosen filters. In other words, a peptide or protein must satisfy ALL selected filters in order to be displayed. The APPLY button starts the application of the filters and a Wait dialog box appears to monitor its progress.

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 Filtering Control bar. A red cross button also appears, clicking the button cancels the applied filter resetting the protein list to the original one.

Figure 4-11: Filtering Control Bar: Icon appearing when Advanced filter is applied



Star Filter

The Star Filter box contains four toggle buttons located next to the "Show Hidden" checkbox. Each button is characterized by one of the four possible star states the user can trigger for a specific analyte group or cluster, by clicking the icon shown under the "Star" column in the Samples table. This action is called "starring an analyte".

Figure 4-9: Filtering Control Bar: Star filter.

One or more of the stars may be selected for filtering. When a star in the filter bar is clicked, a red bar appears over that star and the program will filter out all rows of the Samples Table that are tagged with that particular star or combination of stars. To assign stars to proteins in the Samples View, see Tagging Proteins of Interest, the star function.

Each star button has two possible filtering states:

- **Unselected** - The star appears at the center of a squared box and analytes tagged with that type of star are included in the Samples table. When the button is clicked, a red diagonal bar appears across the box and the analytes that are tagged by this specific star disappear from view. The Cross icon becomes active, indicating that a filter has been applied.
- **Selected** -The star appears at the center of a squared box barred with a red diagonal and the analytes tagged with that type of star are filtered from the list. Clicking the button clears the red bar and returns the analytes to the list. If no other filters are in effect, the Cross icon is grayed out.

It is possible to select one or more star filter buttons at the same time. The analytes tagged with the selected stars will be hidden from the analyte list. Selecting the uncolored star leaves only starred analytes in the list. For more information, see Tagging proteins of interest, the star function.

Text Search Box

The **Text Search box** filters the list of analytes in the Samples Table according to what has been typed in the box. The filter searches for the typed characters in the Analyte Name and Accession Number columns in the Samples Table.

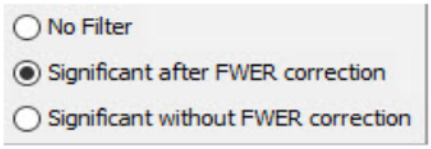
It is possible to allow the use of regular expressions in the search field by going to the menu option **Edit > Preferences** and selecting the appropriate option in the **Searching** tab.

P-Value Filter

The filter is active only when a statistical test has been applied. When applying a test, the user may select the p-value threshold below which a result is considered statistically significant, and may also apply a FWER correction to this significance level to adjust for multiple comparisons, see Multiple Test Correction.

. When a statistical test has been applied, a column containing the resulting p-values appears in the Samples Table, with p-values above the specified significance level colored green. P-values that meet the uncorrected significance level, but do not meet the adjusted significance level after applying a selected FWER correction are colored yellow. The p-value filter allows the user to filter out all analytes whose p-values do not meet significance criteria, either with or without correction.

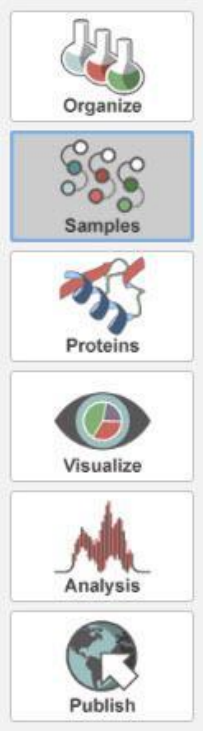
Figure 4-12: p-value Filter



Navigation pane

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

Figure 4-13: Scaffold DIA Navigation pane View selection



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

- The Organize View
- The Samples View
- The Proteins View
- The Visualize View
- The Analysis View
- The Publish View

FDR Dashboard

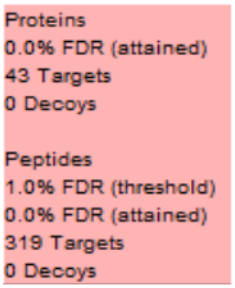
The **Dashboard** or **Information Box** is located under the navigation pane on the left lower corner of the Scaffold DIA main window. The box contains information about the peptide and protein FDR at the current FDR Thresholds together with the number of target and decoy spectra and proteins. It also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

The tooltip associated with the Dashboard also displays the pi0 value calculated by Percolator.

When a quantitative threshold is applied, the threshold is displayed in the Dashboard.

numbers of Identified Analytes, Consensus Features, and total Features at the current level of thresholding. The information reported in the Info Box when highlighted can be copied by simply using the standard CTRL C key strokes.

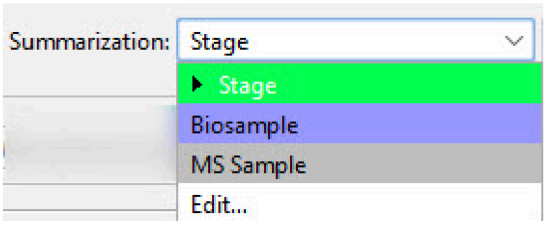
Figure 4-14: Dashboard



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 4-15: Scaffold DIA Summarization bar



The last item in the list is the command **Edit...**, which, when selected, opens the The Configure Sample Organization and Statistical Analysis Dialog through which it is possible to add Categories to the drop down list and to define a different hierarchical order.

Display pane

The information included in the different views appears in the Scaffold DIA 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 included in Scaffold DIA 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 4-16: Viewing information in a collapsed tool-tip

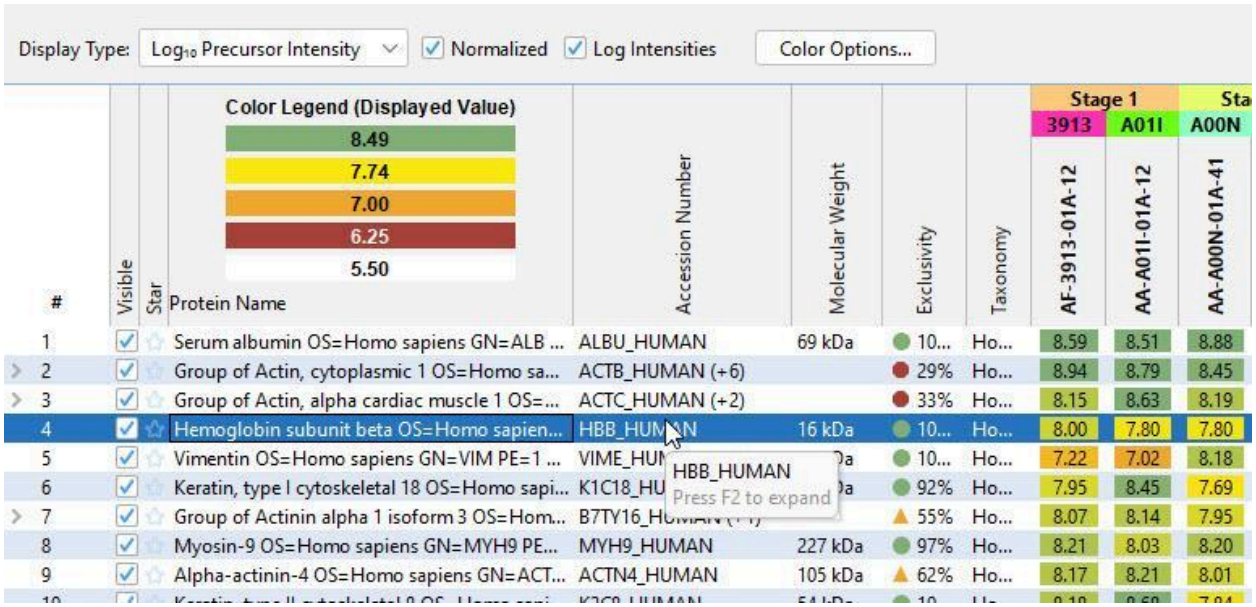
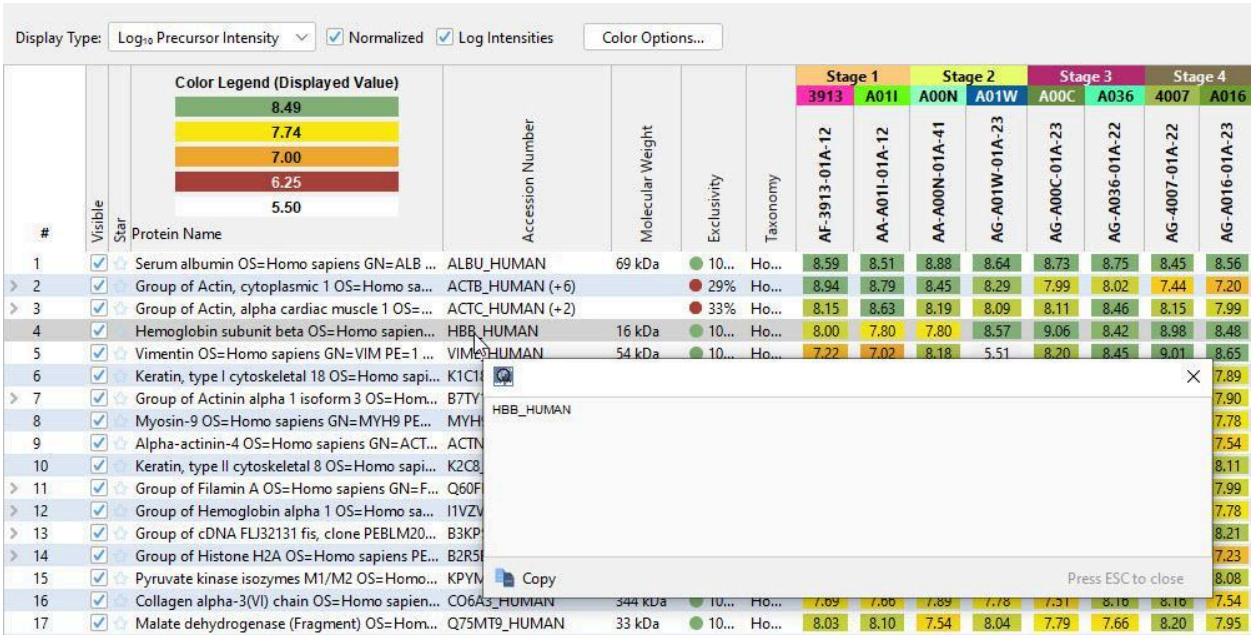


Figure 4-17: Viewing information in an expanded tool-tip



Resizing of columns and panes

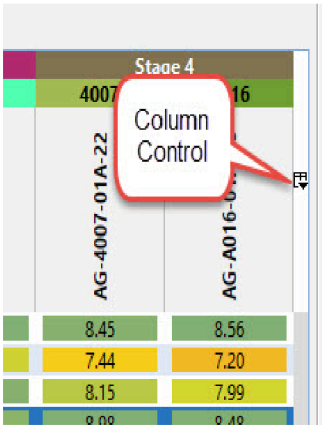
The user can resize columns and different panes in each of the views to better suit his/her working needs. For example, in the Samples Table, the user can change the width of a column by resting the mouse pointer on the right side of a column heading until the pointer changes to a double-headed arrow,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

Figure 4-18: Changing the width of a column in the Samples View

Table Column Control

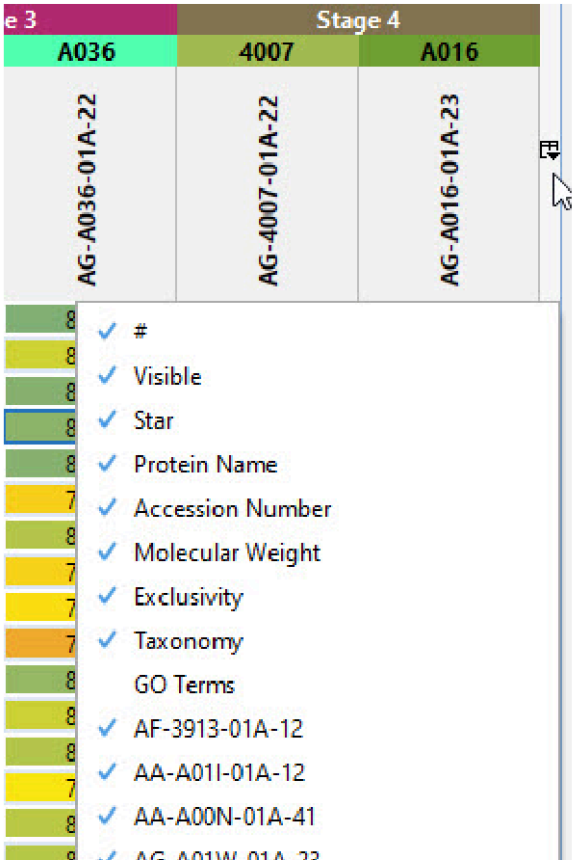
All tables throughout Scaffold DIA have a feature called Column Control. It is a vertical button placed on the right side of every table lined up with the column headers. When the user clicks the button or selects the option **View > Show/Hide Columns** from the main menu, a drop down list of all the columns opens. Each column is associated with a checkbox and at the bottom of the list there are three included group commands.

Figure 4-19: Scaffold DIA



Column Control button

Figure 4-20: Column Control pull down list



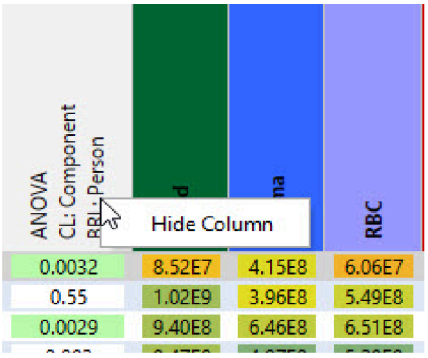
- Unchecking columns from the list will hide them from view in the Samples table.
- The Horizontal Scroll command if checked will add a scroll bar at the bottom of the Samples table.

- 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 make the column reappear use Table Column Control.

.

Figure 4-21: Hiding a column with Mouse Right-click



Moving columns

In all tables throughout Scaffold DIA, 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 is being moved and drags it to the new location. The new position will be retained when the user switches to another view and then returns.

Figure 4-22: Moving columns in tables

Molecular Weight	Exclusivity	Taxonomy	Stage 1 A01I AA-A01I-01A-12	Accession Number	Stage 1 3913 AF-3913-1A-12	Stage 2 A00N AA-A00N-01A-41	Stage 2 A01W AG-A01W-01A-23	Stage 3 A00C AG-A00C-01A-23	Stage 3 A036 AG-A036-01A-22
69 kDa	100%	Homo sapiens	8.51	ALBU_HUMAN	8.59	8.88	8.64	8.73	8.75
	29%	Homo sapiens	8.79	ACTB_HUMAN (+6)	8.94	8.45	8.29	7.99	8.02
	33%	Homo sapiens	8.63	ACTC_HUMAN (+2)	8.15	8.19	8.09	8.11	8.46
16 kDa	100%	Homo sapiens	7.80	HBB_HUMAN	8.00	7.80	8.57	9.06	8.42
54 kDa	100%	Homo sapiens	7.02	VIME_HUMAN	7.22	8.18	5.51	8.20	8.45
48 kDa	92%	Homo sapiens	8.45	K1C18_HUMAN	7.95	7.69	7.73	7.90	7.47
	55%	Homo sapiens	8.14	B7TY16_HUMAN (+1)	8.07	7.95	8.24	7.76	8.18
227 kDa	97%	Homo sapiens	8.03	MYH9_HUMAN	8.21	8.20	8.06	7.59	7.50
105 kDa	62%	Homo sapiens	8.21	ACTN4_HUMAN	8.17	8.01	8.28	7.81	7.65
54 kDa	100%	Homo sapiens	8.68	K2C8_HUMAN	8.18	7.84	8.38	7.92	7.04
	97%	Homo sapiens	7.98	Q60FE6_HUMAN (+3)	8.12	7.85	7.78	7.44	8.37
	100%	Homo sapiens	8.23	I1VZV6_HUMAN (+1)	8.47	8.62	8.57	8.31	8.04

Column sorting feature

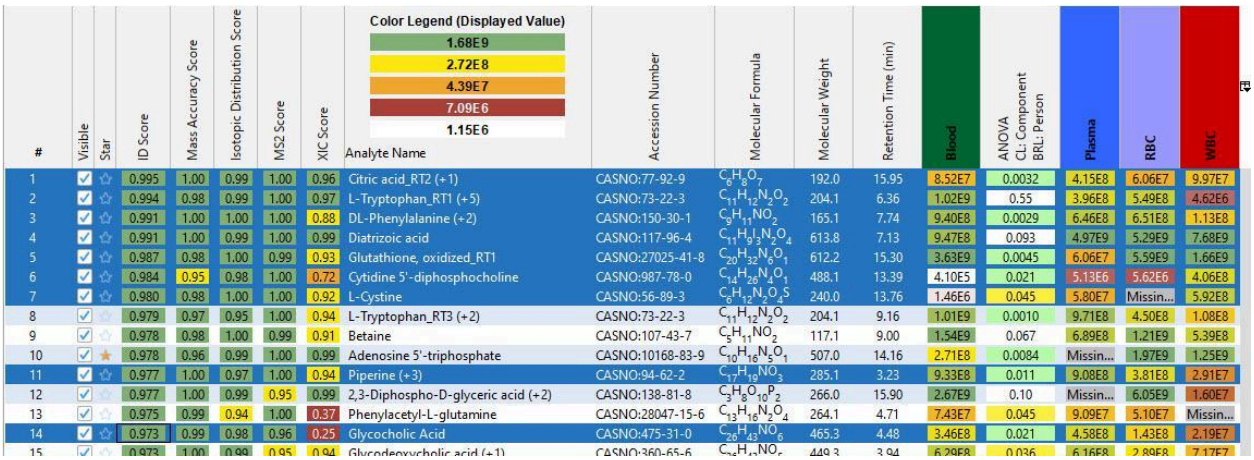
In all tables throughout Scaffold DIA, the user can use the tri-state column sorting feature and sort the display by clicking on any column header. For example, to sort the proteins based on their accession number, the user can click the Accession Number column header to initially select the column. 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 in 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 DIA 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, such as assigning a star to the selected group of proteins in the Samples table, for example.

Figure 4-23: Row multi-selection in the Samples Table



Color Legend (Displayed Value): 1.68E9, 2.72E8, 4.39E7, 7.09E6, 1.15E6

#	Visible	Star	ID Score	Mass Accuracy Score	Isotopic Distribution Score	MS2 Score	XIC Score	Analyte Name	Accession Number	Molecular Formula	Molecular Weight	Retention Time (min)	Blood	ANOVA CL: Component BRL: Person	Plasma	RBC	WBC
1	☑	☆	0.995	1.00	0.99	1.00	0.96	Citric acid_RT2 (+1)	CASNO:77-92-9	$C_6H_8O_7$	192.0	15.95	8.52E7	0.0032	4.15E8	6.06E7	9.97E7
2	☑	☆	0.994	0.98	0.99	1.00	0.97	L-Tryptophan_RT1 (+5)	CASNO:73-22-3	$C_{11}H_{12}N_2O_2$	204.1	6.36	1.02E9	0.55	3.96E8	5.49E8	4.62E6
3	☑	☆	0.991	1.00	1.00	1.00	0.88	DL-Phenylalanine (+2)	CASNO:150-30-1	$C_9H_{11}NO_2$	165.1	7.74	9.40E8	0.0029	6.46E8	6.51E8	1.13E8
4	☑	☆	0.991	1.00	0.99	1.00	0.99	Diatrizoic acid	CASNO:117-96-4	$C_{11}H_9I_3N_2O_4$	613.8	7.13	9.47E8	0.093	4.97E9	5.29E9	7.68E9
5	☑	☆	0.987	0.98	1.00	0.99	0.93	Glutathione, oxidized_RT1	CASNO:27025-41-8	$C_{20}H_{32}N_6O_1$	612.2	15.30	3.63E9	0.0045	6.06E7	5.59E9	1.66E9
6	☑	☆	0.984	0.95	0.98	1.00	0.72	Cytidine 5'-diphosphocholine	CASNO:987-78-0	$C_{14}H_{26}N_4O_1$	488.1	13.39	4.10E5	0.021	5.13E6	5.62E6	4.06E8
7	☑	☆	0.980	0.98	1.00	1.00	0.92	L-Cystine	CASNO:56-89-3	$C_6H_{12}N_2O_4S$	240.0	13.76	1.46E6	0.045	5.80E7	Missin...	5.92E8
8	☑	☆	0.979	0.97	0.95	1.00	0.94	L-Tryptophan_RT3 (+2)	CASNO:73-22-3	$C_{11}H_{12}N_2O_2$	204.1	9.16	1.01E9	0.0010	9.71E8	4.50E8	1.08E8
9	☑	☆	0.978	0.98	1.00	0.99	0.91	Betaine	CASNO:107-43-7	$C_5H_{11}NO_2$	117.1	9.00	1.54E9	0.067	6.89E8	1.21E9	5.39E8
10	☑	★	0.978	0.96	0.99	1.00	0.99	Adenosine 5'-triphosphate	CASNO:10168-83-9	$C_{10}H_{16}N_5O_1$	507.0	14.16	2.71E8	0.0084	Missin...	1.97E9	1.25E9
11	☑	☆	0.977	1.00	0.97	1.00	0.94	Piperine (+3)	CASNO:94-62-2	$C_{17}H_{19}NO_3$	285.1	3.23	9.33E8	0.011	9.08E8	3.81E8	2.91E7
12	☑	☆	0.977	1.00	0.99	0.95	0.99	2,3-Diphospho-D-glyceric acid (+2)	CASNO:138-81-8	$C_3H_8O_{10}P_2$	266.0	15.90	2.67E9	0.10	Missin...	6.05E9	1.60E7
13	☑	☆	0.975	0.99	0.94	1.00	0.87	Phenylacetyl-L-glutamine	CASNO:28047-15-6	$C_{13}H_{16}N_2O_4$	264.1	4.71	7.43E7	0.045	9.09E7	5.10E7	Missin...
14	☑	☆	0.973	0.99	0.98	0.96	0.25	Glycocholic Acid	CASNO:475-31-0	$C_{26}H_{43}NO_6$	465.3	4.48	3.46E8	0.021	4.58E8	1.43E8	2.19E7
15	☑	☆	0.973	1.00	0.99	0.95	0.94	Glycodeoxycholic acid (+1)	CASNO:360-65-6	$C_{26}H_{43}NO_5$	449.3	3.94	6.29E8	0.036	6.16E8	2.89E8	7.17E7

Mouse Right-click Context Menus

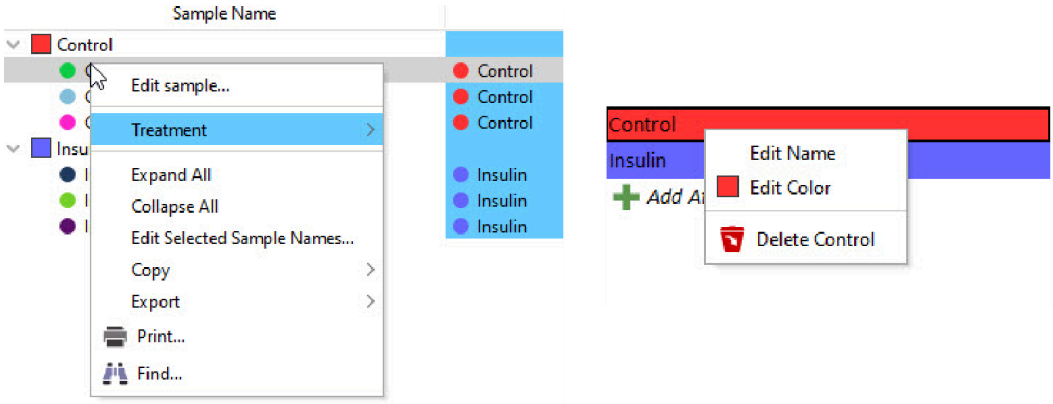
When the user hits the right-click button of the mouse while hovering over the Display Pane of a View, a menu with various options appears close to the working arrow. Depending on the selected view the list of options available in the menu varies. A description of the mouse right-click command is provided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

Organize View

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

- When a row in the Organize View Table is selected, a context menu appears. The menu contains a 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.
- When a row in the Categories tree table is selected, Context menu B appears. The Edit Name command opens up the Bulk Edit Sample Names and the Delete option deletes a selected attribute.

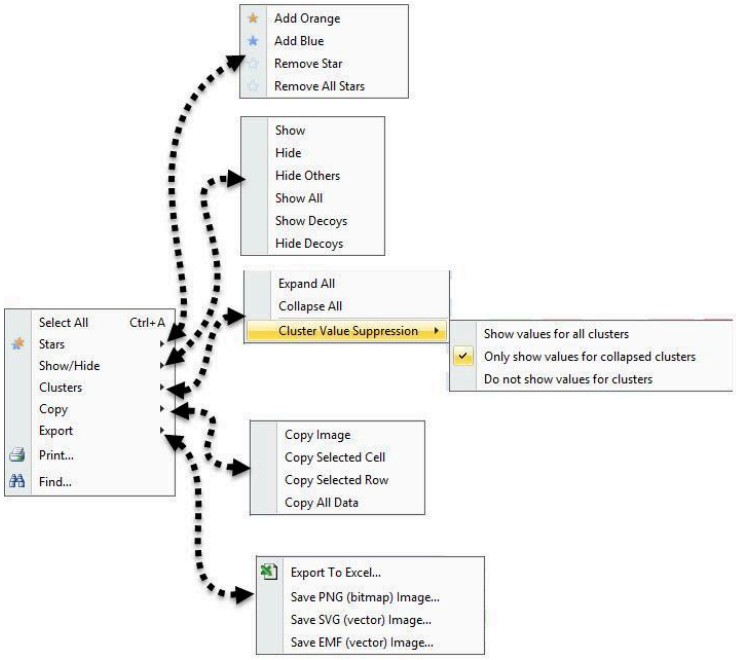
Figure 4-24: Context Menu B and Context Menu A



Samples View

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

Figure 4-25: Context Menu C



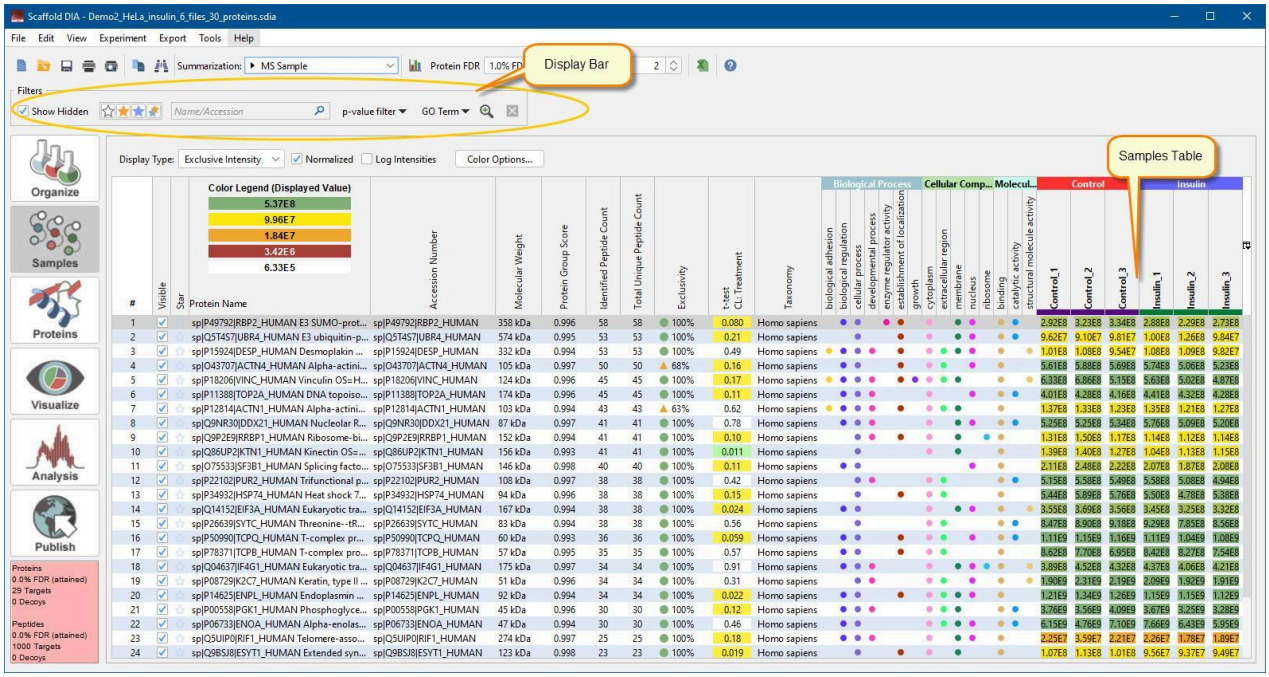
Chapter 5: The Samples View

The Samples View provides a series of tools to help the user summarize and interpret experimental results at the protein level. The Samples View consists of two components: the Samples Table. and the Display Bar.

In addition, the Summarization Bar, found in the main Scaffold DIA window, allows the user to change summarization and display levels to provide various views of the results. For more information, see Experimental Design.

The Samples View is the first view appearing in the Scaffold DIA window when the application finishes loading data or when the user opens a Scaffold DIA *.SDIA file. 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. Several Display Type options are available, including Log_{10} Intensity, Log_2 Fold Change and Quantified Peptide Count.

Figure 5-1: The Scaffold DIA Samples View



Samples Table

Each row of the Samples Table displays quantitative information for a protein group or a protein cluster in every MS sample or group of samples at a selected level of summarization, indicated as a column.

The level of data summarization is chosen from the Summarization Bar available in the Scaffold DIA main window. The Display Type pull down menu determines the type of quantification reported in the table.

The protein list can be filtered using the tools in the Filtering control bar in the Scaffold DIA Main Window.

Features of the Samples Table

Like any table in Scaffold DIA, the Samples Table includes the features and tools described in the Display pane section.

Initial thresholding when loading data

Every time data is loaded into Scaffold DIA, protein clustering and thresholds are applied. Initial grouping and thresholding options may be selected through the Workflow Dialog. After loading, the user may modify the clustering settings through the Experiment>Protein Clustering menu option and may change the thresholds through the Thresholding Controls in the main window.

Initial Sorting of Columns

When the Samples View first opens, all of the protein groups in the protein list are sorted by the Protein Group Score. This score is the maximum of the peptide probabilities of the peptides which are exclusive to the protein group. The protein groups with the most evidence appear 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 table to the original order.

Applying Quant Thresholds

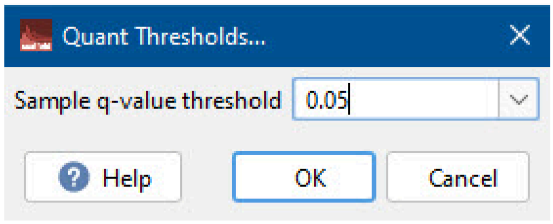
In Scaffold DIA, peptide identifications are made on an experiment-wide basis, and if the peptide is not detected in some samples, an attempt is made to extract any signal that corresponds to fragment ions of the peptide in the corresponding position in those samples. If the peptide is missing in that sample, it is expected that this will result in the integration of background noise, which will give a very low value, appropriate for calculation of ratios.

While this is helpful for quantifying proteins in samples in which the proteins are present in low levels, occasionally this results in quantitative values that are based on interference and the Samples View can display significant quantities of a protein in samples that do not contain the protein. This can be

misleading when an experiment attempts to assess presence or absence of proteins under different conditions. To avoid this situation, a sample level quantitative threshold may be applied.

Sample q-value thresholds are based on the q-values calculated by Percolator in the individual samples. These sample-specific q-values are roughly equivalent to FDR values for identification of the peptide in a specific sample, and are displayed in The Peptide Match Pane in the Proteins View. To apply a sample q-value threshold, Select Apply Thresholding in the Experiment menu. This will offer the option to apply Quant Thresholds. When this option is selected, a Quant Thresholds dialog appears. Select a value from the dropdown or type in a specific q-value threshold and click **OK**. Any peptide matches with sample-specific q-values above that threshold will be treated as if they had no quantitative intensity values. By default, the threshold is set to No bound, which does not filter out any peptide quantitative values.

Figure 5-2: The Quant Thresholds Dialog



When a q-value threshold is applied, the setting is displayed in the FDR Dashboard.

Exclusive Peptide Quantification Option

All of the quantitative values shown in the Samples Table are affected by the "Quantify on Exclusive Peptides" setting in the Experiment menu. By default, this option is checked, which means that only those peptides which are associated with a single protein group in the experiment are used in computing the intensity values and ratios and the quantified peptide counts. Peptides which are shared by two or more protein groups are ignored, although they still appear in the Proteins View.

Exclusivity is calculated independently for protein clusters, so a peptide may be exclusive to a cluster but not to any specific protein group within the cluster.

The user may change this setting through the checkbox in the Experiment menu.

Peptide Count Columns

Identified Peptide Count

This column provides the total number of peptides identified in the experiment that are associated with each protein. Peptides with the same sequence but different modifications are counted separately.

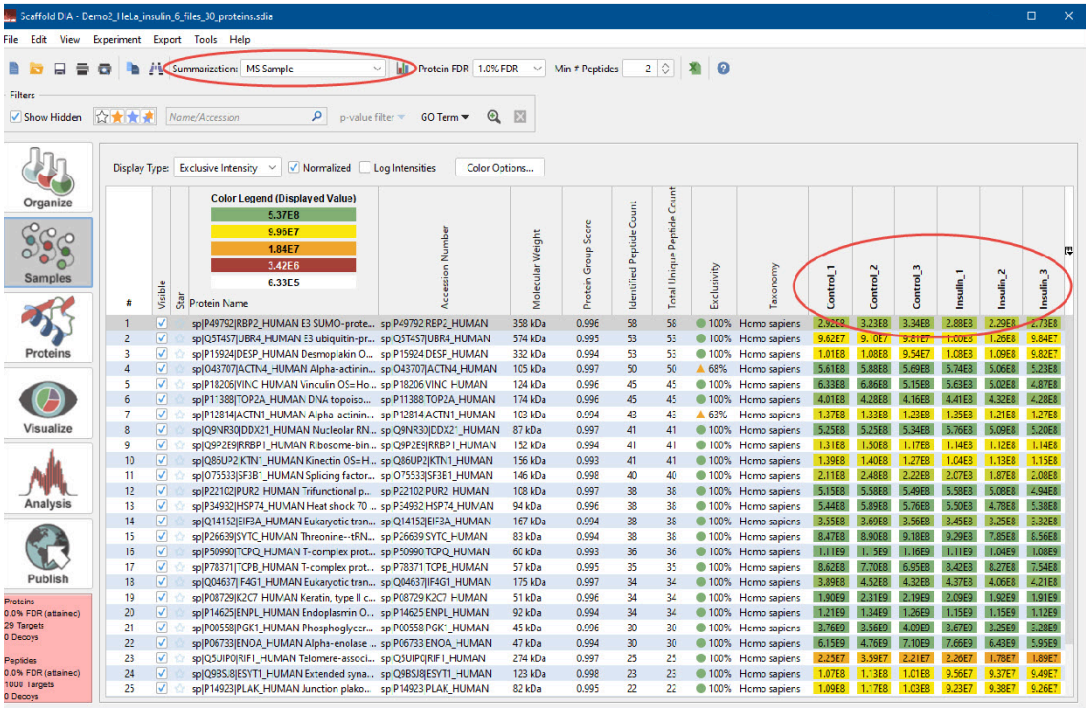
Total Unique Peptide Count

This column provides the total number of peptides, as defined by their amino acid sequences, identified in the experiment that are associated with each protein. Peptides with the same sequence but different modifications are counted as a single peptide.

Summarization Level in the Samples View

After a new experiment is completely loaded, the Scaffold DIA Samples View appears with the Samples Table initially summarized at the lowest level of summarization which is the MS sample level.

Figure 5-3: Scaffold DIA Samples View - The Samples Table initial summarization level

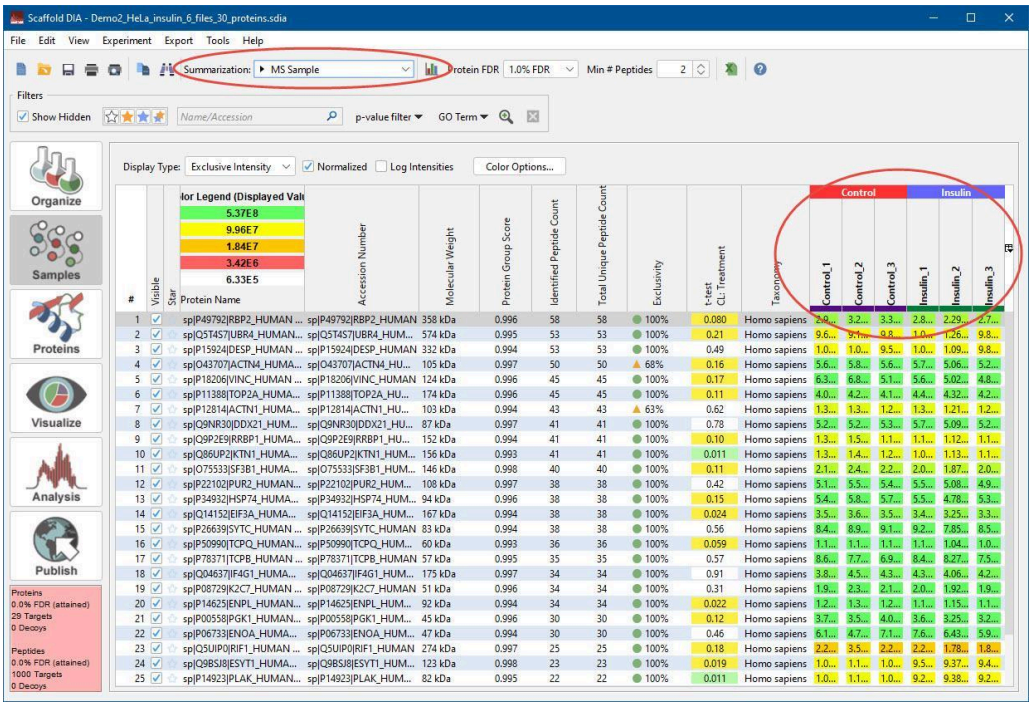


The user can add Attribute Groups and assign them to the various MS samples in the Organize View. It is then possible to set up the desired Summarization Hierarchy using any of the defined Attribute Groups by choosing **Edit...** in the Summarization Bar.

The selected Summarization Hierarchy and the choice of the summarization level will be reflected in the Samples Table column headers. In Figure 5-4 the selected level is still MS Sample but a more complex hierarchy has been set up in the Summarization Pane.

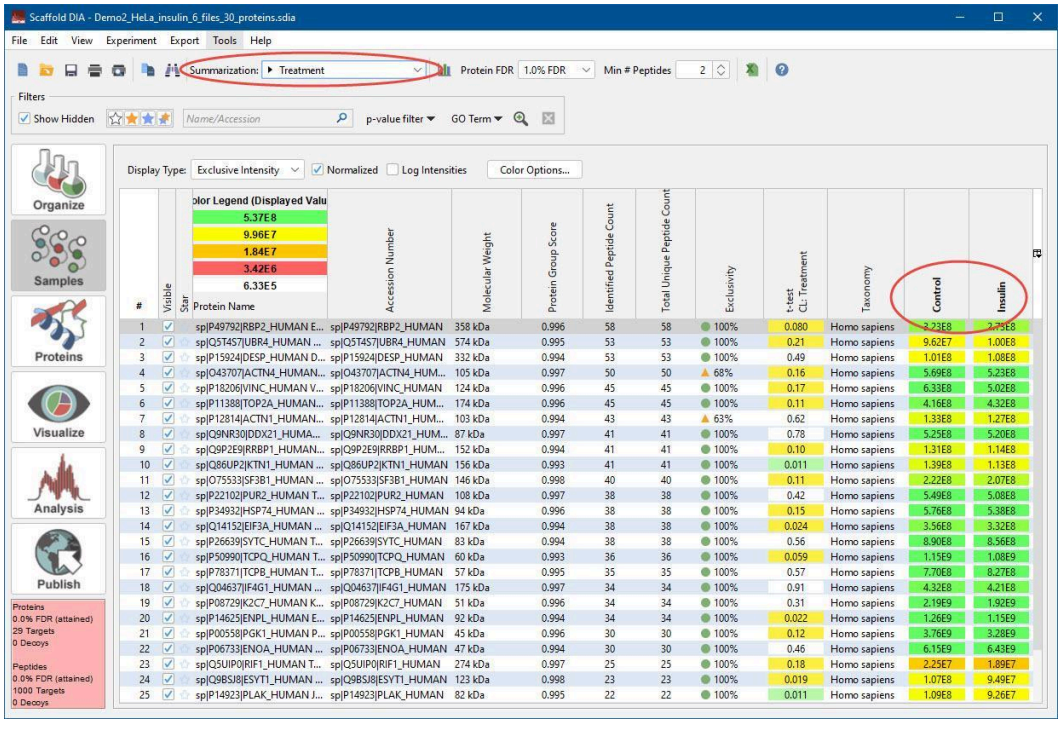
The headers of the sample columns show the Attributes for all Attribute Groups in the Summarization Hierarchy. They are colored according to the values assigned to the Attributes through the Organization View.

Figure 5-4: Scaffold DIA Samples View - The Samples Table levels of summarization added



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

Figure 5-5: Scaffold DIA Samples View - The Samples Table highest level of summarization



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, 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 the Color Options Button located in the Display Bar.

Tools for limiting the Protein List

Protein lists can be very long. Scaffold DIA includes a series of tools that allow the user to simplify the list and filter it to focus on the proteins of interest. Grouping and clustering proteins that share similarities reduces the number of independent rows in the list. Thresholding the protein 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:

- Protein Grouping and Clustering
- Tagging Proteins of Interest, the star function
- Hidden Proteins
- Applying Confidence Thresholds to the Protein list
- Applying filters to the Protein List

Protein clusters appearing in the initial Protein List

For a protein or protein group to be included in the Protein list it must have a valid exclusive peptide in at least one of the MS Samples. Additional levels of protein clustering, based on the degree to which the proteins share peptide evidence, are also available in Scaffold DIA. The level of clustering in the initial display is selected by a parameter in the Analysis tab of the Workflow dialog during loading.

The initial thresholds are also specified in the Workflow dialog. Thresholds may be changed using the controls in the main Scaffold DIA window.

Cluster options may be changed through the Experiment menu or through the right-click context menu in the Samples View, see Protein Grouping and Clustering.

Tagging proteins of interest, the star function

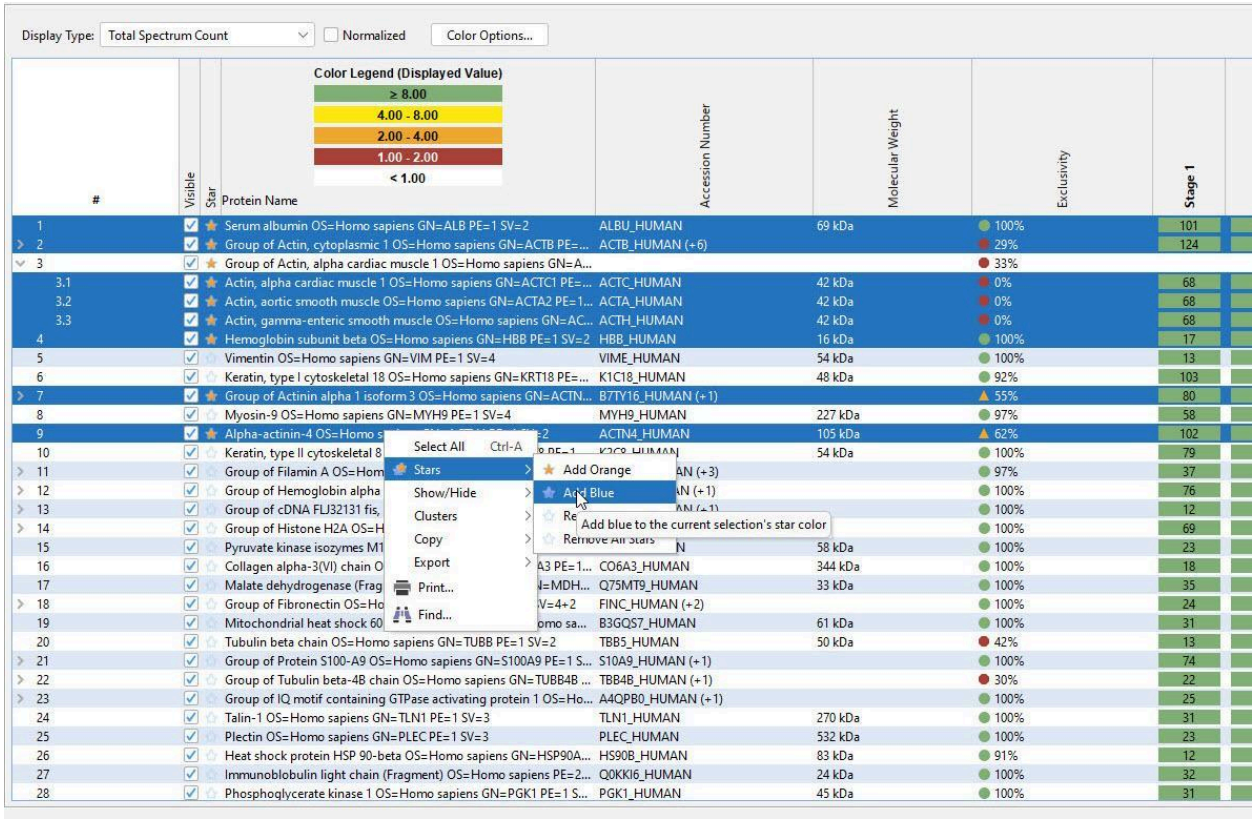
The user can mark proteins in an experiment that are of special interest by clicking the Star icon  in the **Starred?** column for the protein. Three different colored stars, blue, orange and a combination of the two colors may be applied by clicking multiple times on the same star or by selecting the 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 proteins to the top of the display by clicking the **Starred?** column header. To return to the default display order, click the column header twice more.

Groups 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 DIA Main Window. Clicking a specific colored star in the filters bar removes all proteins tagged with that star color from the table.

Figure 5-6: Samples Table- Starring proteins



Hidden Proteins

The user can easily remove proteins displayed in the Samples Table that are of no interest and/or are contaminants. An entire protein entry in the Samples Table can be eliminated by simply clearing the Visible option for the protein. This can be done for a single protein by clicking the Visible checkbox, or for a group of proteins 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 checkbox Show Hidden in the Filtering pane in the Scaffold DIA Main window toggles the display of hidden proteins.

Applying Confidence Thresholds to the Protein list

When data is initially loaded into Scaffold DIA, the list of proteins in the Samples Table reflects the thresholds set through the Workflow Dialog during loading.

Within Scaffold DIA, it is possible to set the Protein False Discovery Rate (FDR) threshold and the minimum number of peptides in a protein through the Threshold Controls in the main window.

Only the Protein FDR threshold may be adjusted because the Peptide FDR threshold set in the Workflow Dialog is used by Percolator as part of the library search process. To change this value, it is necessary to perform a new search.

The thresholds specified represent upper bound values for the FDR Scaffold DIA computation. The thresholds are applied as described in Appendix B, Computation of protein and peptide FDR in Scaffold DIA. The actual computed values after thresholding are shown in the FDR Dashboard located in the lower left corner of the Scaffold DIA main window. The number of target and decoy proteins and the number of target and decoy peptides are also reported in the box, along with the total number of proteins that pass thresholds.

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

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.

Applying filters to the Protein List

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

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

Display Bar

Through the Display Bar, the user can specify the type of values 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.

Figure 5-7: Scaffold DIA Display bar

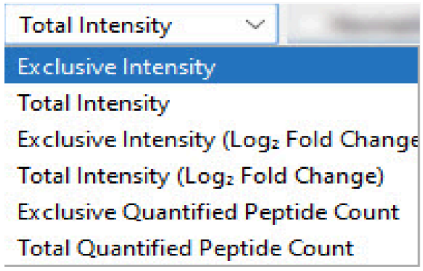
The Display bar contains the following features:

- Display Type
- Normalized checkbox or Ref: pull down list
- Color Options button
- Search Box

Display Type

The Display Type drop down list offers a range of sample statistics. When a particular quantitative type is selected from the list, the values in the cells of the table are updated to reflect the appropriate values.

Figure 5-8: List of Display Type available options



- **Total Intensity/Log$_{10}$ Total Intensity** -- The summarized intensity value or the base 10 log of the summarized intensity value if the Normalized box is checked.
- **Exclusive Intensity/Log10 Exclusive Intensity --** The summarized intensity value of only the peptides that are associated with this protein group and no other or the base 10 log of this value if the Normalized box is checked.
- **Total Intensity (Log$_2$ Fold Change)** -- When this display option is selected, an additional control appears to allow selection of the reference value (See Ref: pull down list). The ratio of each value to the corresponding value in the reference value is calculated, and the base 2 log of the ratio is computed and displayed.

- **Exclusive Quantified Peptide Count** -- Peptide Counts represent the number of peptides in a specific protein group that had measurable fragment intensity in a specific sample or sample group. This should not be viewed as a quantitative value, but rather as a measure of the reliability of the intensity values shown by the other display types. It is important to recognize that Scaffold DIA will extract any signal that is found in the expected RT window at the expected m/z for a peptide, even if it may only represent noise, so quantifying a certain peptide in a certain sample may not indicate that the peptide was actually present in that sample. For this reason, statistical tests are not available for this display type. The Exclusive Quantified Peptide Count only includes the peptides that are associated with a single protein or protein group.
- **Total Quantified Peptide Count** -- The values displayed represent the number of peptides in a specific protein group that had measurable fragment intensity in a specific sample or sample group. This should not be viewed as a quantitative value, but rather as a measure of the reliability of the intensity values shown by the other display types. It is important to recognize that Scaffold DIA will extract any signal that is found in the expected RT window at the expected m/z for a peptide, even if it may only represent noise, so quantifying a certain peptide in a certain sample may not indicate that the peptide was actually present in that sample. For this reason, statistical tests are not available for this display type.

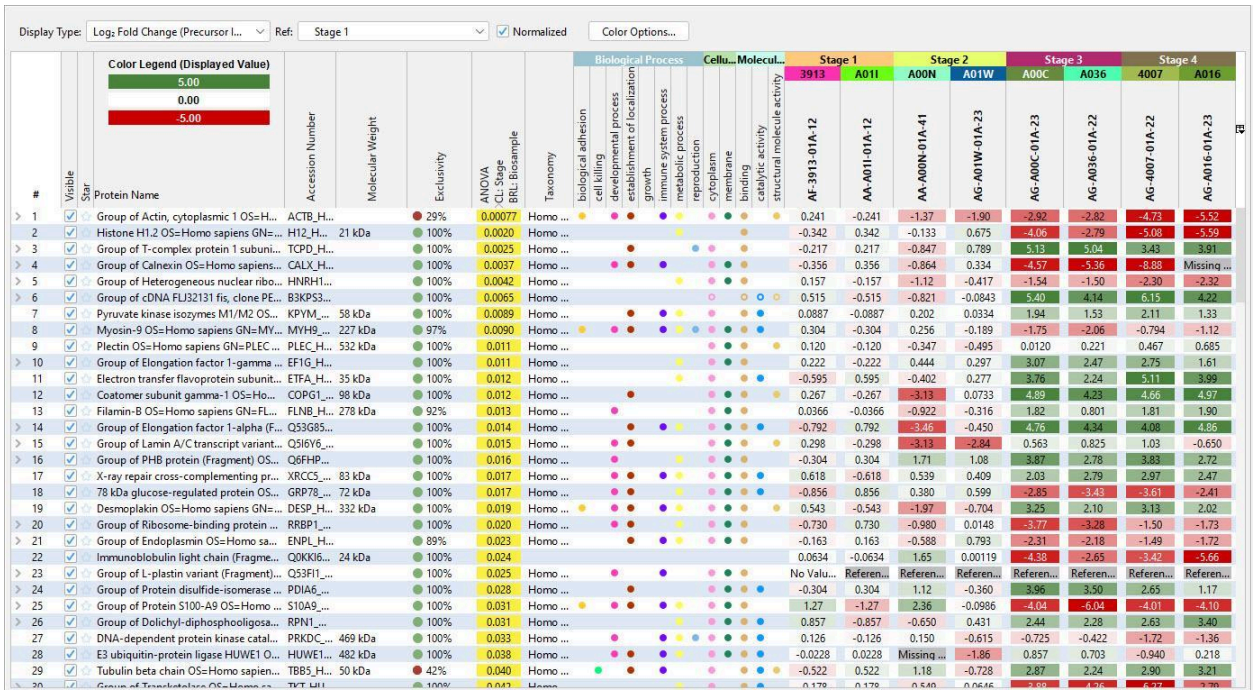
The coloring reflects the selected Display Type. For each type the coloring can be customized by selecting Color Options from the Table Tab Display pane.

Log_2 Fold Change and (Log_{10}) Intensity missing value tags

When selecting a Display Type that represents the log_2 fold change, 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 5-9: 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 DIA 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 Organize View and hierarchically ordered using the Sample Hierarchy Tab. 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.

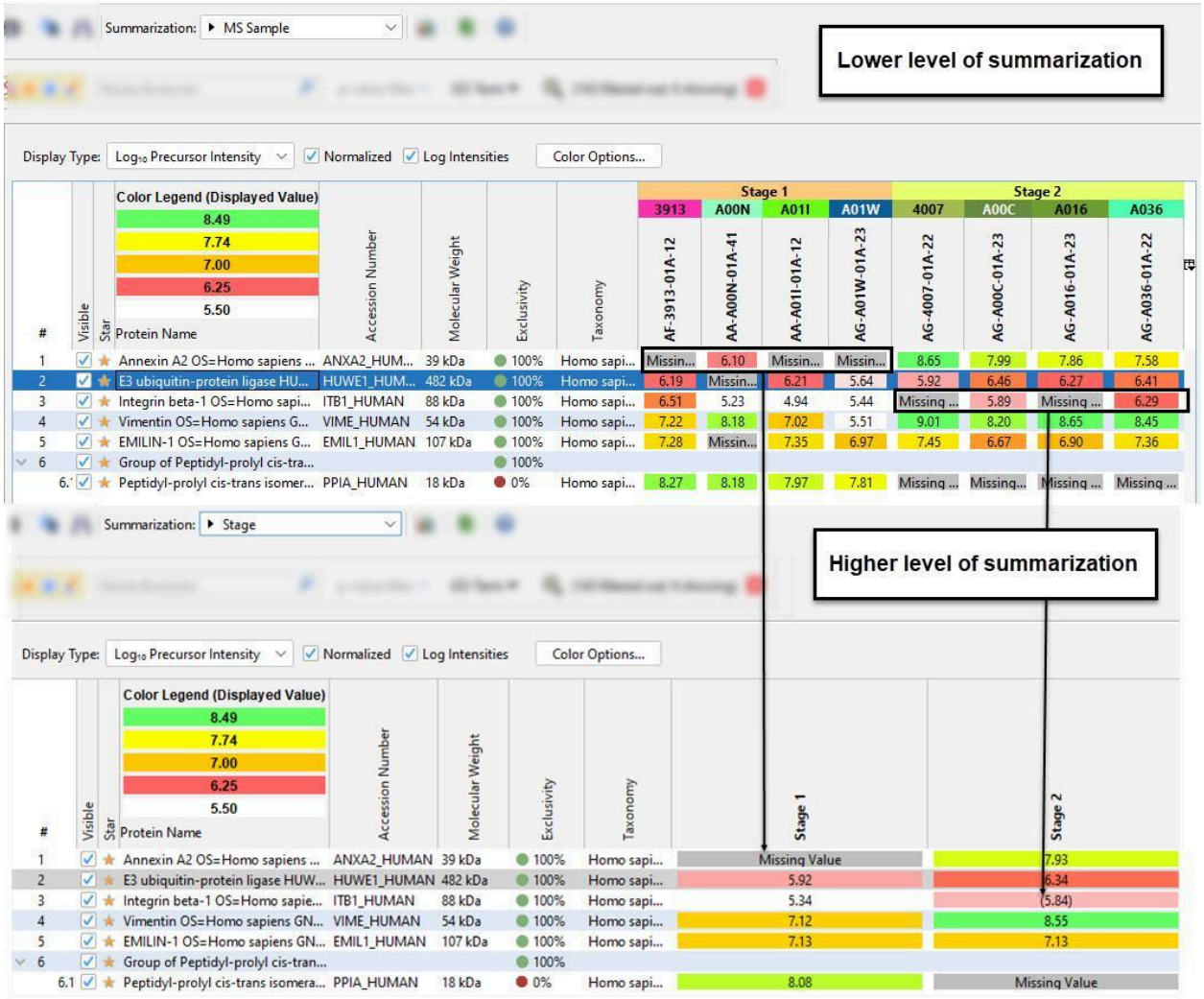
Rolling up of (Log_{10}) Intensity

Intensity values appearing in the Samples table, such as **Log_{10} Intensity**, 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, a missing value imputation method is applied. For a description of these functions see Summarization: Rolling up Values.

At the lowest level of summarization, missing values are highlighted using the Missing Value tag. At a higher level of summarization, when a value results from application of a missing value imputation method (See Appendix D., “Missing Values") it is reported in parentheses. Values 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 below.

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



*The **Log_2 Fold change (Intensity)** is calculated using the Log_{10} Intensity values with the Log_{10} value of the selected reference subtracted.*

Normalized checkbox

When the **Normalized** checkbox is checked, the values shown in the Samples Table for the selected Display Type will appear normalized.

The normalization algorithms applied are described in section Quantitative Methods.

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** checkbox. 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.

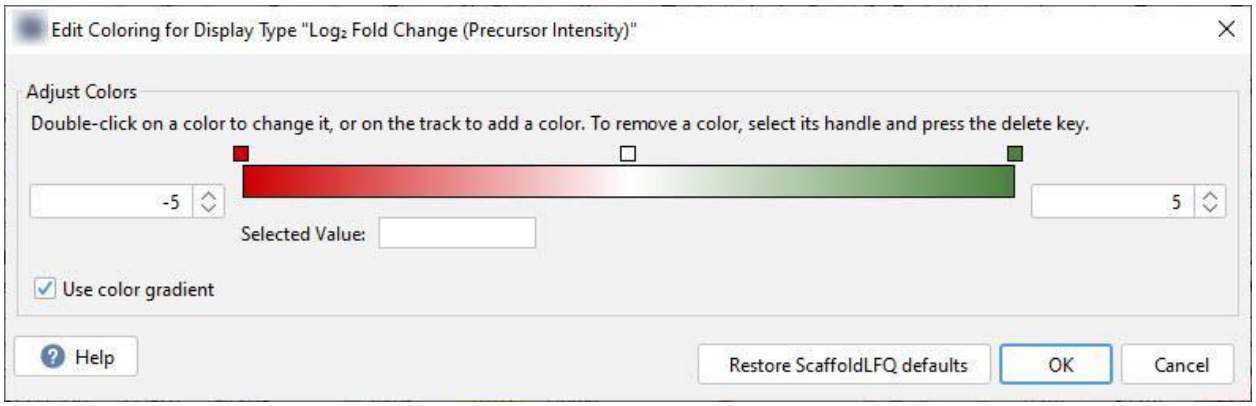
Log Intensities checkbox

When a Display Type representing an intensity value (e.g. Total Intensity, Exclusive Intensity, Total Log_{10} Intensity or Exclusive Log_{10} Intensity) is selected, an additional checkbox appears to allow the user to choose whether or not the intensities should be displayed as log10 values. This control affects display of intensity values throughout Scaffold DIA, including charts in the Proteins View and the Visualize View. Display format of intensity values may also be adjusted through the Log Intensities option in the View menu.

Color Options button

Selecting the Color Options button opens the Edit Coloring for Display Type dialog. This dialog offers coloring adjustments tools for each Display Type.

Figure 5-11: Edit Coloring for Display Type dialog



The Adjust Colors pane 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.

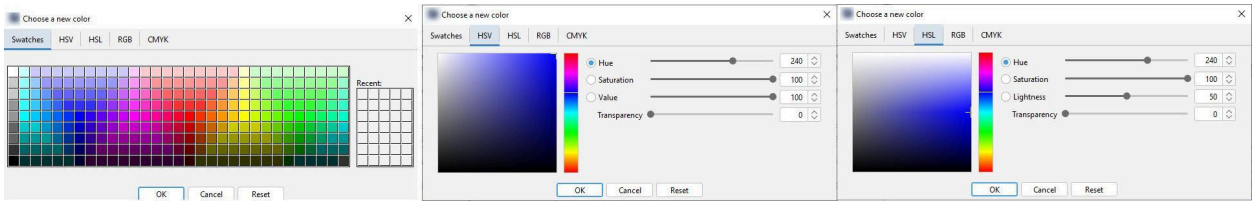
Sliding one of the colored squares located above the legend changes the range of the selected color. The color ranges can also be set by typing a value in the selected value box.

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

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



At the bottom of the Adjust Color pane, the **Restore Scaffold DIA Defaults** button resets the legend to the default Display Type colors.

Search Box

This box allows the user to search the protein list using the protein description or the protein accession number. To allow the use of regular expressions, the user must make the option available by selecting the box **Use regular expressions in search fields** in the Preferences General tab.

Chapter 6: The Organize View

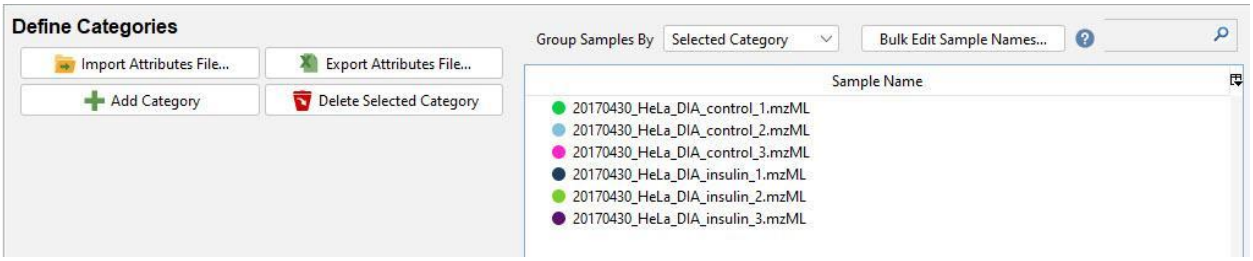
The Organize view displays the loaded samples in a tree structure that can be customized to reproduce the organizational structure of the experiment to be analyzed in Scaffold DIA. It works in conjunction with the “Configure Sample Organization and Statistical Analysis” dialog to support data analysis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

Organizing data in Scaffold DIA

The Organize View in Scaffold DIA provides an easy method to organize even complex experiments involving multiple variables. A variable or factor in Scaffold DIA 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 DIA, and the researcher may experiment with analyzing the data on the basis of any one or a combination of these categories.

Figure 6-1: 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.

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.

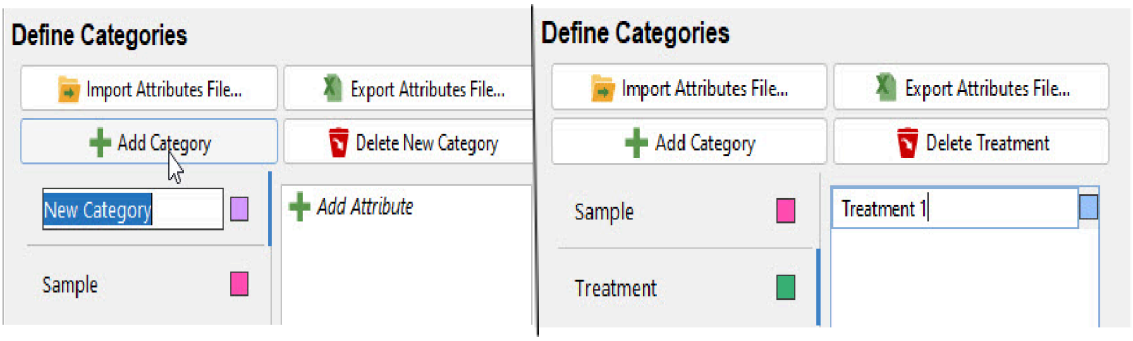
Tools in the Organize View

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"

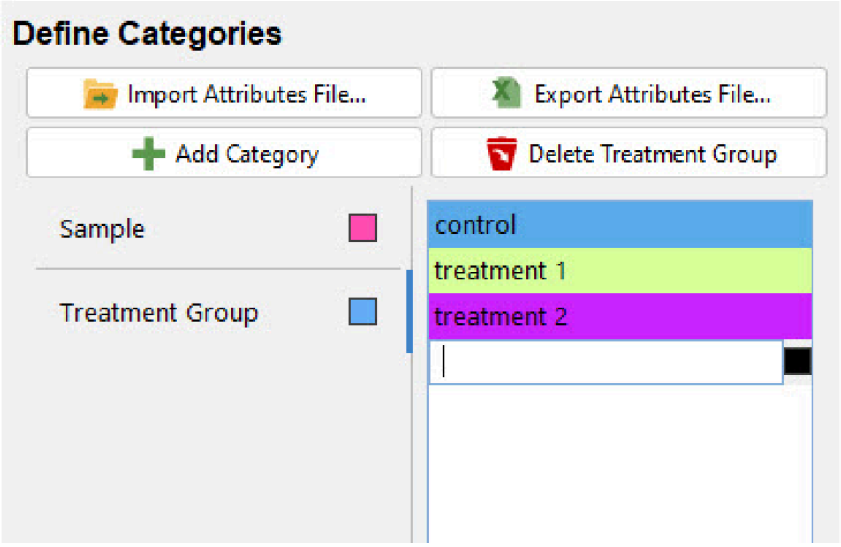
Figure 6-3: Creating a new category



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.

Add Attribute -- clicking on this control and typing a name creates an Attribute, or value, for the Category. When the Attribute name has been typed, pressing Enter creates the Attribute and brings up a field for entry of the next Attribute. Attributes may be ordered in the list using the arrow buttons or sort button (labeled A-Z) at the bottom of the pane, and the order shown in the Attribute list will be respected throughout the program.

Figure 6-4: 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).

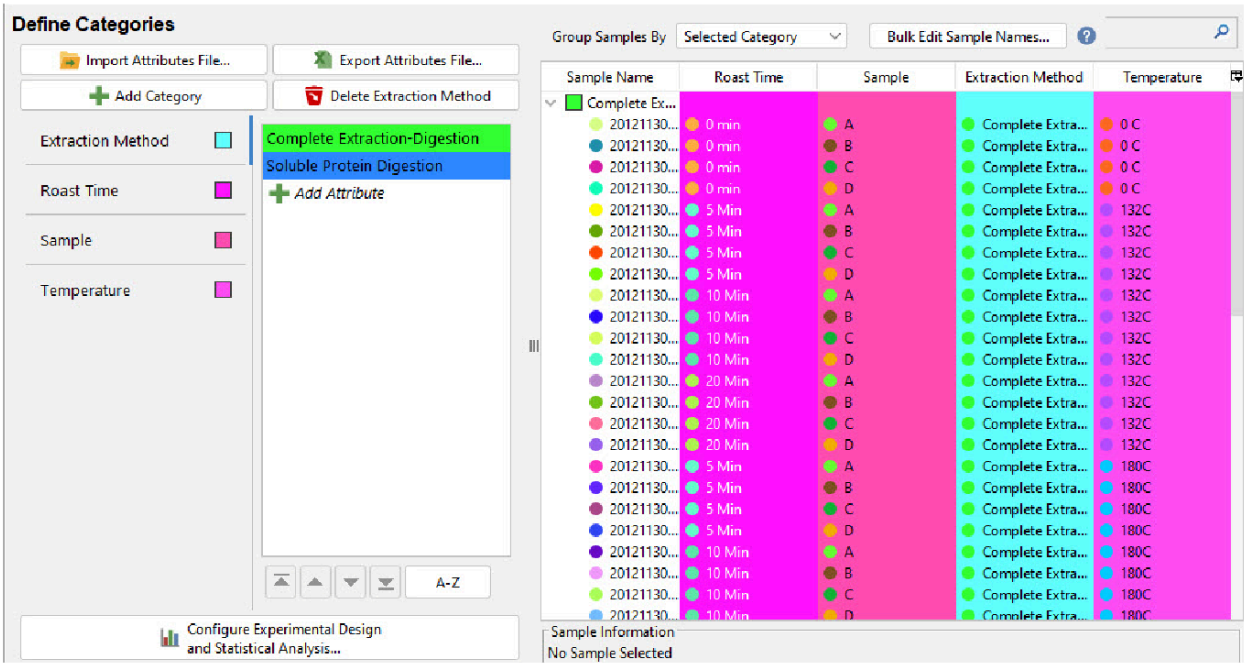
T**he Sample Organization Pane -** allows the user to assign attributes to the samples. It consists of:

- **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 Specifying the Design of an Experiment). 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 Configure Sample Organization and Statistical Analysis Dialog button which allows creation and editing of the Summarization Hierarchy.

Sample Organization Table

Scaffold DIA 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 Hierarchy tab, described in Specifying the Design of an Experiment.

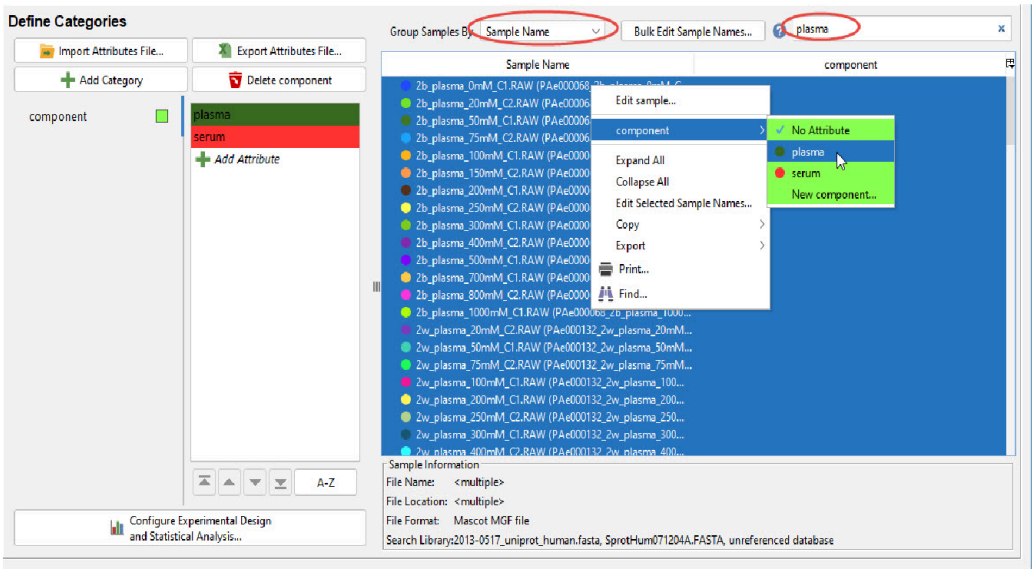
Figure 6-5: Organize View with added attributes



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.
- Filter Box - A helpful method when organizing large experiments is to use the Filter 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 Filter Box approach allows the user to leverage this information to quickly organize the samples.

Figure 6-6: 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.

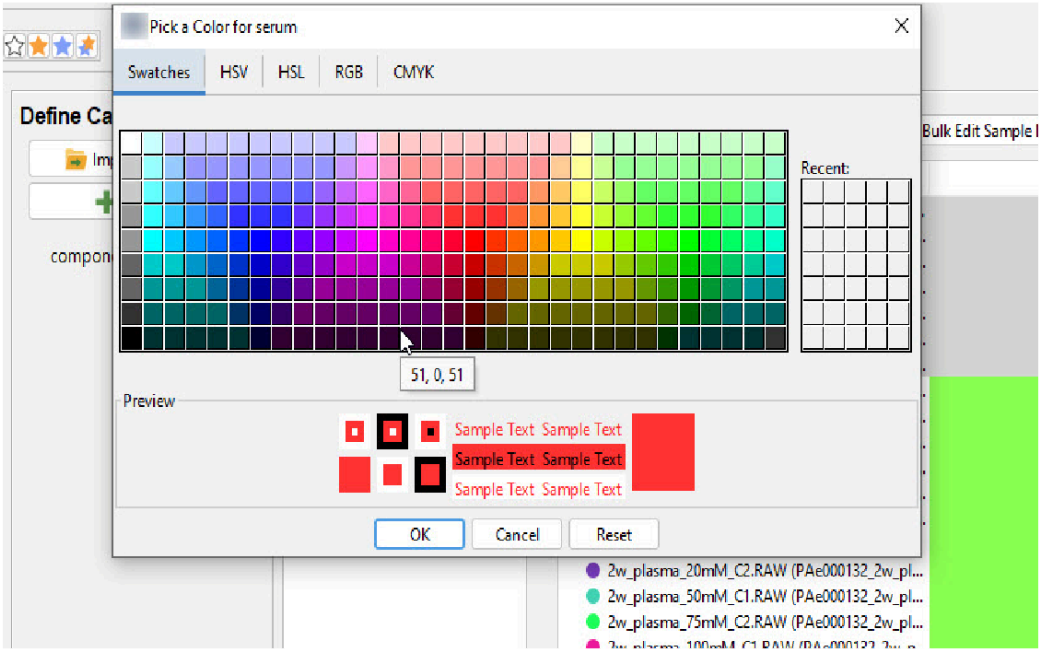
When the Group Samples By option selected is Selected Category or Experimental Design the table shows the list of Categories as collapsible folders:

- 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 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 option opens a dialog that allows the user to select a new color to be assigned to the Attribute.

Editing Category Colors

Figure 6-7: 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).

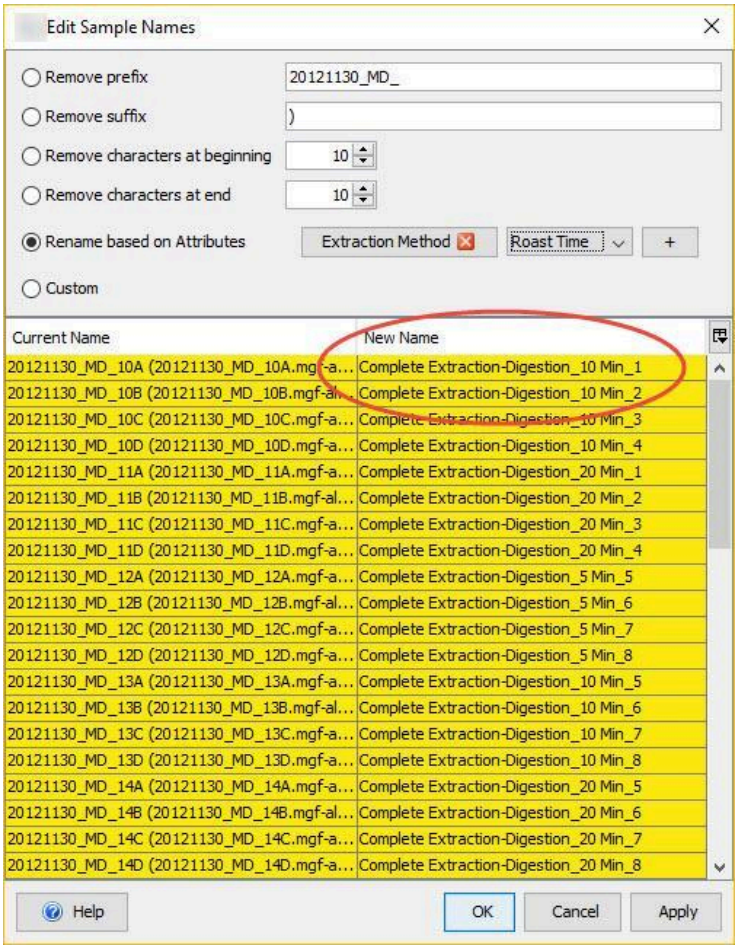
Bulk Edit Sample Names

Often sample names are quite long and difficult to distinguish. This can cause problems when viewing the data in the other Views of Scaffold DIA. 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 6-8).

Figure 6-8: 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 DIA 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.

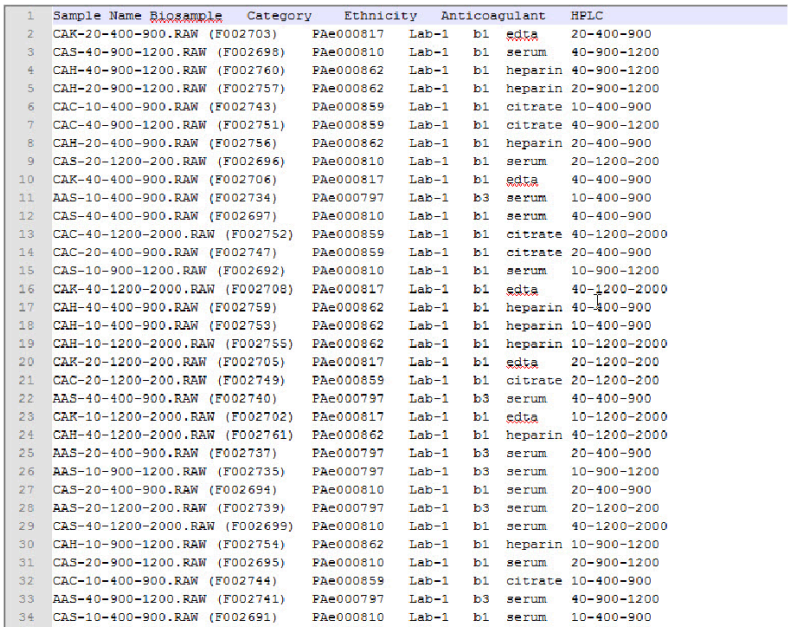
Compiling an Attributes File

The first line, or header line, in a Scaffold DIA 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 DIA.

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 6-9 below).

Figure 6-9: Example of a Scaffold DIA 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 DIA 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

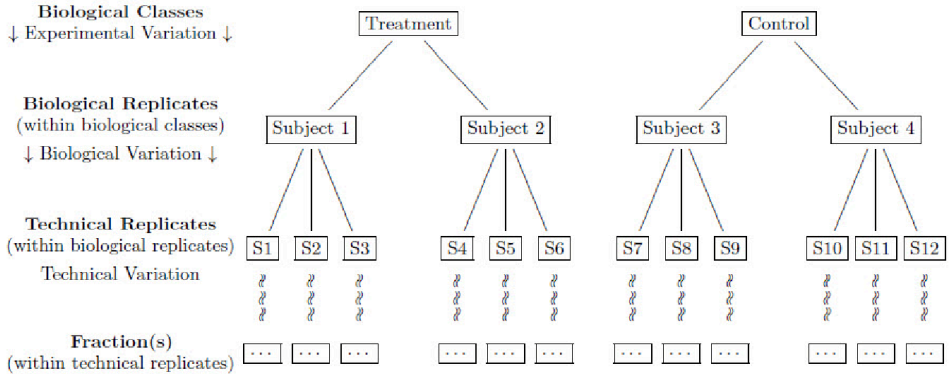
Scaffold DIA supports quantification and statistical analysis for several types of experimental designs.

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 6-10: Example of a basic experimental design



The diagram above shows an example experimental design which can be easily represented in Scaffold DIA with two categories in a simple hierarchy, for example:

Group -> {Treatment, Control}

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

It 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 DIA 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 DIA, 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.

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[9]. 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 DIA only supports complete randomized block designs, i.e. those which contain one value per cell in the Design Matrix.

After samples have been organized into Categories (see Organizing data in Scaffold DIA), 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.

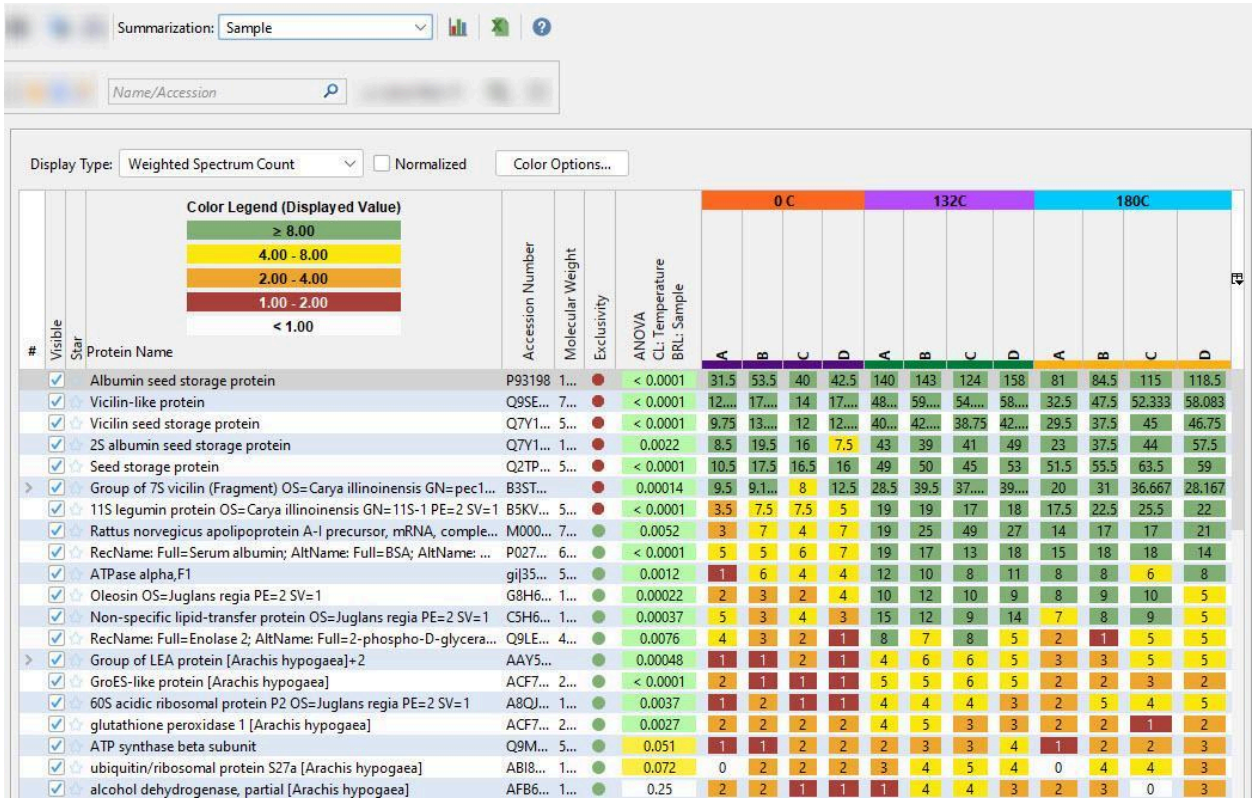
The Configure Sample Organization and Statistical Analysis Dialog

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

[9]https://www.statsdirect.com/help/analysis_of_variance/randomized_blocks.htm

Figure 6-11: 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 protein-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 6-12: The Samples View with a Summarization Hierarchy defined

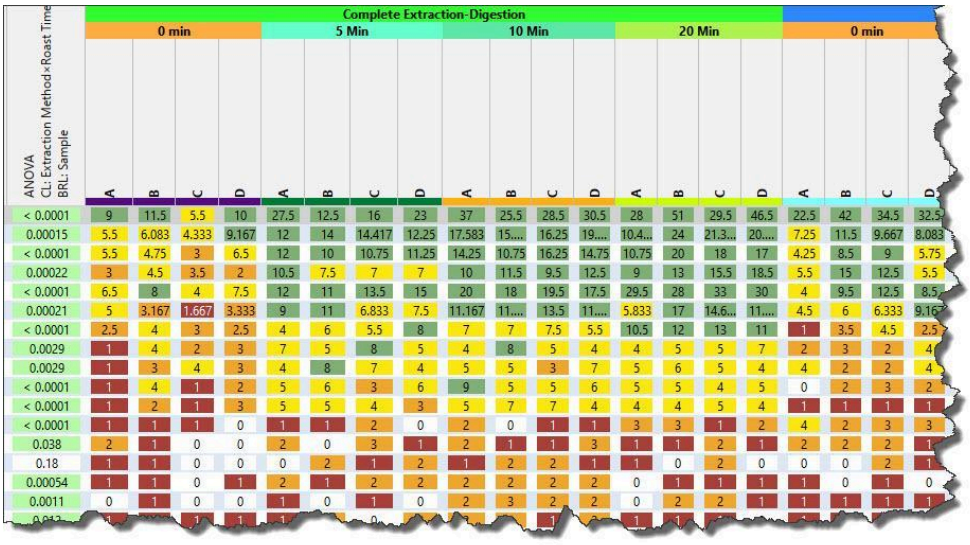
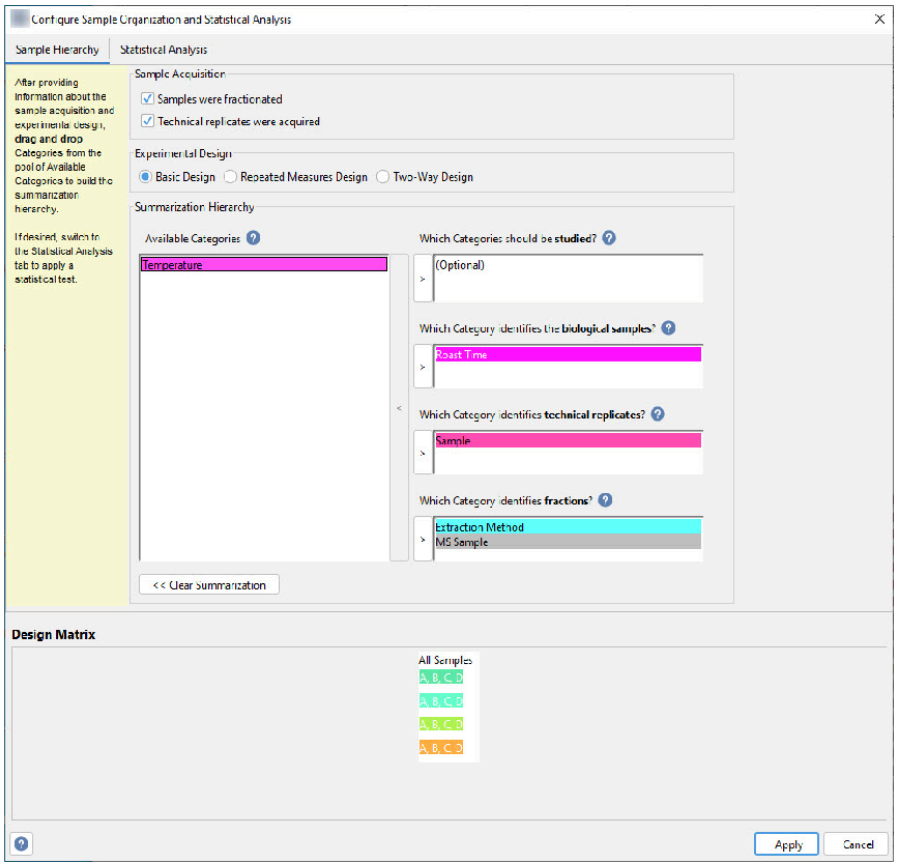


Figure 6-13: The Sample Hierarchy dialog when Basic Design is selected, showing fractions and technical replicates



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.

Design Matrix

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 checkboxes. 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.

Different experimental designs require different classes of Categories to be specified:

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.

Figure 6-14: Selection of two Categories to be studied

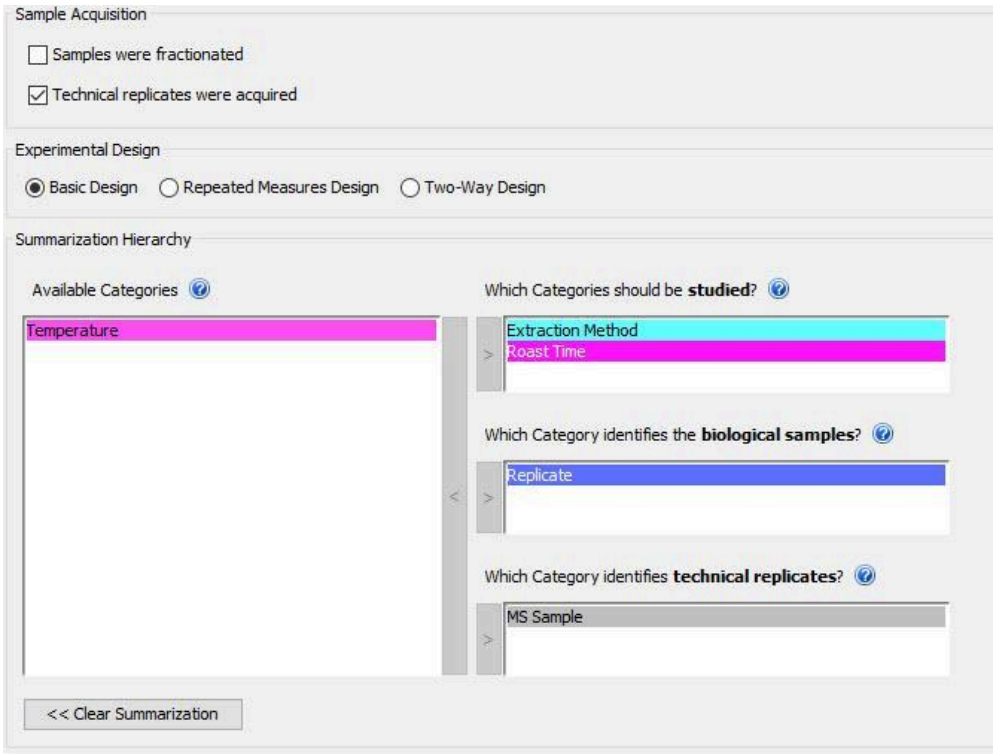
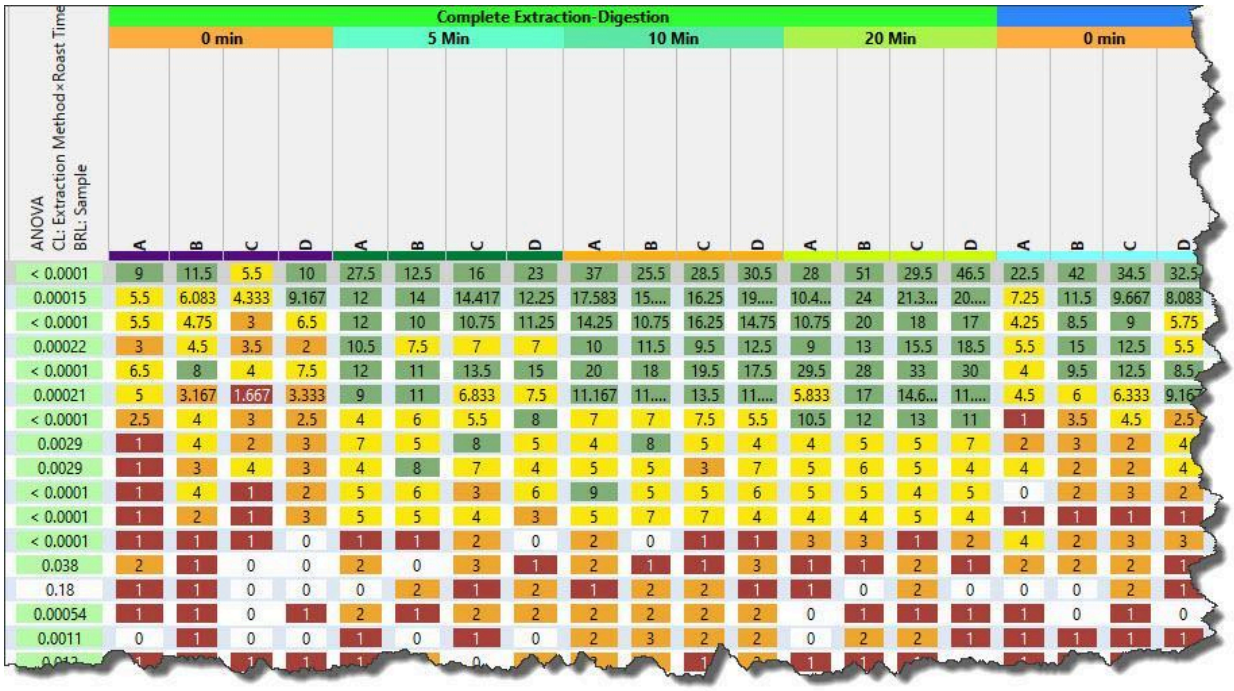


Figure 6-15: The resulting Samples View, shown at the Replicate level



ANOVA CL: Extraction Method×Roast Time BRL: Sample	Complete Extraction-Digestion																			
	0 min				5 Min				10 Min				20 Min				0 min			
	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
< 0.0001	9	11.5	5.5	10	27.5	12.5	16	23	37	25.5	28.5	30.5	28	51	29.5	46.5	22.5	42	34.5	32.5
0.00015	5.5	6.083	4.333	9.167	12	14	14.417	12.25	17.583	15....	16.25	19....	10.4...	24	21.3...	20....	7.25	11.5	9.667	8.083
< 0.0001	5.5	4.75	3	6.5	12	10	10.75	11.25	14.25	10.75	16.25	14.75	10.75	20	18	17	4.25	8.5	9	5.75
0.00022	3	4.5	3.5	2	10.5	7.5	7	7	10	11.5	9.5	12.5	9	13	15.5	18.5	5.5	15	12.5	5.5
< 0.0001	6.5	8	4	7.5	12	11	13.5	15	20	18	19.5	17.5	29.5	28	33	30	4	9.5	12.5	8.5
0.00021	5	3.167	1.667	3.333	9	11	6.833	7.5	11.167	11....	13.5	11....	5.833	17	14.6...	11....	4.5	6	6.333	9.16
< 0.0001	2.5	4	3	2.5	4	6	5.5	8	7	7	7.5	5.5	10.5	12	13	11	1	3.5	4.5	2.5
0.0029	1	4	2	3	7	5	8	5	4	8	5	4	4	5	5	7	2	3	2	4
0.0029	1	3	4	3	4	8	7	4	5	5	3	7	5	6	5	4	4	2	2	4
< 0.0001	1	4	1	2	5	6	3	6	9	5	5	6	5	5	4	5	0	2	3	2
< 0.0001	1	2	1	3	5	5	4	3	5	7	7	4	4	4	5	4	1	1	1	1
< 0.0001	1	1	1	0	1	1	2	0	2	0	1	1	3	3	1	2	4	2	3	3
0.038	2	1	0	0	2	0	3	1	2	1	1	3	1	1	2	1	2	2	2	1
0.18	1	1	0	0	0	2	1	2	1	2	2	1	1	0	2	0	0	0	2	1
0.00054	1	1	0	1	2	1	2	2	2	2	2	2	0	1	1	1	1	0	1	0
0.0011	0	1	0	0	1	0	1	0	2	3	2	2	0	2	2	1	1	1	1	1

Biological Samples

- 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.

Technical Replicates

- 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 DIA, they should be considered as technical replicates if the intent is to use them as multiple measurements of the same biological subject.

Fractions

- Which Category identifies fractions: this field appears only if the fractionation checkbox 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 DIA 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:

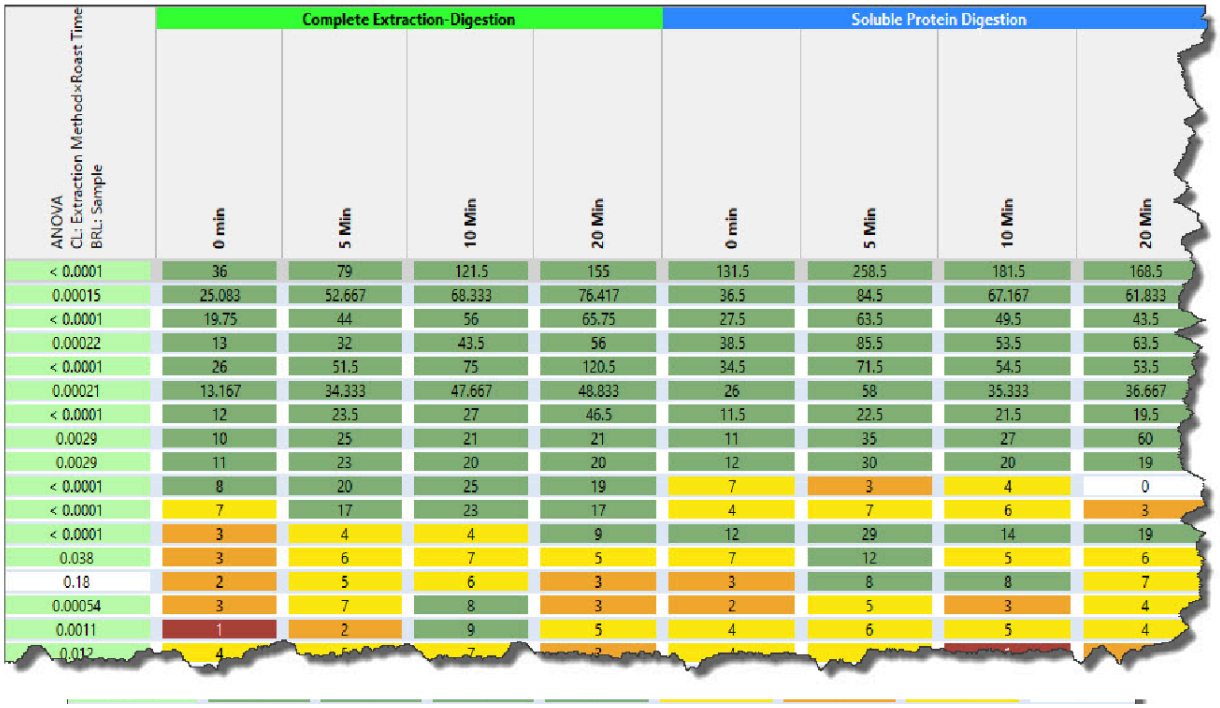
- Which Category should be considered as the Primary Analysis Category. This is generally the treatment or condition that is the main focus of the experiment.
- Which Category identifies the Secondary Analysis Category. Often this will be a category that defines a condition which is to be controlled 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 6-16.

Figure 6-16: 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

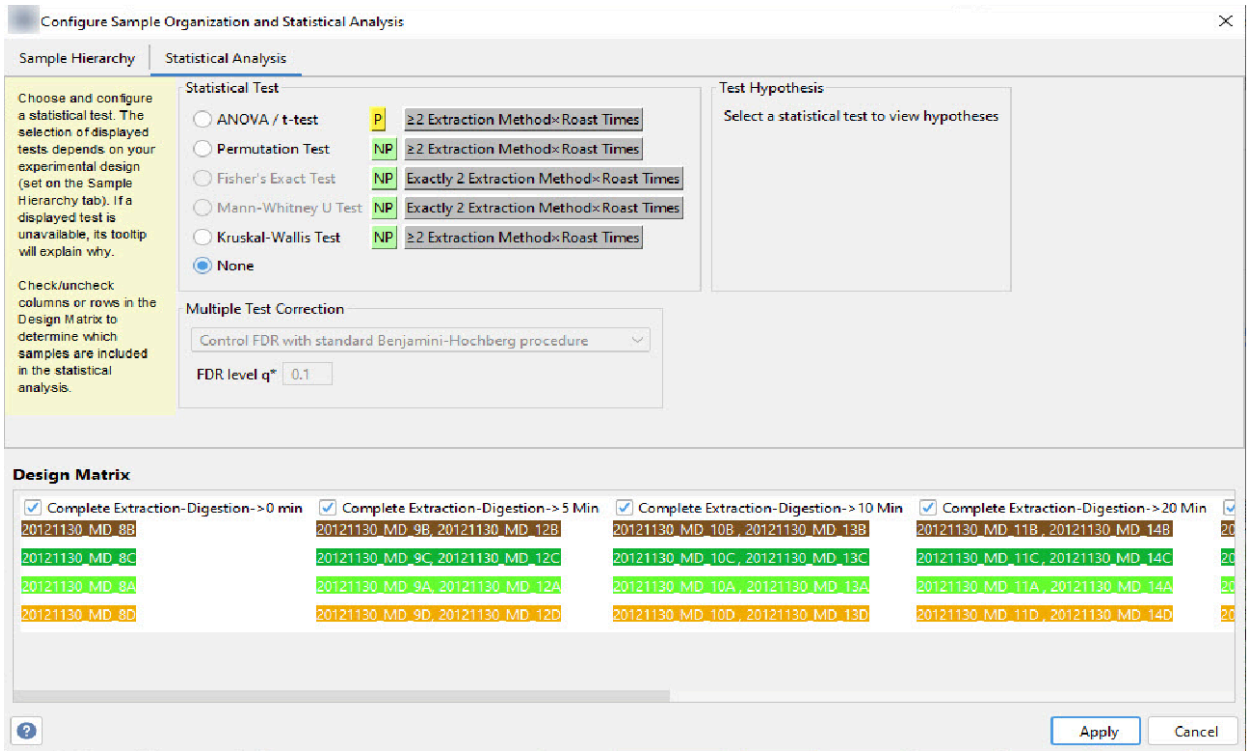
If the Technical replicates were acquired box is checked, the user must specify a category that represents the technical replicates. If each biological sample has been divided and the resulting sub-samples have been analyzed in the mass spectrometer separately, the sub-samples are technical replicates. 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 DIA, 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 Samples if 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 in cases in which the same biological samples have been analyzed in different ways and the results should be considered as a single sample for protein identification.

Statistical Analysis Tab

Figure 6-17: 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 drop down 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 drop down 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 7: The Proteins View

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

Inspection and Validation of peptides

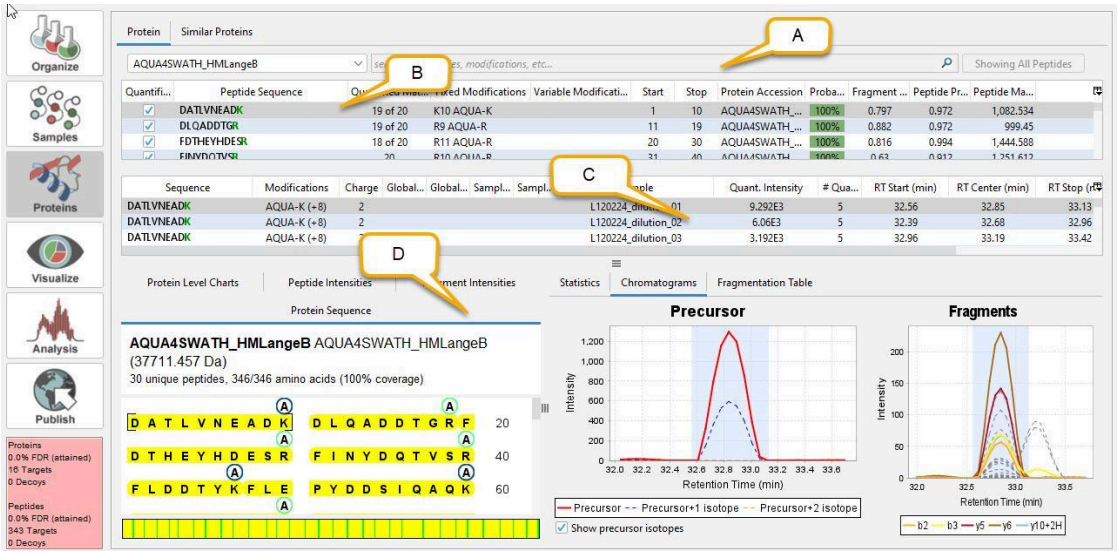
The Proteins View builds on the validation and visualization features of the Scaffold DDA Suite of programs, but has been enhanced to provide information about Similar Proteins, i.e. those that share peptides with the protein under consideration.

The Protein Tab

The tab Proteins View consists of four major panes:

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

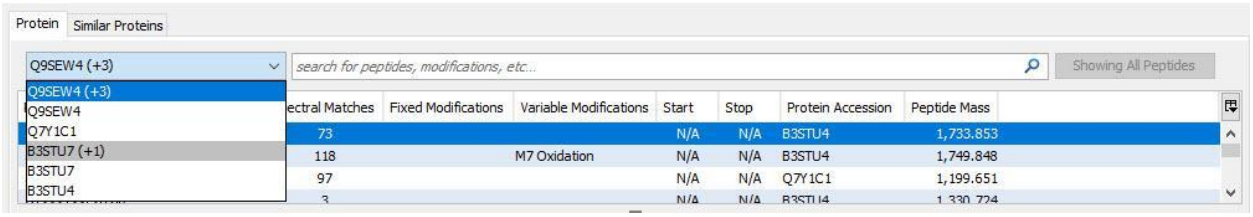
Figure 7-1: The Proteins View



The Peptides Filtering Pane

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

Figure 7-2: Peptides View filtering pane



Filtering Status

At the upper left 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 “Show All Peptides” and **clicking this button clears the filter(s)**.

Protein menu

Scrolling through the protein list allows examination of various proteins within a protein group or cluster 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.

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 number of quantified peptide matches, whether the peptide exclusively matches one protein, variable and fixed modifications, start and stop positions in the protein

sequence the Protein Accession, peptide probability, a quality metric based on fragment consistency (see Fragment Consistency Score), a quality metric based on consistency of the peptide's intensity profile with other peptides in the same protein (see Peptide Profile Correlation), 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.

Manual Validation

The details in the Peptide Match Pane and the tabs below it provide the opportunity for the user to examine the peptide evidence and the fragments that contribute to its quantitative values. If after examining this evidence, the user determines that the peptide is not a reliable basis for quantification of the proteins in which it appears, the peptide may be removed from quantitative calculations by unchecking the **Quantified Checkbox** in the Peptides table above. This removes the peptide from quantitative calculations across the entire experiment.

Peptide Quality Metrics

Two metrics are provided to assist in manual validation of peptides for inclusion in quantification. Each of these is designed to address a different potential problem:

Fragment Consistency Score

This metric is designed to help identify peptides which may be unsuitable for quantification because the pattern of the fragments used to calculate their intensity values is inconsistent. This may indicate that the peptide has been incorrectly identified, or that it is not reliably quantified. A figure to help in visualizing this problem is provided in the Fragment Intensities tab in the lower pane of the Proteins View.

The metric is calculated as follows:

For each peptide detected in an experiment consisting of n DIA samples, the intensities of each of the fragments assigned to the peptide are adjusted so that the sum of the intensities in each sample is one. From these intensities a matrix of fragment intensities is constructed, where each cell (M_{ij}) contains the adjusted intensity of the j^{th} fragment in the i^{th} sample.

Each column of the matrix constitutes a "feature vector" which represents the intensity of a single fragment across samples. An intensity-weighted average vector is computed, with the weights based on the log of the total intensity of each vector. For each feature vector, the euclidean distance from the average vector is computed. The intensity-weighted average of the distances is then calculated and converted to a meaningful score using the formula:

$$fcs = (1 - ((avg)/(\sqrt{2})))^2$$

If the pattern of fragment intensities is perfectly consistent between samples, the Fragment Consistency Score = 1. On the other hand, if each fragment's intensities are completely independent between samples,

then the Fragment Consistency Score is close to 0. Care should be taken in interpreting the fragment consistency score when there are only two samples.

Peptide Profile Correlation

This metric is designed to identify peptides which may be shared with other, possibly unidentified, proteins. In this case, the intensity of the peptide may not correctly reflect the relative abundance of the protein in question. To address this problem, the pattern of expression across samples for the specific peptide is compared to an average peptide profile for the protein. Peptides whose profiles are not consistent with those of the protein's other peptides receive a low score. A negative score indicates that the specific peptide's intensity pattern is inversely correlated with the others'.

This score is calculated by first summing peptide intensity values to the technical replicate level to support proper combination of fractions when fractionation has been performed. The peptide intensities are then normalized so that all values are in the range 0 to 1, and averaged across all peptides to give an average profile. The reported score for each peptide is the Pearson correlation between its profile and the average profile.

The Peptide Intensities tab in the lower pane of the Proteins View provides a visualization of the peptide profiles for a protein.

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 from there.

The Peptide Match Pane

The Peptide Match Pane provides detailed information about the DIA evidence for the peptide selected in the the Peptides Filtering Pane above in each sample in the experiment.

Displayed information

Columns available for display include the peptide sequence, modifications, charge, Global identification probability and q-value, Sample-specific identification probability and q-value, MS Sample, intensity value, number of quantitative peptides, RT Start, Center and Stop times, the theoretical (m/z) of the peptide and the Attributes that have been applied to this Sample. The columns to be displayed may be selected using the Column control.

The **Global Probability** represents the probability, as calculated by Percolator on an experiment-wide basis, that this peptide was present in the one or more samples in the experiment. Similarly, the **Global q-value** is the false discovery rate calculated by Percolator for this peptide on an experiment-wide basis. These values are shown only for the representative peptide match, selected as the best match among all samples.

The **Sample Probability** is the identification probability calculated by Percolator when considering only the specific sample represented by a given row in the table. The **Sample q-value** is the false discovery rate calculated by Percolator for that specific sample. The Sample q-value is used in Applying Quant Thresholds.

The Visualization Pane

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

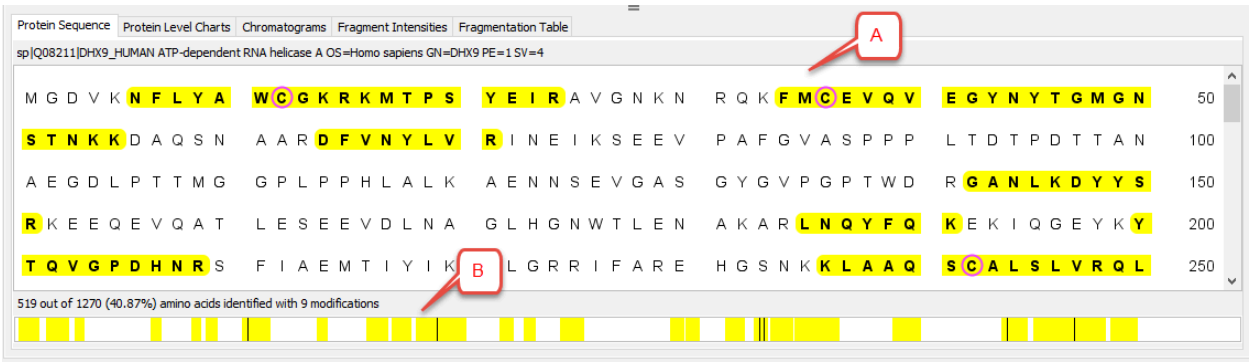
- The Protein Sequence tab
- The Protein Level Charts tab
- The Peptide Intensities tab
- The Statistics tab
- The Chromatograms tab
- The Fragment Intensities tab
- The Fragmentation Table tab

The tabs are arranged in two panes, left and right, and at any given time, one tab can be open in each. Tabs may be rearranged by dragging and dropping, either to rearrange the order within a pane or to move a tab from one pane to the other.

Protein Sequence tab

This tab consists of two different displays: the Protein Sequence display and the Coverage display.

Figure 7-2: Protein Sequence tab in the Proteins View. The Sequence Display (A) highlights the amino acids covered by identified peptides. Modifications are indicated by colored circles around the modified amino acids. The Coverage Display (B) appears below the sequence. It depicts the entire protein sequence, with covered regions highlighted, and modifications indicated by black bands.



Sequence coverage

The sequence coverage diagram displays coverage of the sequence of the selected protein by identified peptides. Since peptides are identified at the experiment level, all samples exhibit the same coverage. The peptide selected in the upper table of the Peptides tab is bracketed in black.

Sequence Display Options

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, while clicking on Show All Peptides at the upper left of the Peptides tab clears the filter. Note that clicking on an amino acid which is not covered by any peptides will result in the upper tables being empty.

Context menu

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

- Edit Modification Colors
- Outline Modifications - determines whether modifications are indicated by an open circle or a fully colored circle identifying the modified amino acid.
- Use Blinking Cursor - turns on and off blinking behavior of the brackets indicating the selected peptide.
- Copy Image - Copy the image to the clipboard which can be pasted into a third party tool such as Microsoft PowerPoint for easy editing and manipulation.
- Copy Sequence—Copies the protein sequence to the clipboard. You can then paste this copied sequence into a third-party application such as Microsoft Word.
- Save PNG... —Saves the currently displayed spectrum in a PNG format and opens the Write PNG bitmap image 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 an SVG vector format and opens the Write SVG vector image 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 an EMF format and opens the Write EMF vector image file dialog box in which you can specify the name and directory
- Print Protein—Opens the Print dialog box in which you can specify the options for printing (printer, number of copies, and so on) the protein sequence.
- 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.

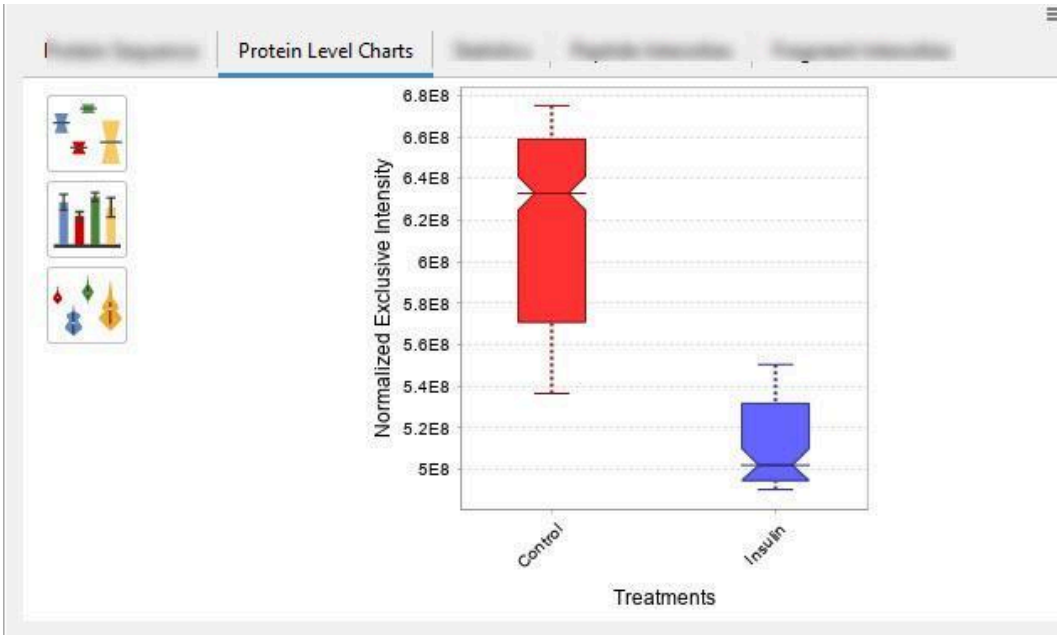
Protein Level Charts tab

The **Protein Level Charts tab** displays a graph that shows the Log_{10} Intensity for a protein in each of the MS samples or selected level of summarization as chosen from the Summarization Bar available in the Scaffold DIA main window.

Two formats are available for this graph—a Box plot and a Bar chart. Different colors are randomly assigned to each MS sample or selected summarization.

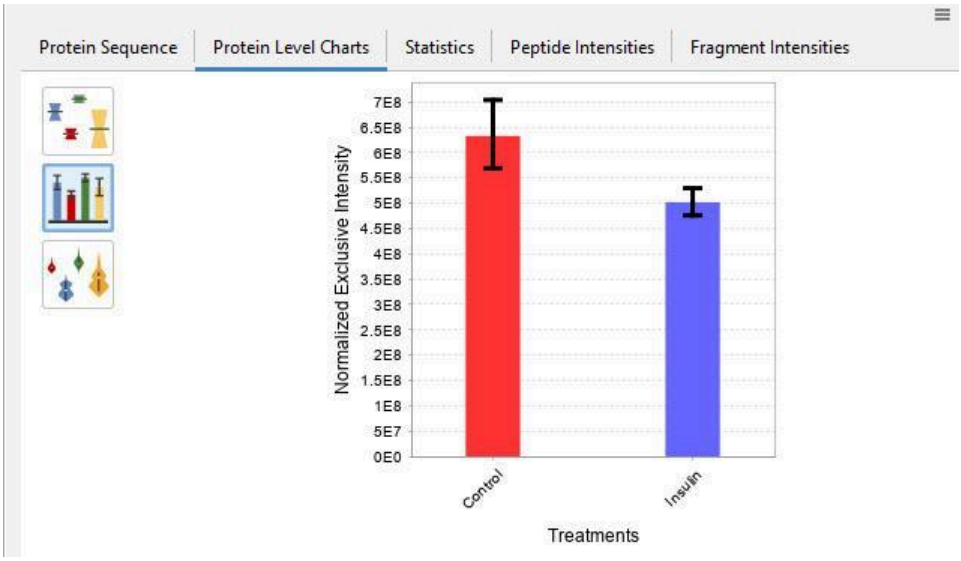
- Box plot—(The default plot) A box plot is a convenient way of graphically depicting groups of numerical data through their quartiles. In Scaffold DIA it displays the median value and range for the Log_{10} Intensity at the selected level of summarization. Placing the cursor on any box plot displays the information about the median and interquartile range for the corresponding quantitative value. When the selected summarization includes a single value in a sample or level of summarization only a line is shown at the corresponding intensity level. The percentage displayed below each box plot close to the x-axis indicates the percentage of missing spectral data. Quantitative analysis of samples that have a high percentage of missing spectral data might be unreliable.

Figure 7-3: Proteins View: Quantitative Charts tab - Box plot.



- Bar chart—The Bar plot displays the median value and range for the Log_{10} 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.

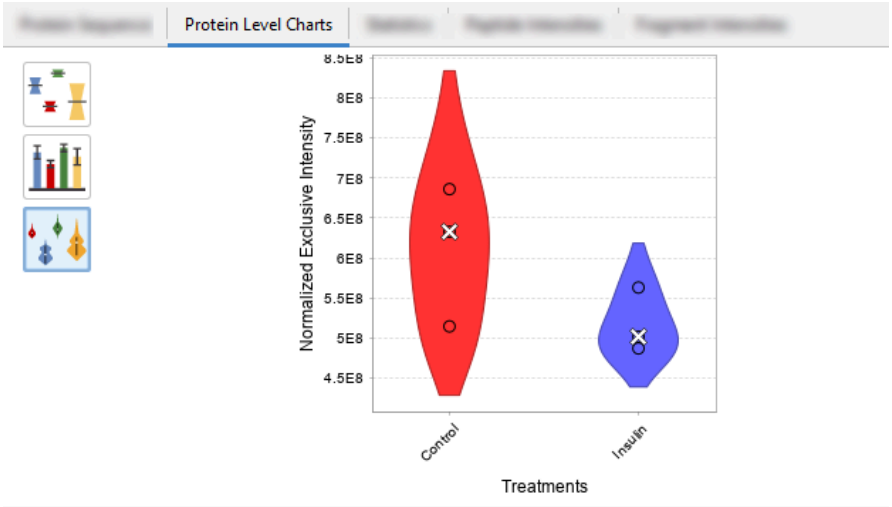
Figure 7-4: Proteins View: Quantitative Charts tab - Bar chart.



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

- 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.

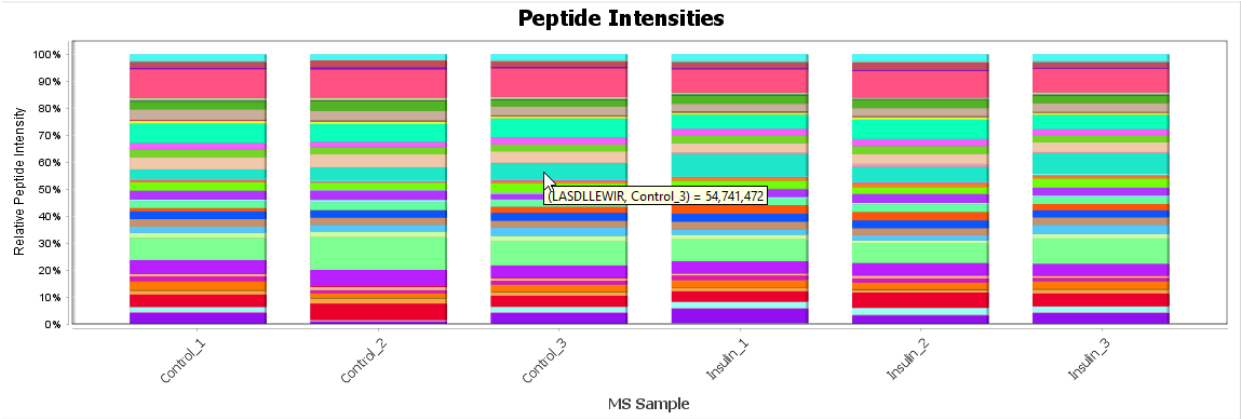
Figure 7-5: Proteins View: Violin Plot



Peptide Intensities tab

The Peptide Intensities tab displays a series of stacked bar charts depicting the portion of total intensity derived from each of the peptides used in quantification of the protein. Each chart represents a category at the selected level of summarization. Comparison of a peptide's contribution across categories may provide a measure of the reliability of that peptide for quantification. If a peptide is seen to be highly variable across categories, it may be excluded from the protein's intensity calculation by unchecking its Quantified checkbox.

Figure 7-6: Peptide Intensities chart



A legend is provided to identify the peptides, but it may become unwieldy if there are many peptides, and can be turned off via the Legend checkbox at the upper right of the tab. Hovering over a band in one of the charts displays a tooltip which identifies the peptide and the category and shows the intensity values for that peptide in that category. Clicking on a band selects the peptide and filters the Proteins View to display only that peptide. The display may be restored by clicking on Show All Peptides at the top of the tab.

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.

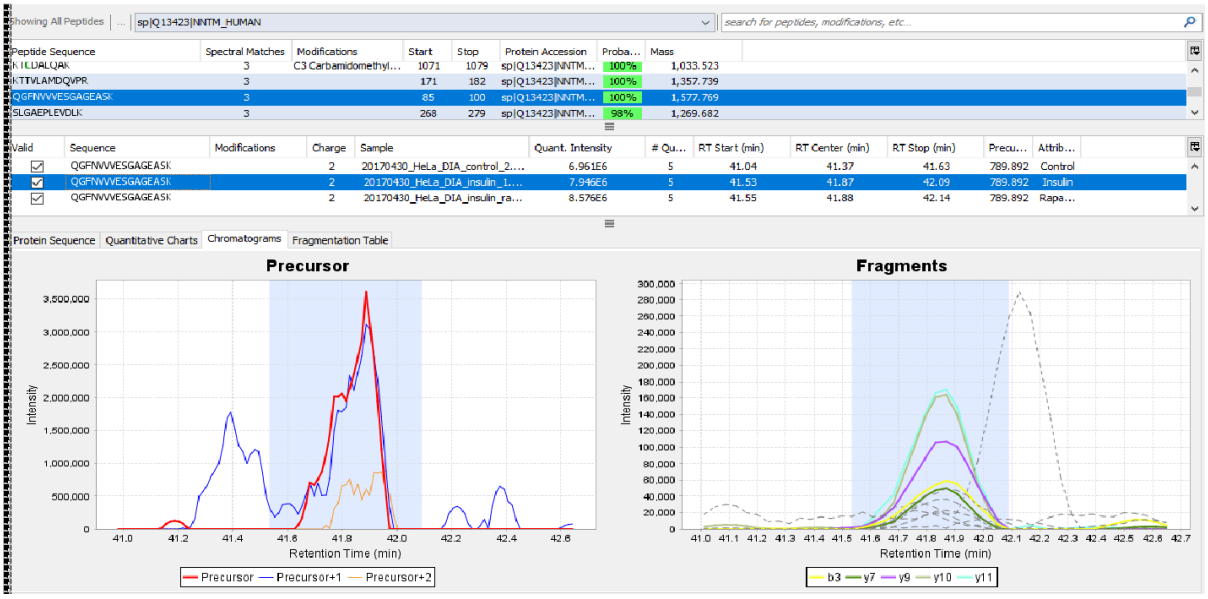
Chromatograms tab

The Chromatogram tab displays two plots. To the left is the Precursor plot, which graphs the intensity at the m/z corresponding to the m/z of the peptide, and of its +1 and +2 isotopic peaks in the Retention Time range in which the peptide identification was made.

To the right is the peptide Chromatogram, which shows the fragment peaks in the Retention Time window in which the peptide was identified. Peaks used for quantification are indicated in color and labeled in the legend below. Hovering over the peak in the graph highlights its label in the legend and vice-versa. Fragment ion colors are consistent when viewing the chromatogram of the same peptide across samples.

Unquantified peptides: Some peptides are identified by Percolator but cannot be quantified, often because they do not contain enough quantifiable fragment ions. In this case, Scaffold DIA shows the extracted chromatogram for the peptide match in the sample which received the highest Percolator score to allow the user to inspect the evidence for identification, but it does not display any chromatogram for the other samples.

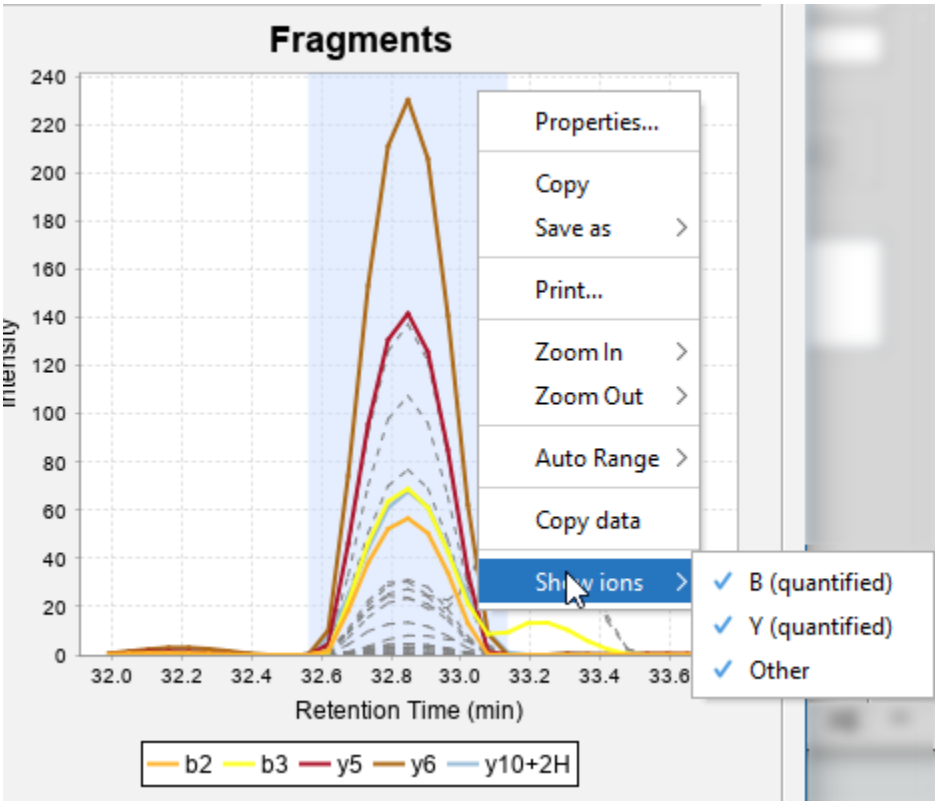
7-7: Proteins View Visualization Pane Chromatogram tab



Interacting with the Chromatogram tab:

- Hovering over a fragment peak highlights the corresponding entry in the legend. Similarly, hovering over a fragment name in the legend highlights the corresponding peak.
- Clicking and holding the left mouse button anywhere on the chromatogram or precursor plot and then dragging the mouse pointer to any position and then 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 7-8: Right-click context menu



- Properties - opens a dialog that allows the user to change graph properties such as title, axes, labeling, etc.
- Copy - 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 or PowerPoint.
- Save as - Allows the user to save bitmap 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 In - Allows zooming in one or both axes.
- Zoom Out—Allows zooming out in one or both axes, even beyond the original scaling.

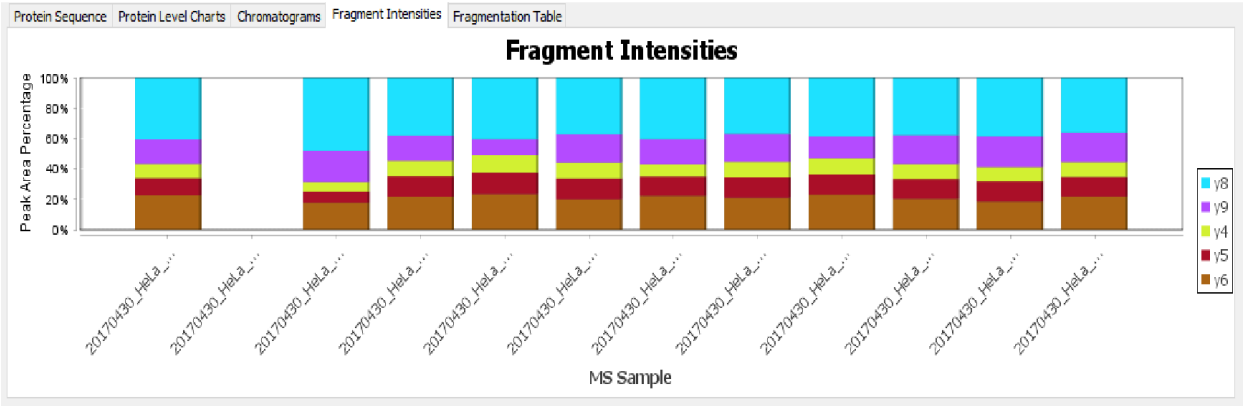
This is identical to a single click of the left mouse button after using the Click and Drag feature.

- Auto Range —Rescale the graph to the original level in the selected dimension.
- Copy data - Copies all of the data necessary to reproduce the graph. The data may be pasted into Excel.
- Show ions - Allows the user to select which ion type(s) should be displayed.

Fragment Intensities tab

This tab displays a stacked bar chart for each sample in the experiment. Each band in one of the bars represents the percentage of the total intensity for the peptide attributable to a specific fragment ion. This display allows the user to assess the consistency of the fragments in the peptide's chromatograms across samples, providing an indication of the reliability of the peptide for quantification. If no quantitative fragments were identified for a peptide, the bar will be missing for that sample.

Figure 7-9 Fragment Intensities chart:



Fragmentation Table tab

This tab displays information about the peptide 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.

The Similar Proteins Tab

The Similarity Table

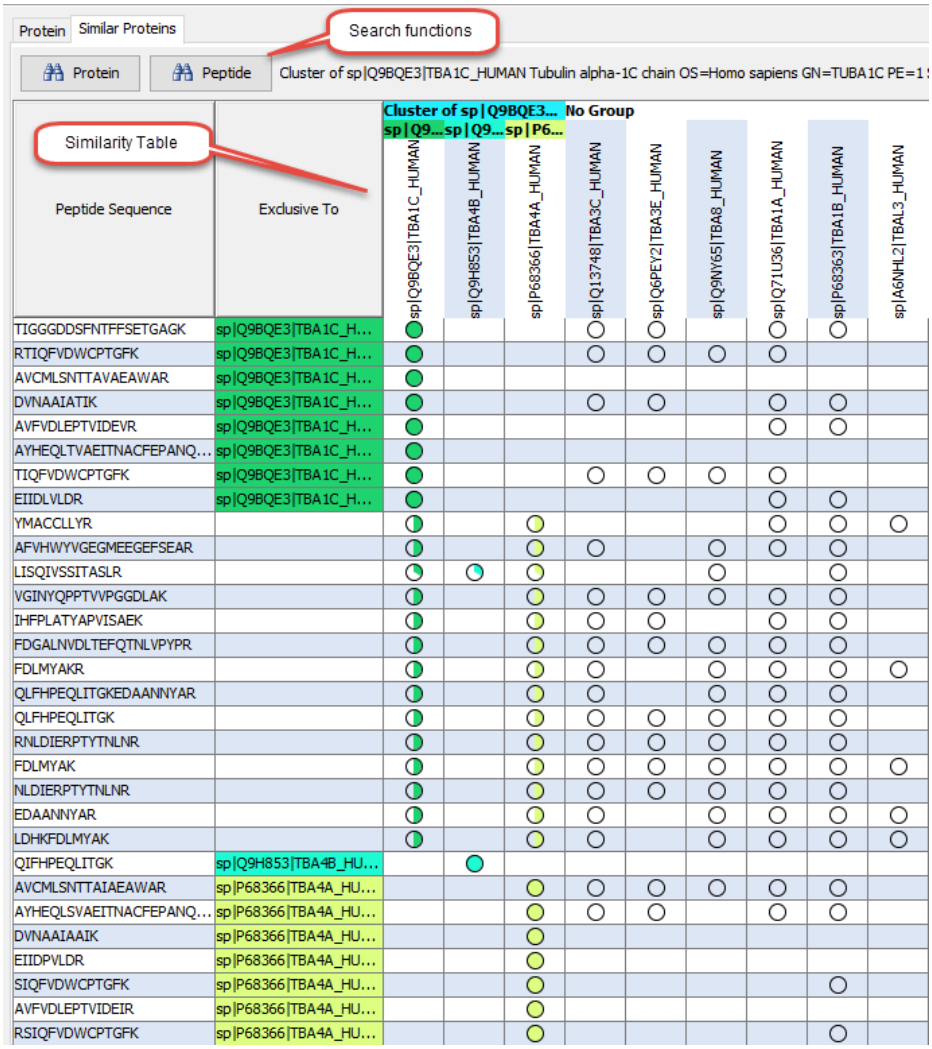
The Similar Proteins tab of the Proteins View allows the user to examine the role of shared peptides in protein identification and quantification. It provides insight into the results of parsimonious protein inference.

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.

Figure 7-10: The Similarity Tab



Search Functions in the Similarity View

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 View 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 8: 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: Principal Component Analysis, Quantitation, and a Heat Map of the filtered analytes list shown in the Samples table.

- Quantitation Tab, 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 and a plot that is intended to help the user assess the quality of the quantitation in the experiment.
- Principal Component Analysis tab, which helps identify the underlying sources of variation in the data set.
- Heatmap tab, 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 8-1.

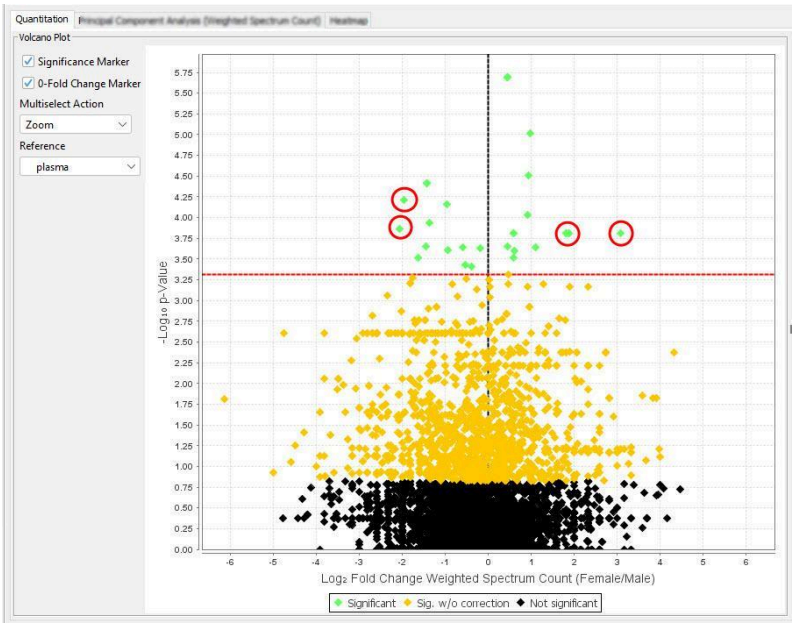
Plot Actions

A checkbox 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 checkbox, marks the point at which there is no difference between the compared values, i.e. the zero fold change line.

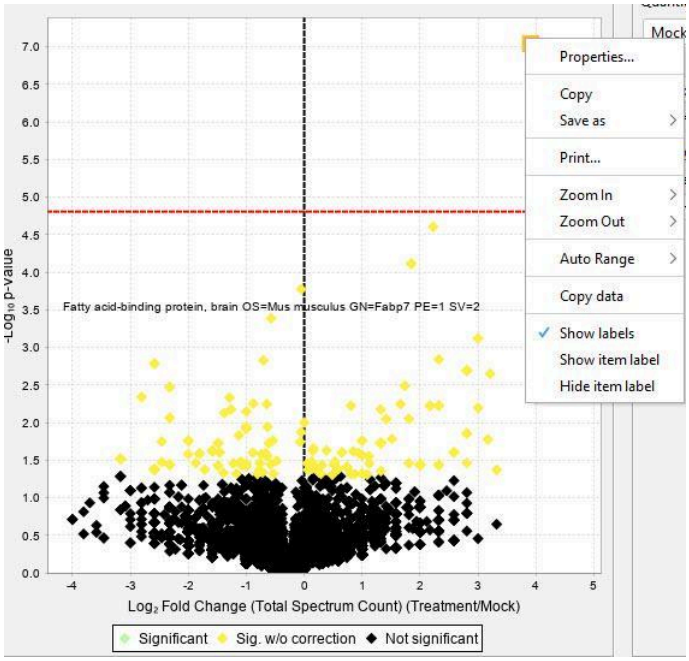
The **Multi-Select Action** pull down menu determines the behavior of the pane when the user selects a rectangular area in the plot by holding down the left mouse button and dragging the cursor. Depending on the option chosen, the graph may zoom in on the selected area or the proteins in the selected 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 8-1: The Volcano Plot



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 8-2: 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.

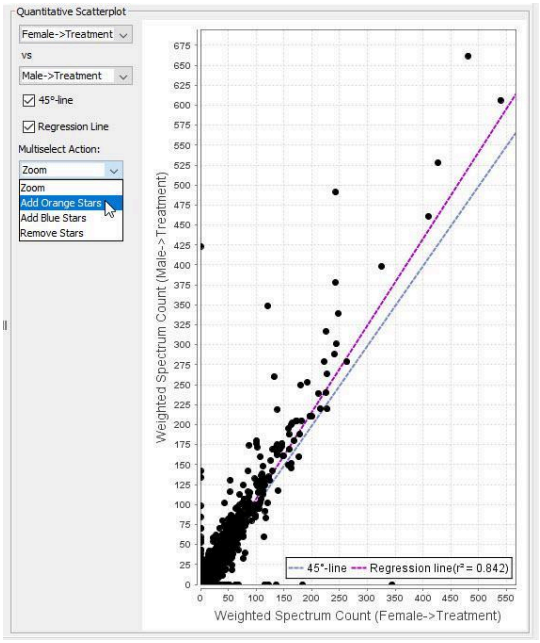
Examination of the Quantitative Scatterplot assists the user in assessing the relationship between the two categories, and in identifying outliers, which may be proteins that are especially important in distinguishing the two groups.

Two lines may be displayed on the graph, each activated by a checkbox 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 8-3: The Quantitative Scatterplot

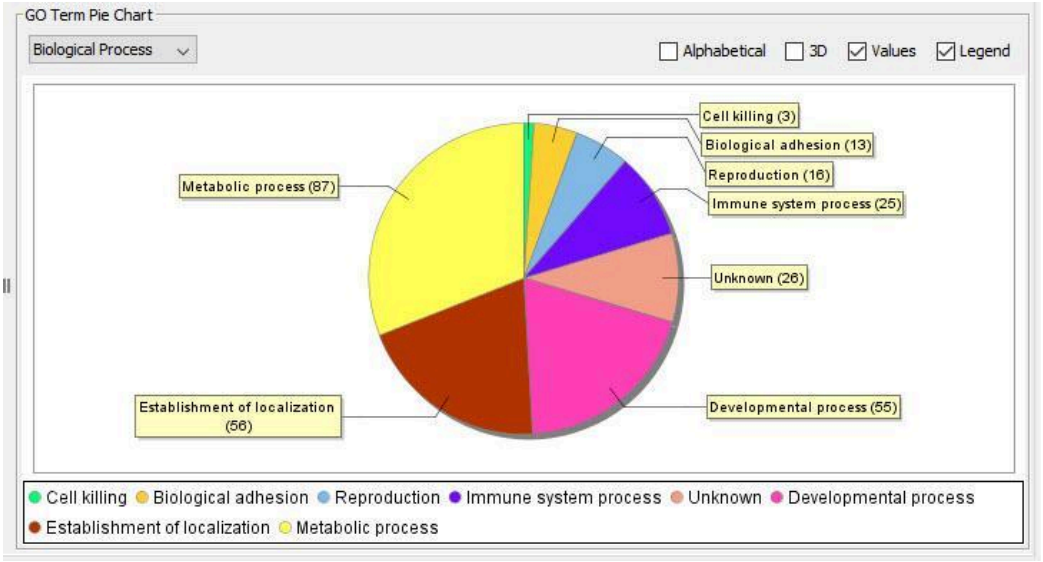


GO Annotations Pie Chart

The GO Annotations Pie Chart displays the relative proportions 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 checkboxes 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 8-5: 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 8-6: The Quantitative Trends Chart

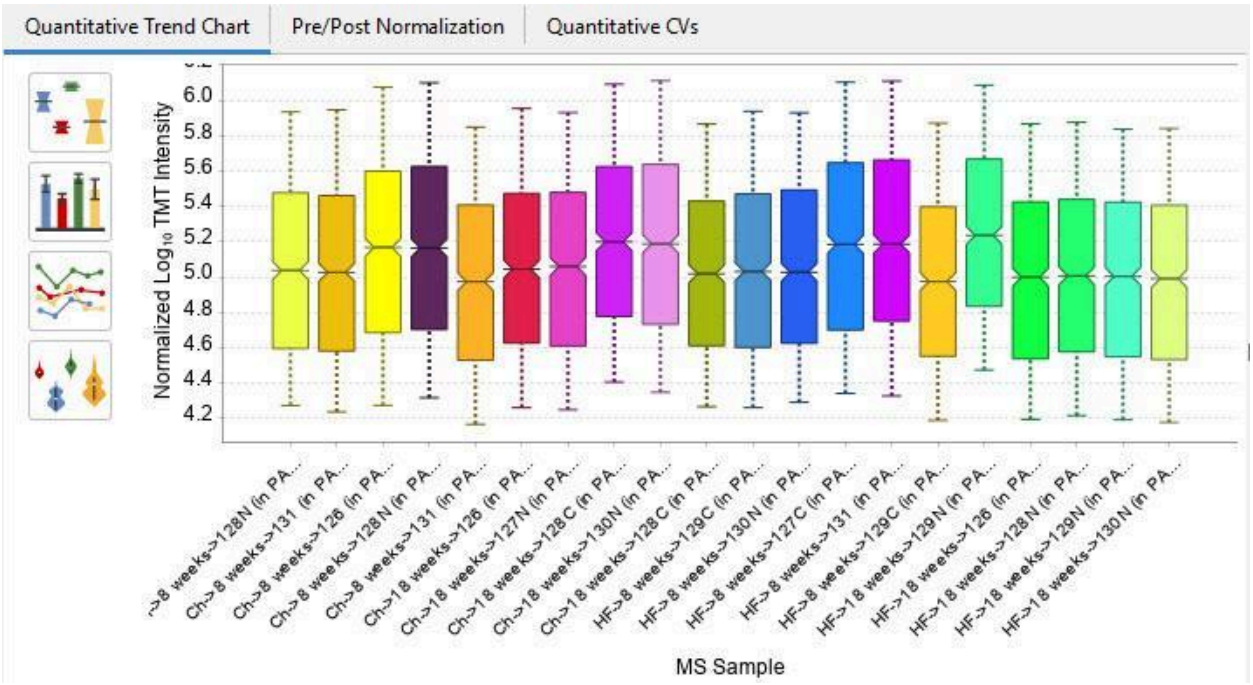
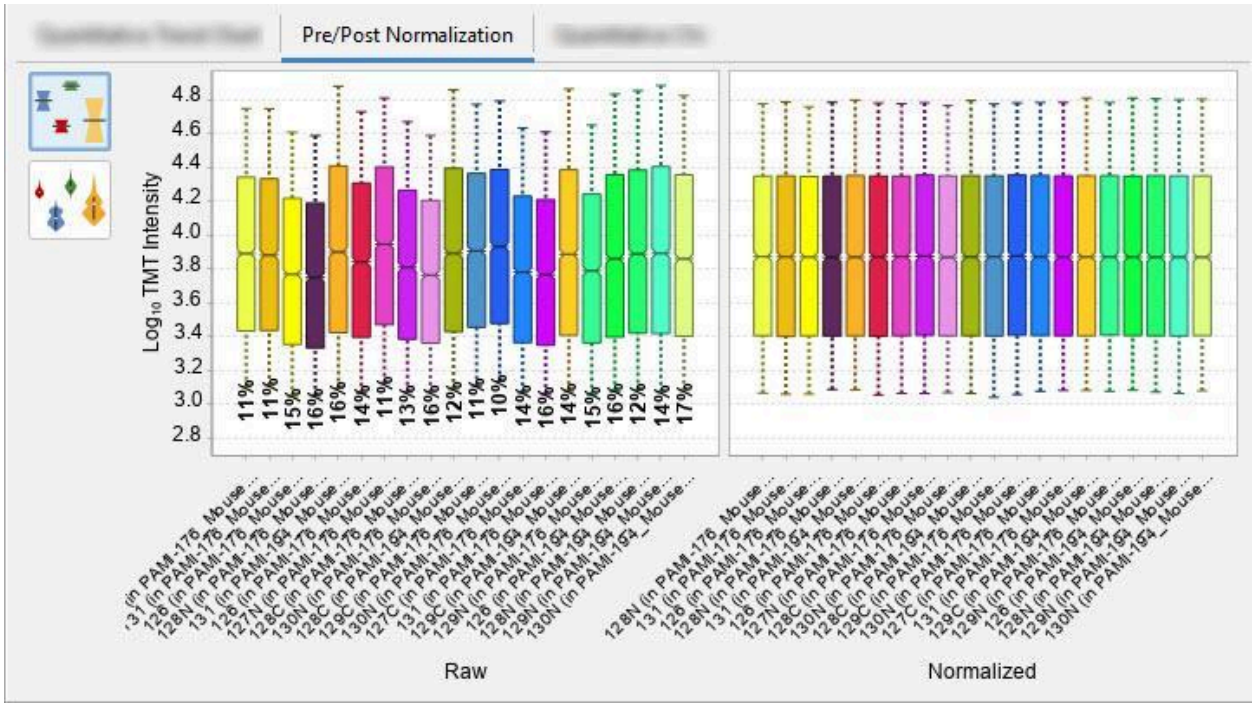


Figure 8-7: Pre/Post Normalization Box Plot



Quantitative CVs Chart

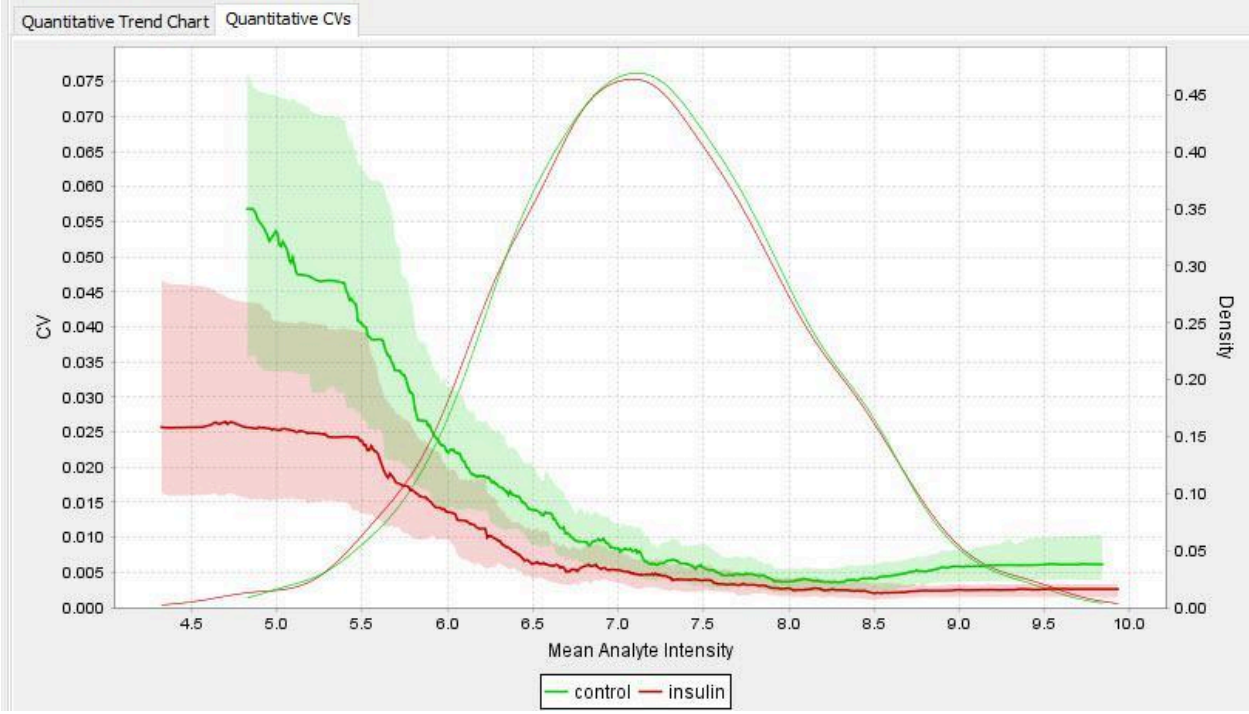
The Quantitative CVs chart contains two distinct types of plots, which, in combination, provide insight into the reliability of the quantitative values calculated in the experiment. The Quantitative CVs chart displays the relationship between mean protein intensity and Coefficient of Variation (CV) for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

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.

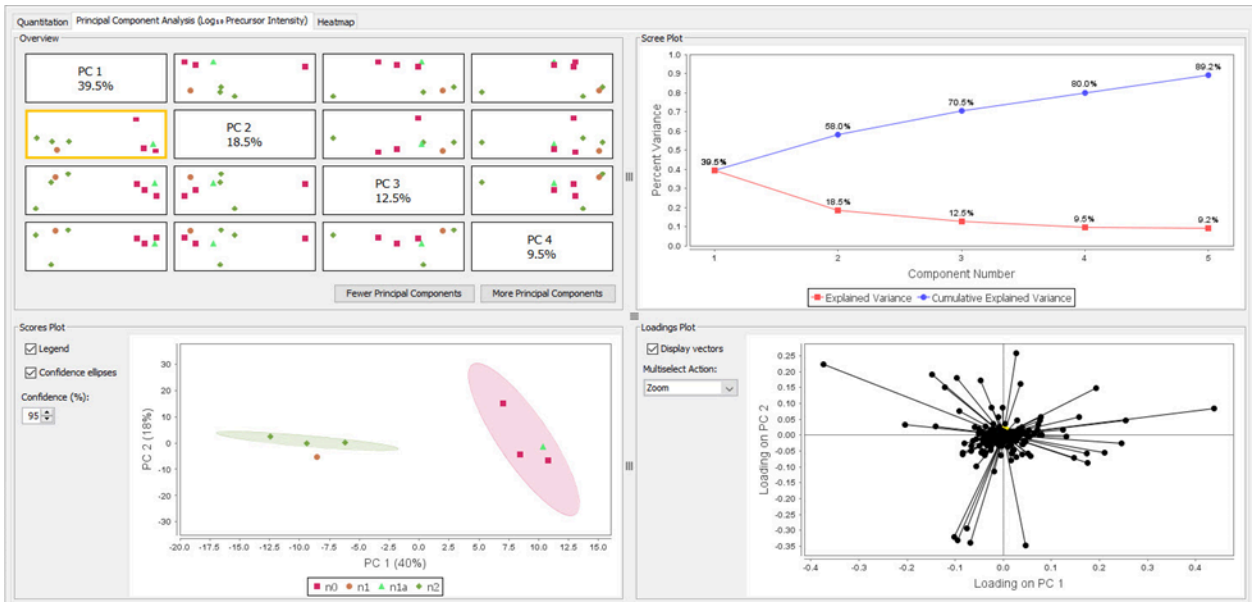
Figure 8-8: The Quantitative CVs Chart



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 DIA, this provides a powerful tool for exploring the biological meaning of quantitative differences observed in an experiment.

Figure 8-9: 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. checkboxes 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 of 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 DIA, see How PCA is Performed in Scaffold DIA.

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[10].

In order to produce a figure that is meaningful, Scaffold DIA 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 DIA constructs a heatmap from the data appearing in the Samples table and displays it in the Visualize View Heatmap tab.

Figure 8-10: Visualize View: Heatmap tab

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

[10]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

- 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. 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 DIA 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.

Figure 8-11: 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

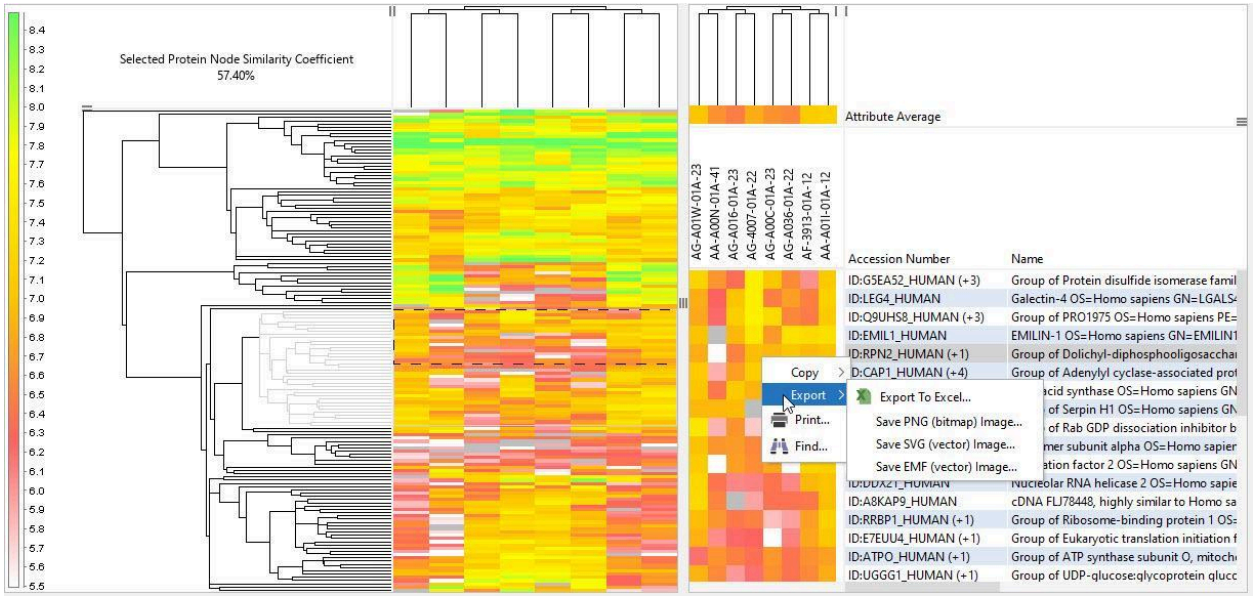
A right-click of the mouse while hovering over the Heatmap landscape provides a context menu with zoom and export options, see Figure 8-12.

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

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

The Export > Save 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 8-12: Heat map Context menu



Heatmap Details pane

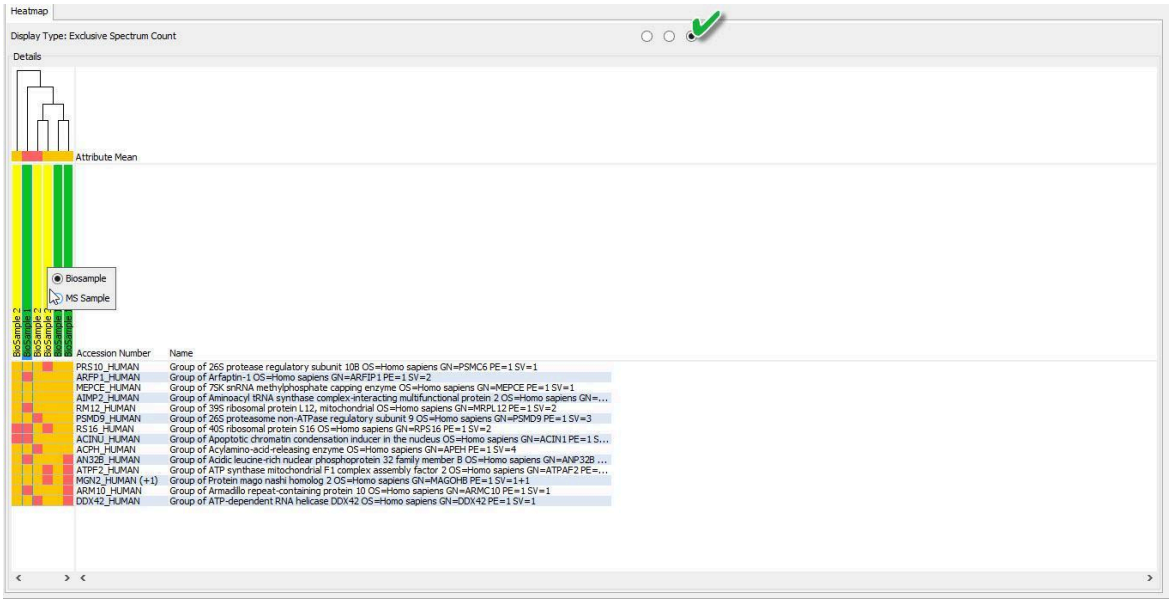
When sections of the Heat map are selected the Heatmap details pane is populated with the selected section of the map and the related information about the protein groups associated with each row and the related MS sample or selected level of summarization as chosen from the Summarization Bar available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

When no selection is active the pane will appear empty with instructions to help the user populate the pane.

Figure 8-13: 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 DIA 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.

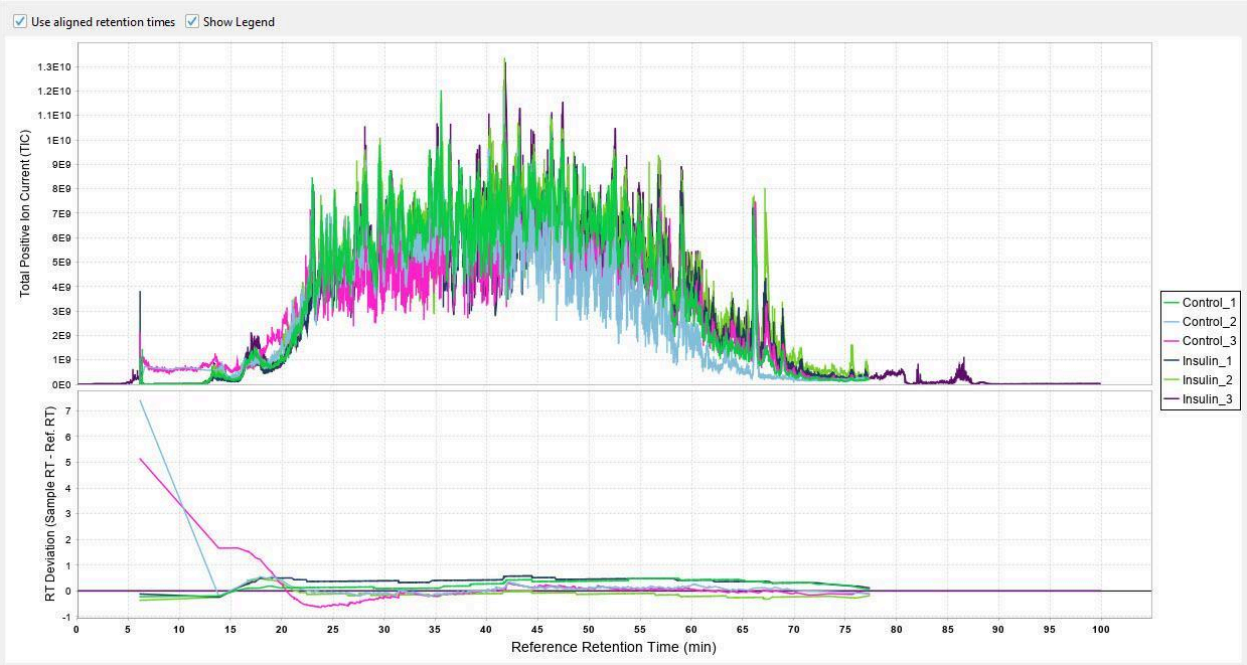
To turn column clustering on or off, select **Edit>Preferences>Heatmap** and check or uncheck the **Cluster Columns for Heatmap** box.

Chapter 9: The Analysis View

The Analysis View provides graphical tools to help the user visualize the samples alignment in retention time.

The view includes two tabs, one for each ionization mode. If the data is acquired in only one mode, the corresponding mode tab will be active while the other tab will be grayed out.

Figure 9-1: The Analysis View



This view contains two graphs, a legend and a check-box.

Total Ion Current (TIC) Chromatogram

The top graph in the tab plots the MS1 TIC for each MS sample in the experiment as a function of retention time. The TIC represents the summed intensity across the entire range of masses being detected at every point in the analysis. The chromatograms are drawn in the colors assigned to their corresponding samples, as described in the legend to the right of the graph.

The chromatograms can be displayed either aligned or unaligned by clicking the “Use aligned retention times” checkbox.

RT Deviation from Reference plot

The lower graph in the tab displays the differences between the raw retention times and the aligned retention times for each sample. This indicates the degree of adjustment that was necessary in order to align the sample to the reference at any given point.

When the “Use aligned retention times” box is checked, the differences are graphed relative to the reference retention times. When the box is unchecked, each sample’s deviations are plotted relative to that sample’s retention time. This allows a comparison between the upper plot and the lower, since they are always plotted on the same time scale.

The colors of the plots correspond to the colors assigned to the samples, as described in the legend to the right of the graph.

Raw samples legend

Provides information about the color assignment of each raw data file in the experiment.

Use aligned retention times checkbox

This checkbox changes the Y axis in both graphs. When the box is unchecked, values are plotted as a function of unaligned Retention Time. The unaligned version of the TIC Chromatogram provides a comparison of the chromatography in the various samples, and the RT Deviation plots below indicate the degree of adjustment that would be required at each point in order to align the chromatograms.

When it is checked, values are plotted against the Reference Retention Time. The TIC Chromatogram provides a view of the aligned chromatograms, and the RT Deviation plot indicates the degree to which each sample was adjusted in order to achieve alignment.

Chapter 10: The Publish View

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

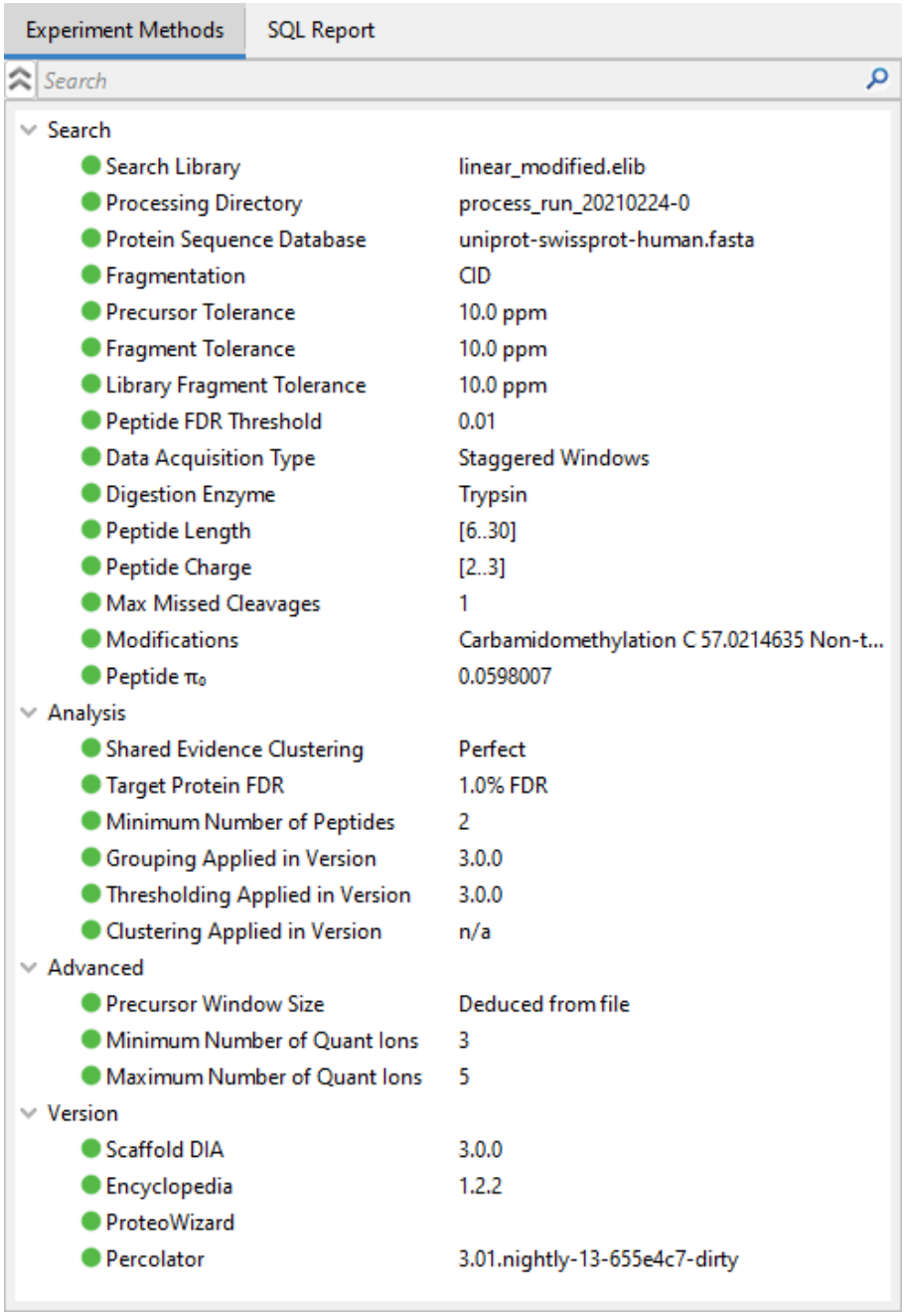
The Publish View has two tabs:

- 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.

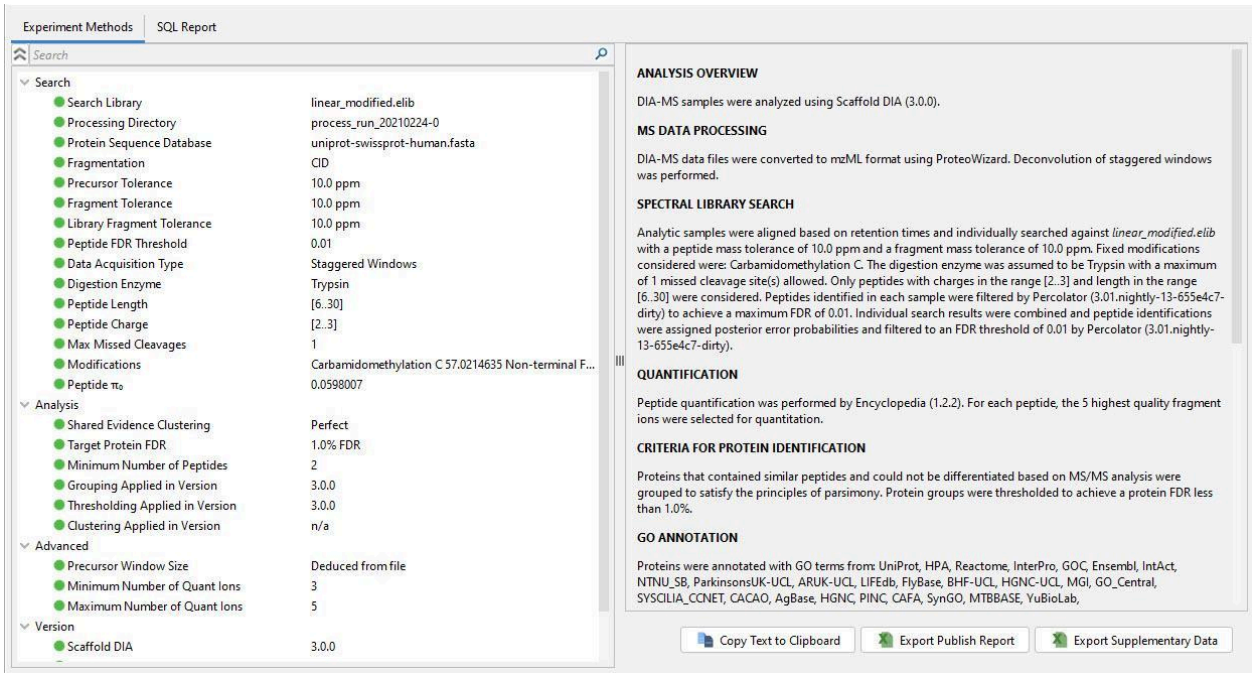
Figure 10-1: Experiment Information tab



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 10-2: The Analysis 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.

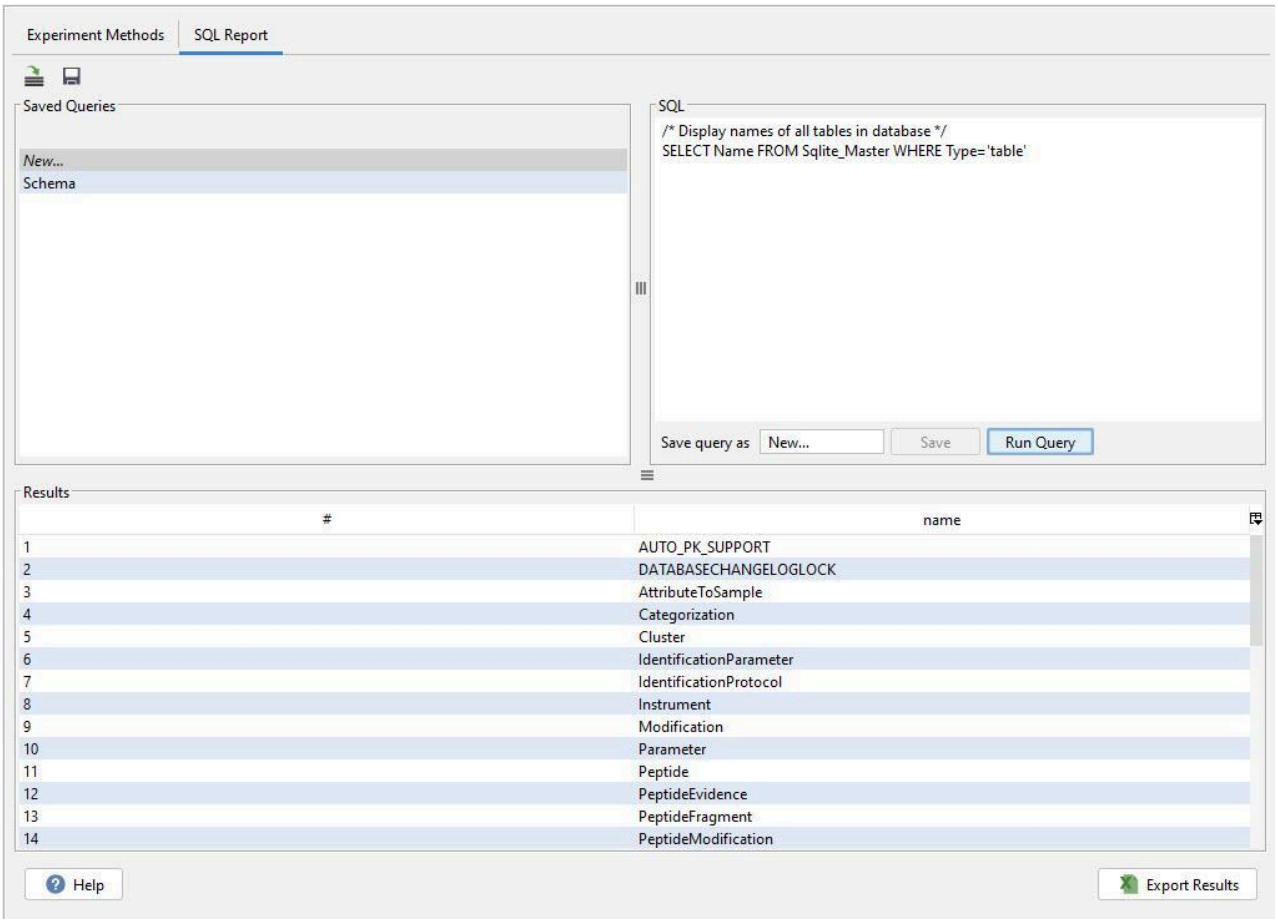
SQL Export tab

The experiment files created by Scaffold DIA, *.SDIA files, are essentially SQLite databases.

The SQL export tab is a SQLite graphical interface where a Scaffold DIA 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 *.SDIA file is shown in the Appendix in the Scaffold DIA 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 10-3: The SQL tab



The SQL Export tab contains four panes:

The SQL pane

Through the SQL pane it is possible to directly explore the information stored in a Scaffold DIA 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 Saved Queries pane.

Example:

List of tables available in *.SDIA 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 11: Quantitative Methods and Tests

Quantitative Methods

Scaffold DIA supports label-free quantitative methods based on fragment 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.

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

Peptide Counting

Scaffold DIA provides two peptide counting methods, selectable through the Display Type pull-down list:

- **Total Quantified Peptide Count** -- The number of different amino acid sequences that are associated with a specific protein including those shared with other proteins. Only the peptides that were Quantified by EncyclopeDIA are counted.
- **Exclusive Quantified Peptide Count** --The number of different amino acid sequences, regardless of any modifications that are associated only with a single protein group. Only the peptides that were Quantified by EncyclopeDIA are counted.

When counting, Scaffold DIA includes every peptide which is considered quantified. Quantified 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.

Peptide Counting Normalization

Scaffold DIA normalizes peptide 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 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 quantified peptides in all of the columns divided by the number of quantified peptides in a column.

Fragment Intensity Quantification

Quantification in EncyclopeDIA, and hence in Scaffold DIA, is based entirely on the intensities of the fragment ions in the chromatograms. Precursor intensity is not considered.

Selecting Fragment Intensities

In order to minimize the effects of fragment ion interference on peptide quantification, EncyclopeDIA performs automated transition refinement to select the fragment ions which are most likely to provide accurate quantitative information for the peptide. Peaks which represent fragment ions of a peptide and which do not exhibit interference would be expected to show a consistent shape across retention times in the detection window. As a result, all potential fragment ions in each sample are assessed for consistency and assigned an interference score. A threshold is applied to determine which fragment ions may be used for quantification. A user-specified parameter (see The Workflow Dialog) determines the number of these fragments that are actually used in quantification, and selection from among the qualified fragments is based on higher intensity.

Occasionally, fragments which were not provided in the spectral library that was searched may be quantified. This is because EncyclopeDIA also considers the theoretical fragmentation pattern of the peptide, and if additional fragment ions from the theoretical spectrum are identified, they may be selected for quantification.

Calculation of Peptide Intensities

The quantitative value of a peptide in a specific sample is the sum of the integrated areas of the selected quantitative fragment peaks. Before this calculation is carried out, however, the fragment ion areas are adjusted by subtracting areas which are attributable to background noise.

Selecting Peptides for Quantification

Peptides are selected for quantification on an experiment-wide basis. The fragment interference scores for a given peptide are summed across samples and the resulting scores are assessed, and the peptide is either accepted for or excluded from quantification in all samples. Peptides which do not meet a user-specified threshold (see The Workflow Dialog) for the number of quantitative fragments are not used for quantification.

Normalization

In Scaffold DIA, normalization is applied at the technical replicate (TR) level as defined by the user in the experimental design dialog. Technical replicate (TR) groups may contain one or more MS Samples. For more information on how to define a TR attribute group see Specifying the Design of an Experiment. Within each TR group, an ion is quantitatively characterized by summing all of its peptide intensity values in all of the MS Samples in the TR group.

When normalization is not applied, the peptide intensities within each TR group are summed and, if requested, the $\log_{10}$ is applied to the result. This is the value that is shown in the Samples table when Normalization is not selected.

When Normalization is selected, the summed intensities from the different MS Samples are first $\log_{10}$ transformed and then normalized by the procedure described in the Normalization procedure.

After normalization is performed, the ion intensities are transformed back into exponentiated form and the peptide 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 requested, 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 Missing Values).

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.

Scaffold DIA provides a method for imputing missing values (see Missing Values).

Quantitative tests

Scaffold DIA 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) and the Statistical Analysis 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.

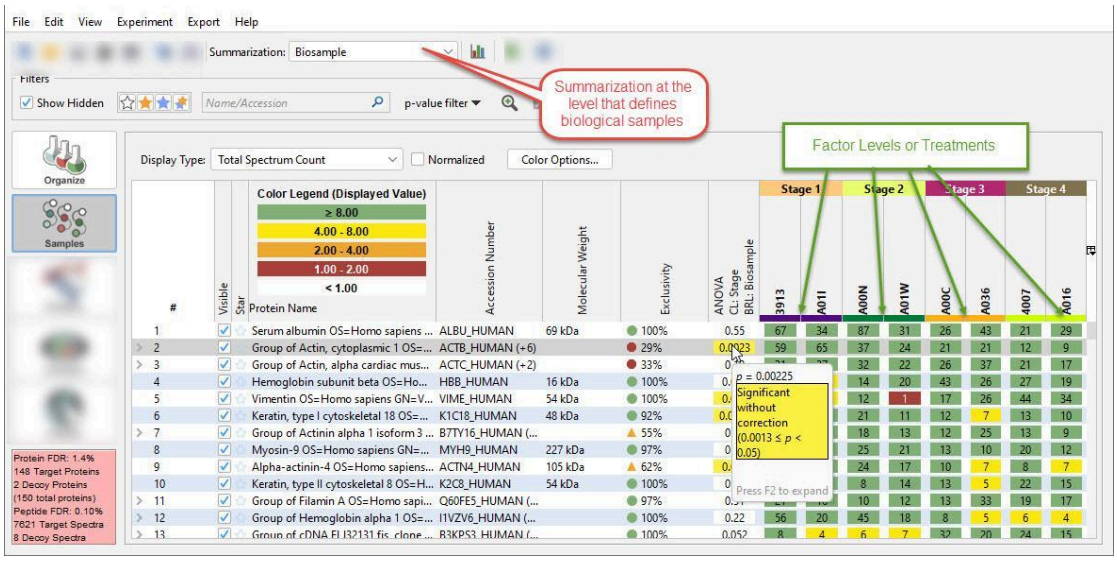
Not all factor levels or treatments need to be used for a specific test; only factor levels with a selected checkbox in The Configure Sample Organization and Statistical Analysis Dialog are used in computing the test. This can be useful if the user wants to exclude one or more treatments from the quantitative 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 11-10: 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 Basic Design experiments

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 DIA 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.

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 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.

For unbalanced experiments, in which groups have different numbers of samples, Scaffold DIAuses a Type III Sums of Squares analysis.

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

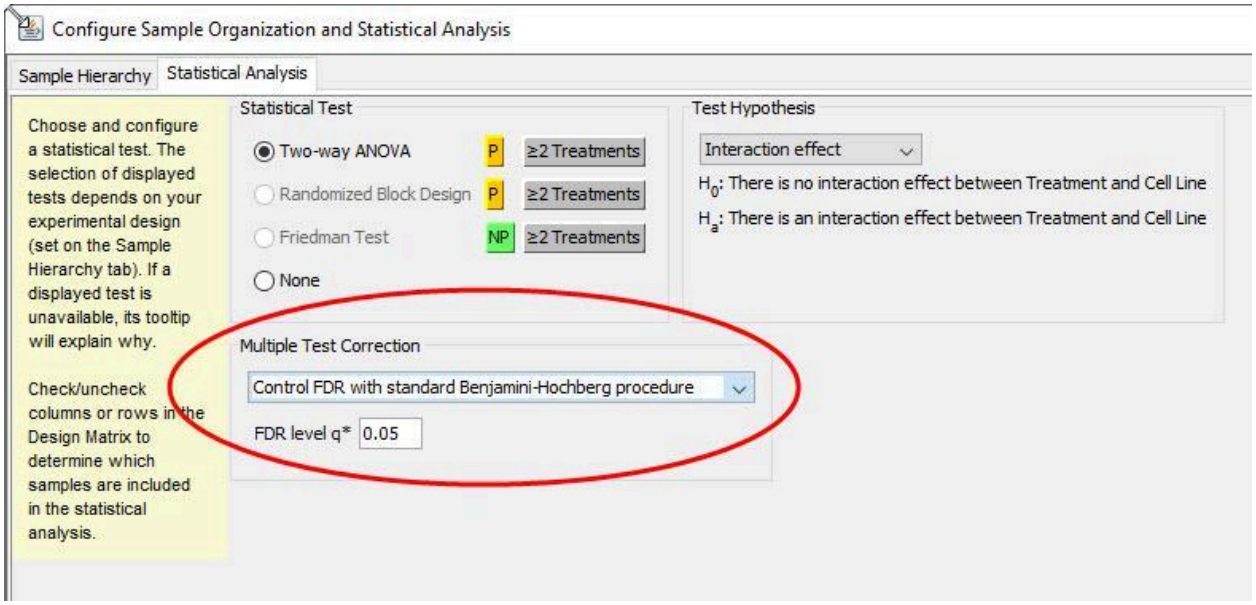
The Friedman test is a nonparametric statistical hypothesis test for assessing whether the population mean ranks of repeated measurements differ when the measurements have been taken under at least three conditions (e.g., time points). It is a nonparametric alternative to the Repeated Measures Analysis 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 11-11: 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 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 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 DIA 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 DIA 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 in the appendix.

When this option is selected the significance level is expressed in terms of o and the related text box appears underneath the pull down list. This text box allows the user to set the significance level 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 DIA computes the FDR using the Benjamini-Hochberg procedure as developed in the original paper[11].

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.
- 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.

[11]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.

- 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 insufficient number of samples to for a two-way ANOVA.

Reference Samples

The Reference Samples section of the Statistical Analysis tab allows the user to specify how values will be calculated for Fold Change display options.

Reference Sample selection

The drop down is used to select which sample or category will be used as the reference. This means that the values for that sample or category will determine the denominator when calculating the ratios that are used in computing fold change values.

When the experimental design specified in the Sample Hierarchy results in a Design Matrix with more than one row, there are two possible methods of computing fold changes:

- **Use all reference samples**

  In some experiments, each category represents a group of independent measurements and fold changes should be calculated by comparing to the average of all values in the reference category. In this case, the user should select Use all reference samples.

- **Use matched reference samples**

  In other experiments, especially in repeated measures studies, the fold change is used to measure the change from the baseline value for a particular individual or category. In this case, the radio button for Use matched reference samples should be selected. Note that if there is no reference value for a given reference sample or category, the fold changes values related to that reference will appear as missing values.

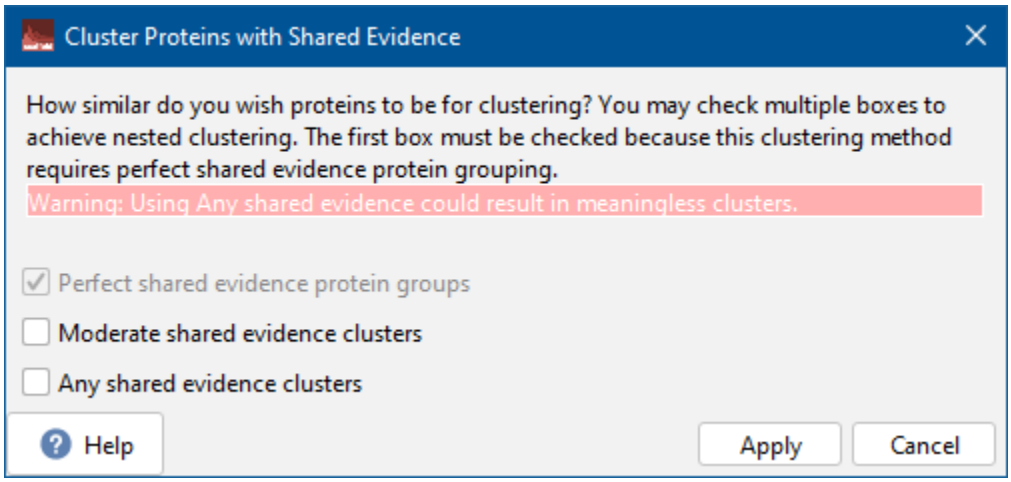
Use matched reference samples is the default value.

Chapter 12: Protein Grouping and Clustering

Grouping and clustering options

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

Figure 12-1: Protein clustering menu options



Experiment >Protein Clustering > Shared Evidence Clustering...

When this menu option is selected, the dialog box **Cluster proteins with shared evidence** opens up. It contains three checkboxes the combination of which defines the level of peptide evidence clustering Scaffold DIA applies to the protein list. The warning in the dialog reminds the user to be cautious when choosing “any shared evidence” since this method often results in large clusters of proteins with little in common.

- **Any shared evidence clustering** -- This option corresponds to a level of clustering $\lambda=0$, where λ is a parameter that defines the level of sharing of peptides among proteins in a cluster, for more details see the following appendix: Level of sharing. A cluster includes any pair of proteins A and B with a level of sharing greater than zero. This option tends to create a smaller number of large clusters.
- **Moderate shared evidence clustering** -- This option corresponds to a level of clustering $\lambda = 1/3$. A cluster includes any pair of proteins A and B with a level of sharing $L(A,B) \geq 1/3$. This option creates smaller clusters.

- **Perfect shared evidence protein groups** -- This option corresponds to a level of clustering $\lambda = 1$. A cluster includes any pair of proteins A and B with a level of sharing $L(A,B) = 1$. It corresponds to the definition of a protein group where two proteins share the same peptides. The option is grayed out because it aways needs to be selected to be able to apply the other levels of clustering. Scaffold DIA automatically groups proteins using the perfect shared protein grouping.

When both the **Moderate and Any shared evidence clustering** options are selected, sub-clusters will be formed within the larger clusters created by the **Any shared evidence clustering** option

Chapter 13: Reports

A variety of reports are available in Scaffold DIA 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 .csv file 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 DIA main menu. Each report is saved in a .csv format.

Reports can be specified in a .workflow file by using the `EXPORT-TYPE` and (optional) `EXPORT-LOCATION` parameters. The `EXPORT-TYPE` should be a comma-separated list of values from the set {`samples`, `peptides`, `quant`, `matches`, `publish`} or `EXPORT-TYPE=all` for all reports. If no `EXPORT-LOCATION` is specified, the reports are written in the same directory as the sdia file.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 DIA 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 DIA report in Excel or another spreadsheet application, or using a text editor.

The following reports are available in Scaffold DIA:

- Export Current View to Excel
- Export Attributes File
- Export Samples Report to Excel
- Export Peptide Report to Excel
- Export Peptide Match Report to Excel
- Export Publish Report to Excel
- Export MzTab Report for PRIDE
- Export Workflow

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.

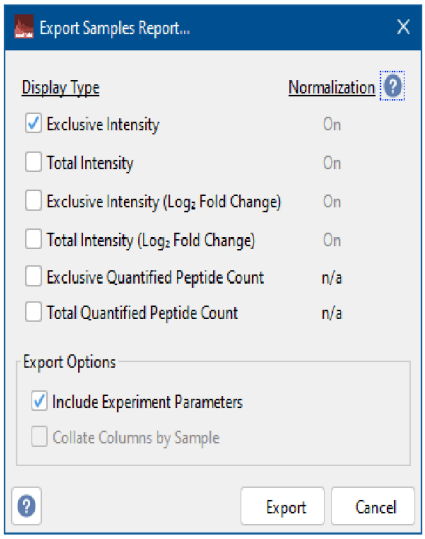
Export Samples Report to Excel...

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.

Normalization

The Normalization column indicates whether normalization is on for each Display Type. Normalization status must be set by selecting the Display Type in the Samples View and toggling the Normalize checkbox. It cannot be changed through this dialog.

Figure 13-1: Export Samples Report dialog



Export Peptide Report to Excel...

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

Export Peptide Match Report to Excel...

Generates a table, in comma-delimited format, which contains information about each peptide-match for all peptides included in the current thresholded experiment. This is similar to the information in the lower table in the Proteins View, for all proteins. This report does not contain a header with experimental parameters.

Export Publish Report to Excel...

Generates a comma-delimited report of the information contained in the Publish View Experiment Information table.

Export MzTab Report for PRIDE...

Generates a tab-delimited text file containing the results of the analysis in the mzTAB[12] format. This is a standard format created by the HUPO Proteomics Standards Initiative. The **PRIDE**[13] repository accepts an mzTAB file, along with .mzml files for all samples included in the analysis, as a "Complete Submission."

Export Workflow

Generates a workflow file capturing all of the user-specified parameters represented in the current experiment. This allows the user to save the workflow used to create an experiment after the data has been analyzed.

Run SQL Query for Export...

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

[12]Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N, Cox J, Neumann S, Fan J, Reisinger F, Xu QW, Del Toro N, Pérez-Riverol Y, Ghali F, Bandeira N, Xenarios I, Kohlbacher O, Vizcaíno JA, Hermjakob H. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. Mol Cell Proteomics. 2014 Oct;13(10):2765-75. doi: 10.1074/mcp.O113.036681. Epub 2014 Jun 30. PMID: 24980485; PMCID: PMC4189001.

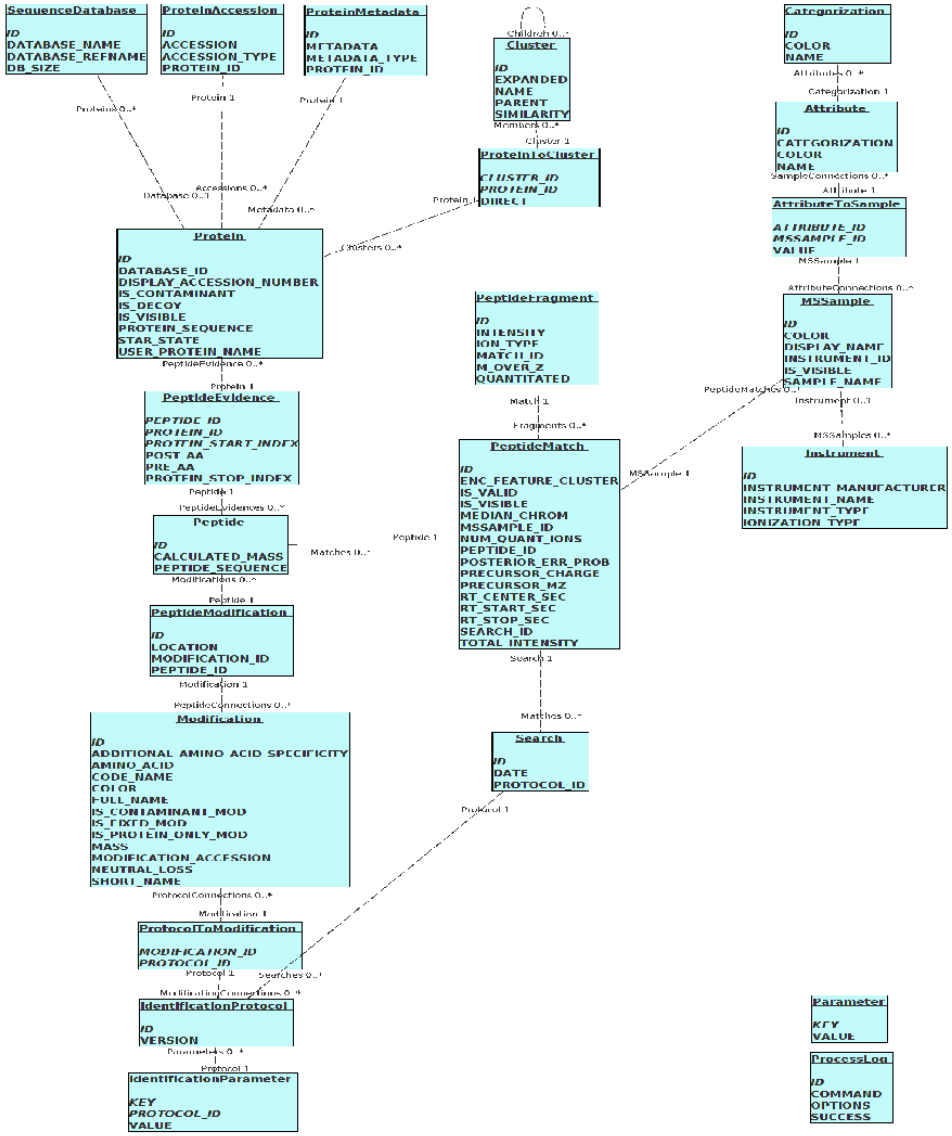
[13]https://www.ebi.ac.uk/pride/

Appendices

- Appendix A. Structure of Scaffold DIA files (*.SDIA)
- Appendix B. Computation of FDR in Scaffold DIA
- Appendix C. Summarization: Rolling up Values
- Appendix D. Missing Values
- Appendix E. Shared Evidence Clustering Algorithm
- Appendix F. Heat map clustering
- Appendix G. Techniques to Control the Family-wise Error Rate
- Appendix H. Using Principal Component Analysis in Scaffold DIA
- Appendix I. How PCA is Performed in Scaffold DIA
- Appendix J. Description of Mouse Right Click Context Menu Commands

Appendix A: Structure of Scaffold DIA files (*.SDIA)

Figure 1: Scaffold DIA experiment files, *.SDIA are SQLite files. This figure shows the general schema of a SDIA file. SQlite queries of this database can be submitted through the Export SQL table dialog box that can be opened from the menu **Export > Export SQL Table to Excel**.



Appendix B: Computation of FDR in Scaffold DIA

The user specifies the target Peptide FDR threshold in the Workflow Dialog, and it is passed to Percolator, which filters the set of identified peptides reported by EncyclopeDIA. The target Protein FDR threshold is set during loading in the Analysis tab of the Workflow Dialog, but may be adjusted through the Threshold controls. The Attained FDR values for peptides and proteins, which are reported in the FDR Dashboard, may differ somewhat from these target thresholds.

The attained values are calculated as follows:

All protein groups in the experiment meeting the identified peptide criterion are listed in descending order by their protein group scores. Some of these matches are target proteins, while others are decoy proteins. Starting from the top of the list, for each row, Scaffold DIA calculates the current protein FDR by dividing the number of decoy proteins in the list so far by the number of target proteins 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 Protein FDR Threshold. All proteins 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 is intended to produce the largest set of proteins possible while maintaining a protein FDR at or below the threshold.

Any peptides that were not associated with a protein in the thresholded protein list are then eliminated from the experiment. Because some peptides that passed threshold are not included in the proteins that pass threshold, Scaffold DIA recalculates the peptide FDR value from the final thresholded set.

For consistency, Scaffold DIA calculates Attained Peptide FDR with the formula used by Percolator:

$$PeptideFDR = \Pi_0 \cdot (T/D) \cdot (DThresh/TThresh)$$

where

Π_0 is the estimated proportion of target PSMS that are incorrect, reported by Percolator

T is the # of target peptides in the experiment

D is the # of decoy peptides in the experiment

DThresh is the # of decoy peptides in the thresholded set

TThresh is the # of target peptides in the thresholded set

Attained Protein FDR is simply calculated as DThresh/TThresh. This is a conservative estimate of protein FDR, which is consistent with the peptide FDR calculation since the number of decoy proteins generated by EncyclopeDIA is equal to the number of target proteins and we assume a conservative protein $\pi_{0 \text{ of } 1.}$

The results are displayed in the FDR Dashboard. Note that occasionally this procedure might result in a peptide FDR that is slightly higher than the selected Peptide FDR Threshold.

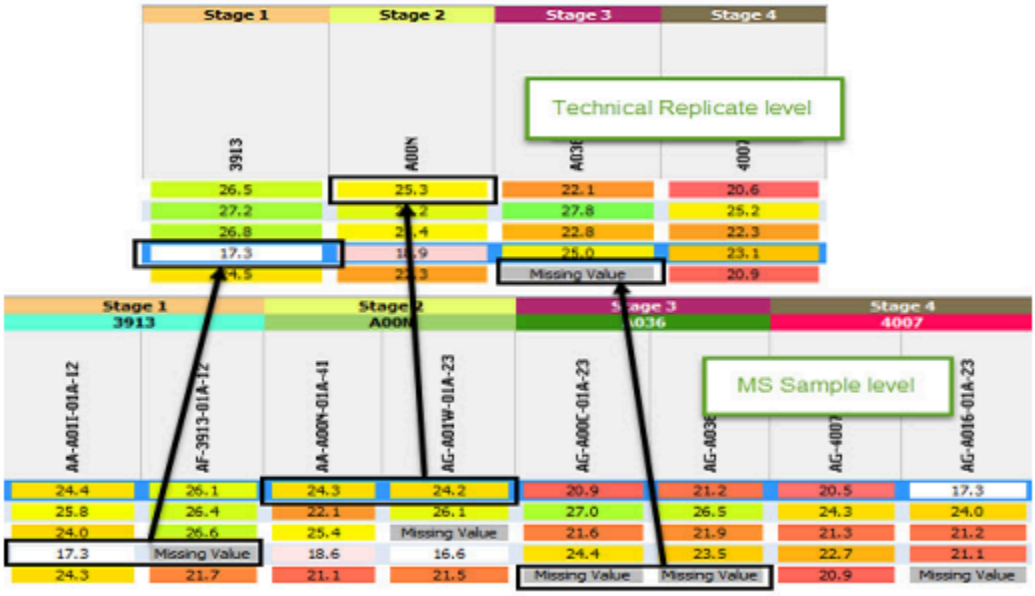
Appendix C: Summarization: Rolling up Values

When the level of summarization is changed, Scaffold DIA rolls up **Log10 Intensity** values to a higher level of summarization, or Attribute Group, in one of two distinct ways, depending on whether the particular level of summarization is at or above the technical replicate (TR) level, see below.

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

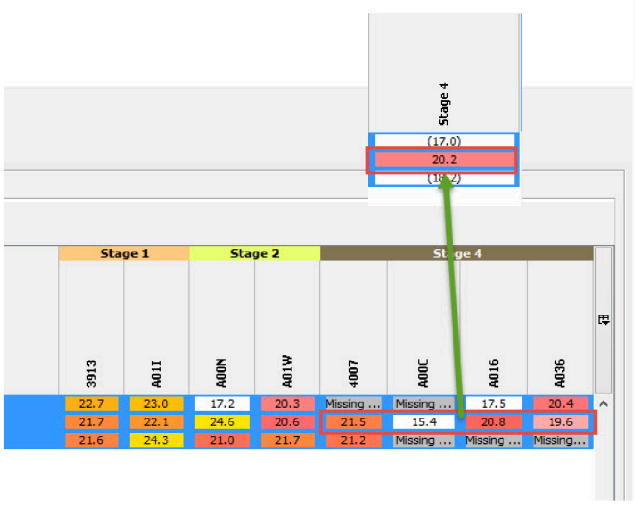
Note: This configuration is recommended when working with MS samples that are fractions of a MUDPIT experiment.



When the summarization level is switched to a level higher than the technical replicate level, Scaffold DIA rolls up the values using the median of the Intensity values of the technical replicates in an attribute

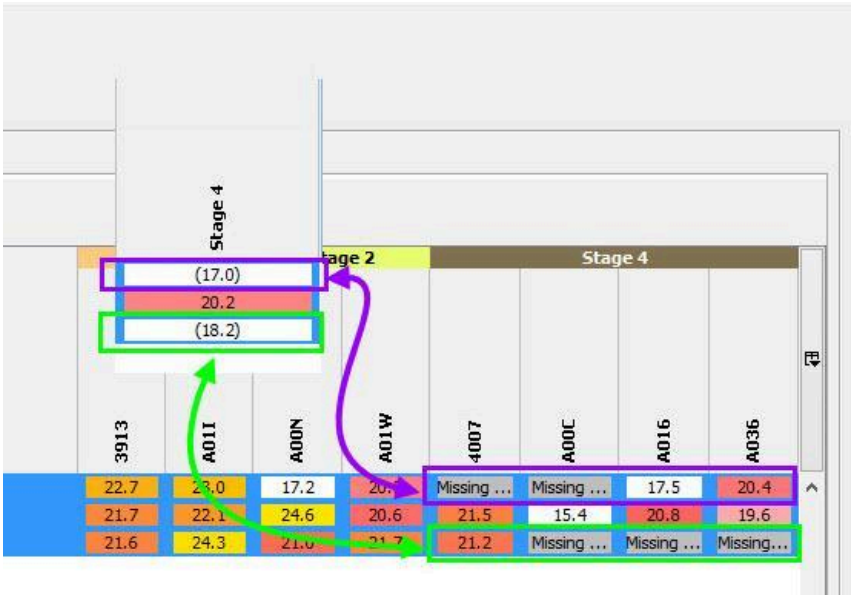
group, see the figure below The picture shows the rolled up value of the $\log_{10}$ intensity values of four biosamples that make up the attribute group Stage 4 and the $\log_{10}$ intensity when the higher summarization level Stage is selected. The value corresponds to the median of the values appearing when the next lower level of summarization is chosen.

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, missing values are 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, as shown below.

Missing values



Appendix D: Missing Values

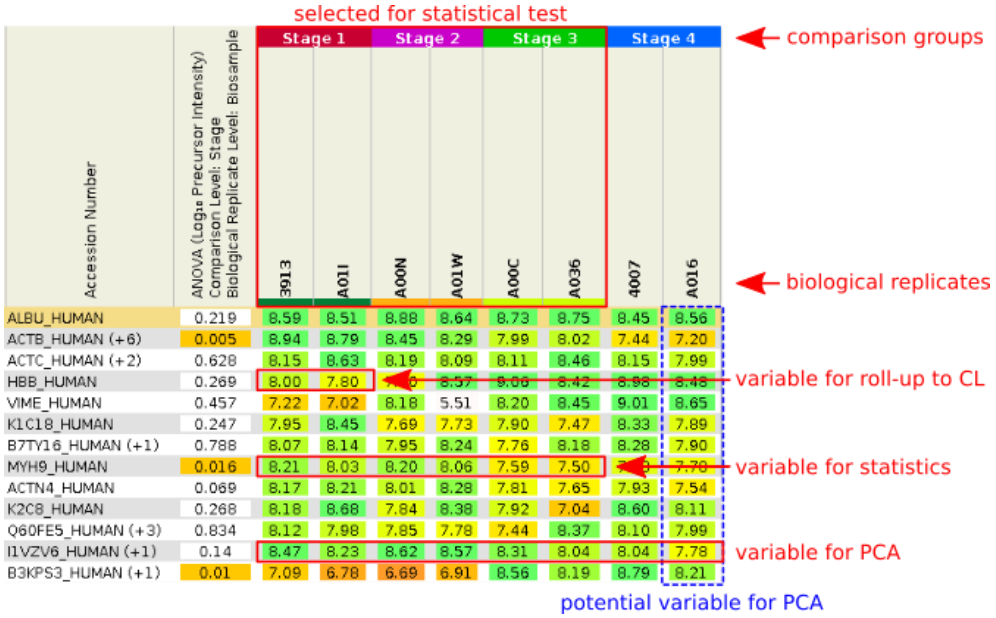
Missing values affect various computations in Scaffold DIA. In order to be able to roll up values to higher levels of summarization, to perform statistical testing and to perform Principle Component Analysis, it is necessary to impute values when no measurement has been obtained. On the assumption that missing values in DIA analysis are generally a result of either absence of the peptide in a sample or presence at very low intensity, Scaffold DIA uses the method of “Quantile Regression for Imputation of Left-Censored data" (QRILC)[14].

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.



[14]Cosmin Lazar (2015): imputeLCMD v2.0". R package version 2.0

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 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 E: Shared Evidence Clustering

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:

$$0 \le L(A,B) \le 1$$

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

$$0 < \lambda \le 1$$

and then consider A and B similar at level λ if

$$L(A,B) \ge \lambda$$

.

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.

Appendix F: Heat map clustering

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 own 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 DIA 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:

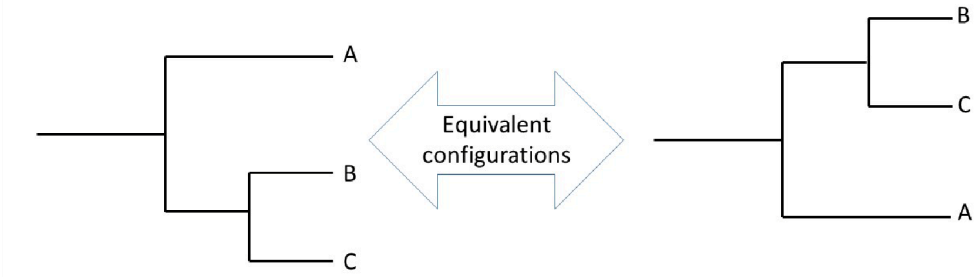
- *A* and *B* are vectors whose coordinates characterize two rows or two columns of the data matrix at a selected Display Type and summarization level, see Summarization Bar; and r_A and r_B are coordinate-wise vectors of A and B. The components of the vector are given by the displayed values shown in 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 DIA 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 display 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, see the figure below.

Binary switching example



The Heat map in Scaffold DIA

Scaffold DIA constructs a heat map using information and data available in the Samples View and displays it in the Visualize View.

Whatever thresholding and filtering are applied to the Samples Table determine the number of rows and columns considered for the data matrix used to create the Heat map shown in The Visualize View. 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 DIA 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 G: Techniques to Control the Family-wise Error Rate

Techniques to Control the Family-wise Error Rate

Currently Scaffold DIA supports control of the family-wise error rate, FWER, using Holm's step-down procedure and Hochberg's step-up procedure.

The way Scaffold DIA 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 H: Using Principal Component Analysis in Scaffold DIA

Using Principal Component Analysis in Scaffold DIA

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 DIA, 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[15] and was obtained from the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org)[16] 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.

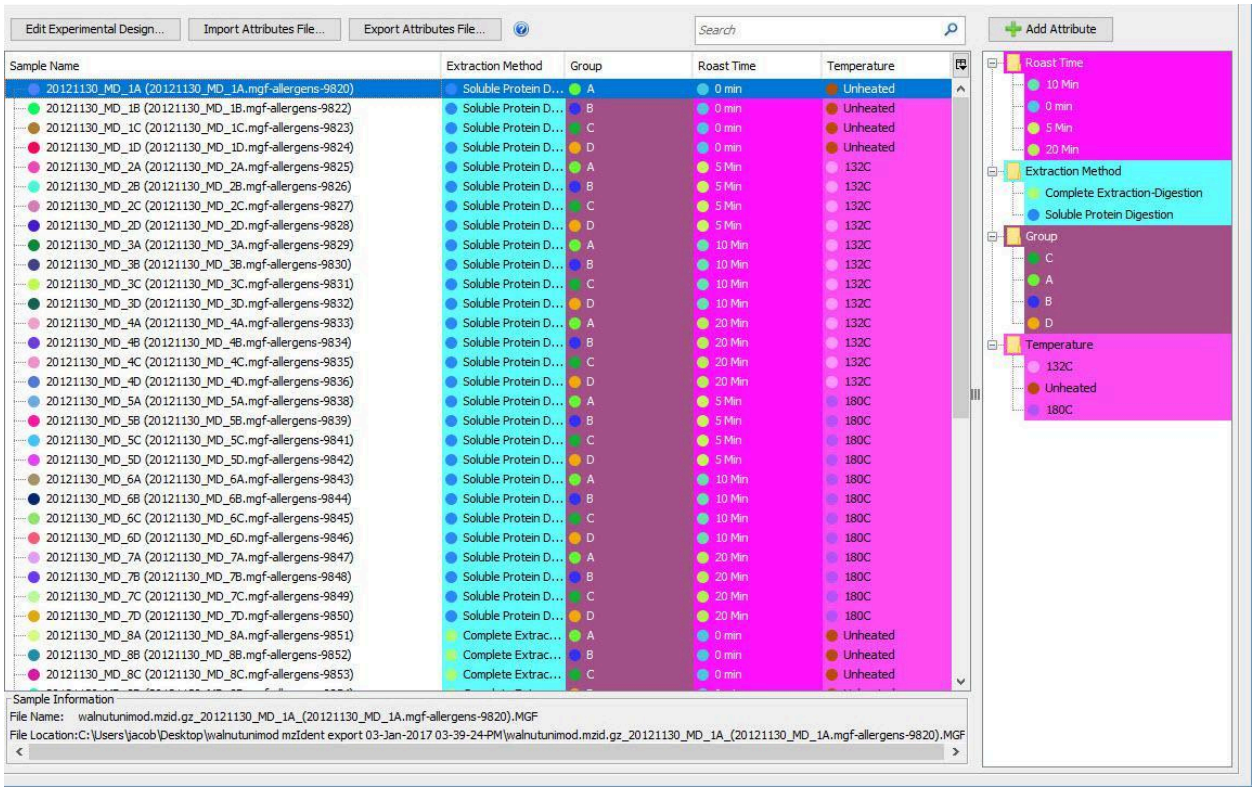
This study used a block design, in which four replicate samples were divided and subsamples of each underwent several treatments. The proteins from each were then extracted in two different ways.

In Scaffold DIA, 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.

[15]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.

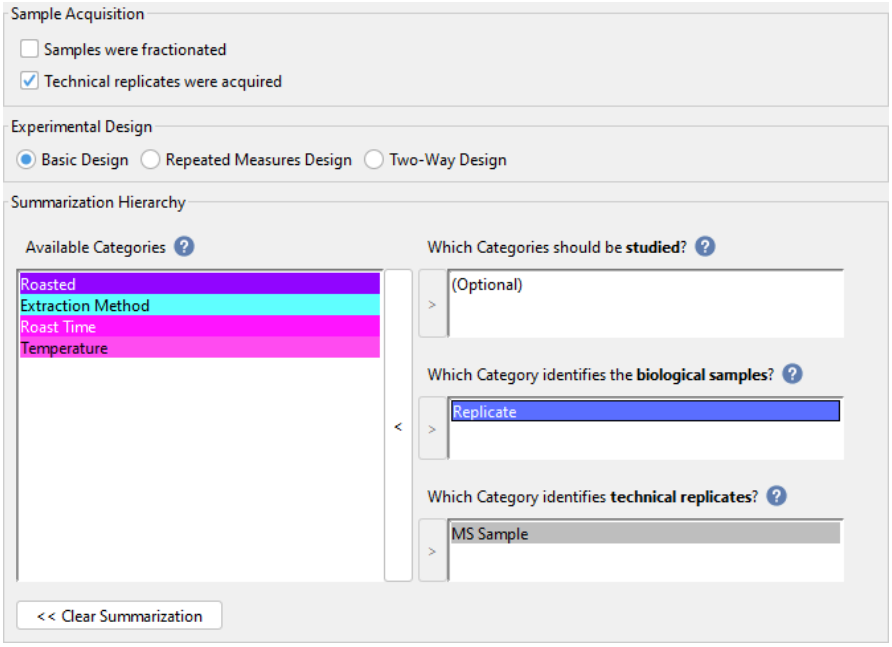
[16]Vizcaíno JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, Xu QW, Wang R, Hermjakob H. 2016 update of the PRIDE database and related tools. Nucleic Acids Res. 2016 Jan 1;44(D1): D447-D456. PubMed PMID:26527722.

Figure 7: Samples with Attributes assigned



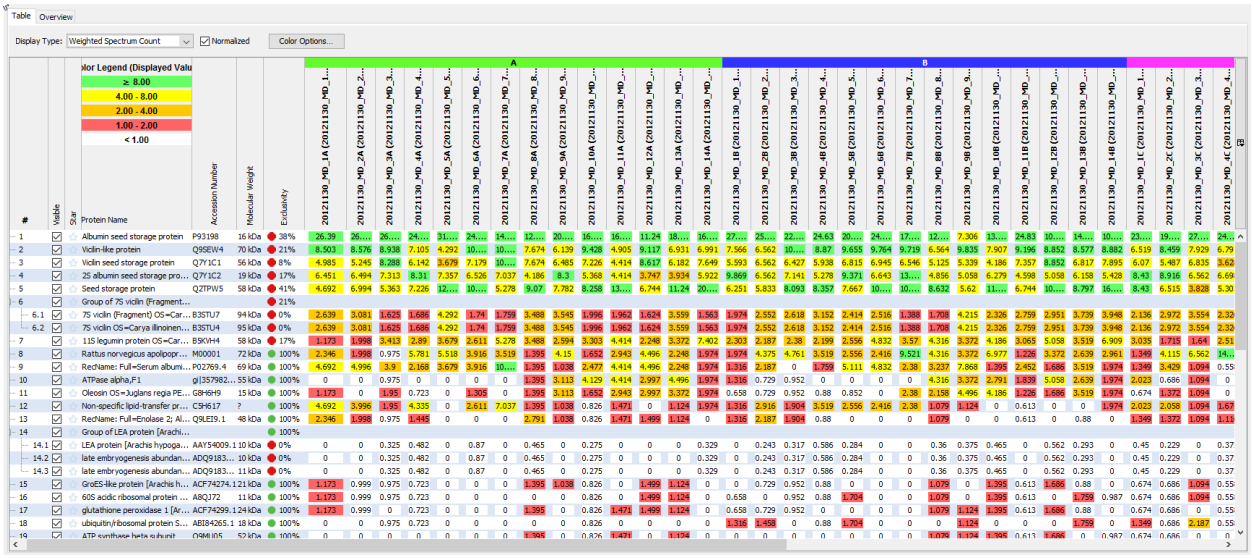
We might initially set the Summarization Hierarchy to specify our technical and biological replicates:

Figure 8: 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 9: 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.

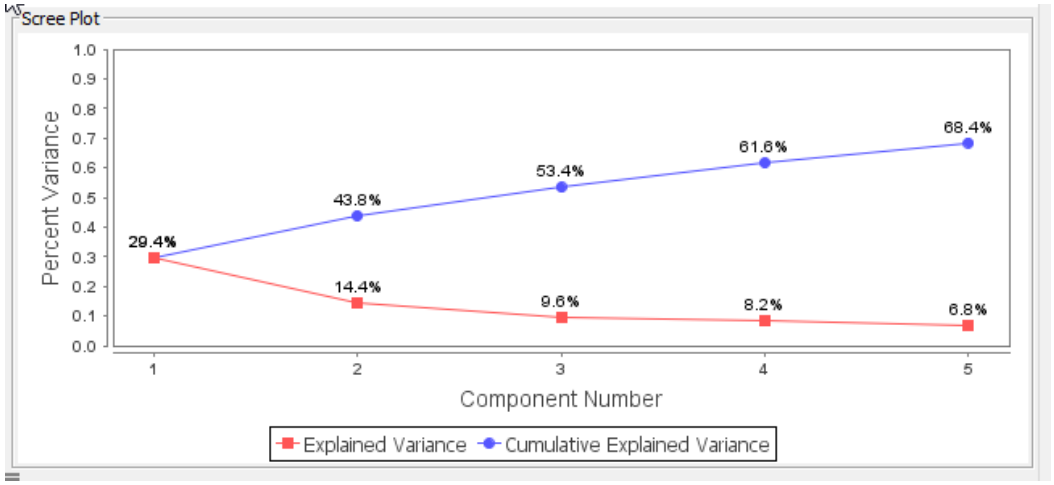
PCA analyzes data in order to find patterns in the expression levels of the various proteins that differentiate samples or groups of samples. It constructs a weighting function that, when applied to the quantitative values of the proteins, results in the greatest separation of the samples, or, put another 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.

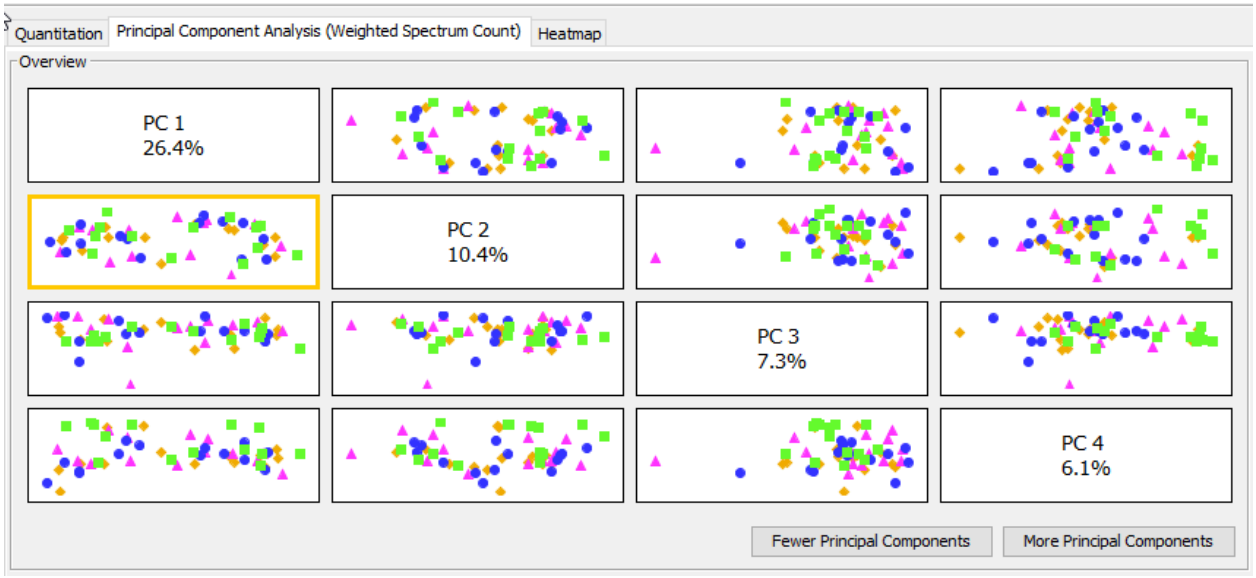
Figure 10: The Scree Plot



Overview

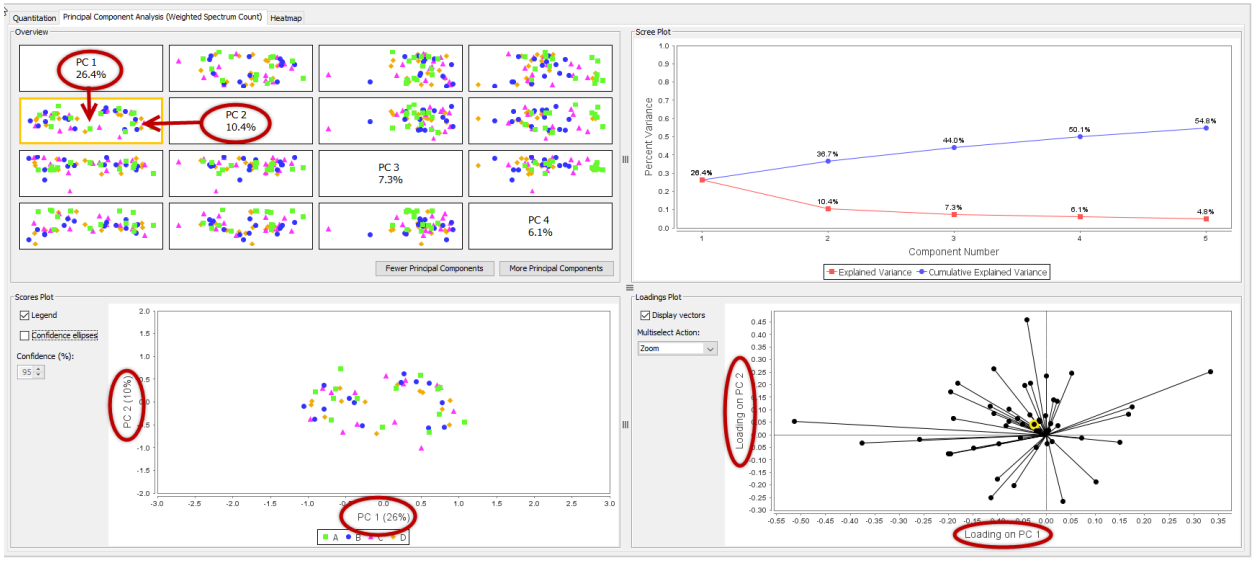
The Overview is a series of graphs where one Principal Component is plotted against another. The points in these graphs represent samples, and the X and Y coordinates are the values computed from the corresponding Principal Component functions. We can see that the samples tend to cluster in different ways depending on the Principal Components applied.

Figure 11: 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.

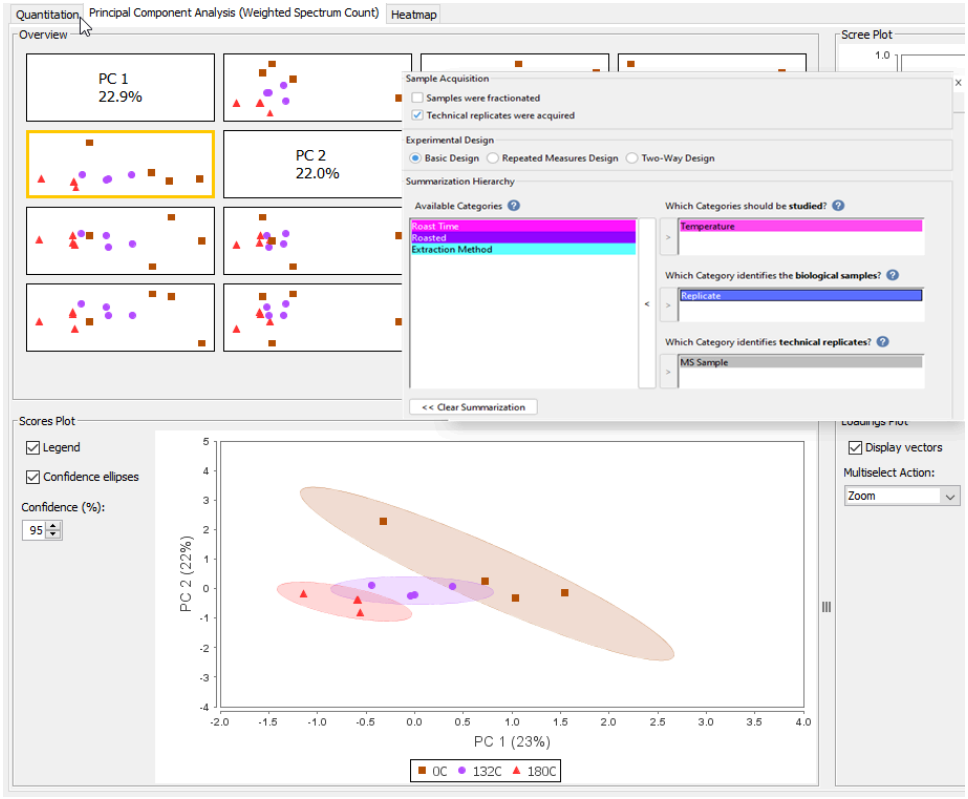
Figure 12: Selecting Principal Components with the Overview



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 13: Exploring the relationship of Temperature and PC1 and PC2

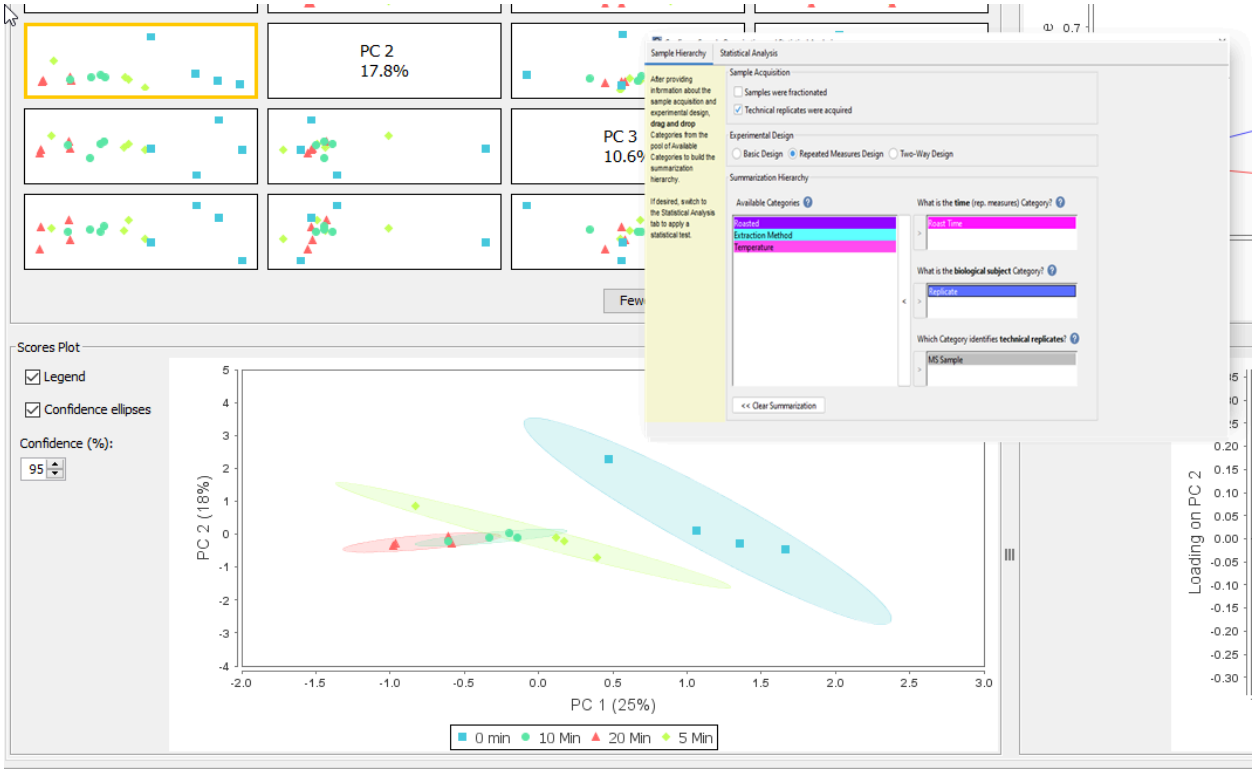


Here we see clustering in the Scores Plot. Roasting Temperature correlates with PC1, as the samples subjected to higher temperatures appear to the left in the chart (lower PC1 values), those roasted at lower temperatures are in the middle, and unroasted samples are on the right (higher PC1 values).

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:

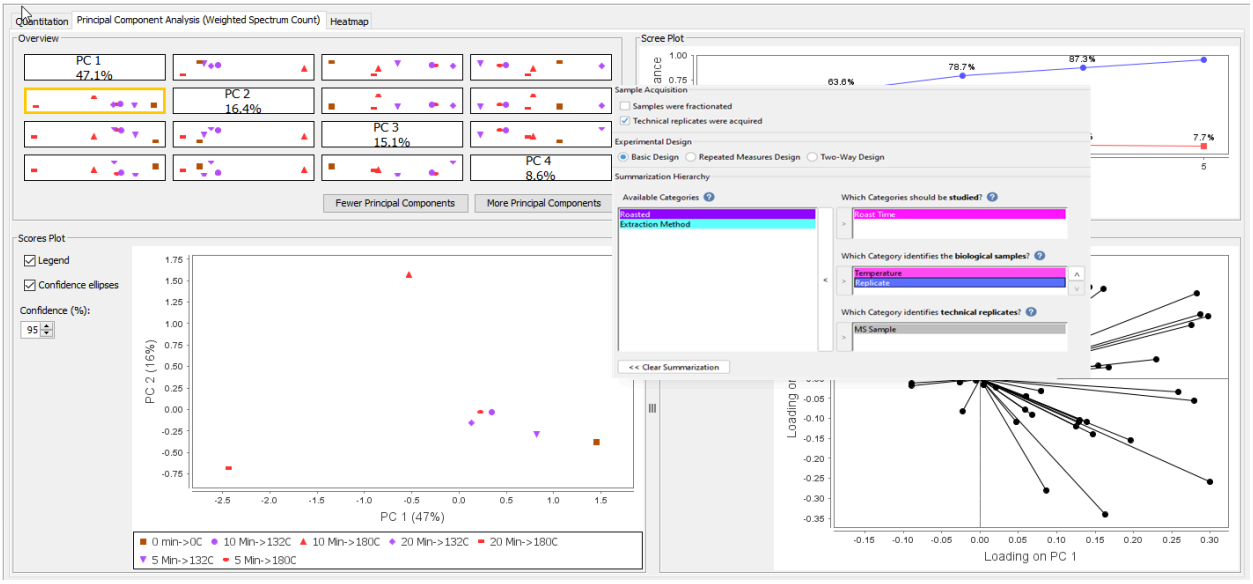
Figure 14: Exploring the Relationship of Roasting Temperature with PC1 and PC2



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 15: 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.

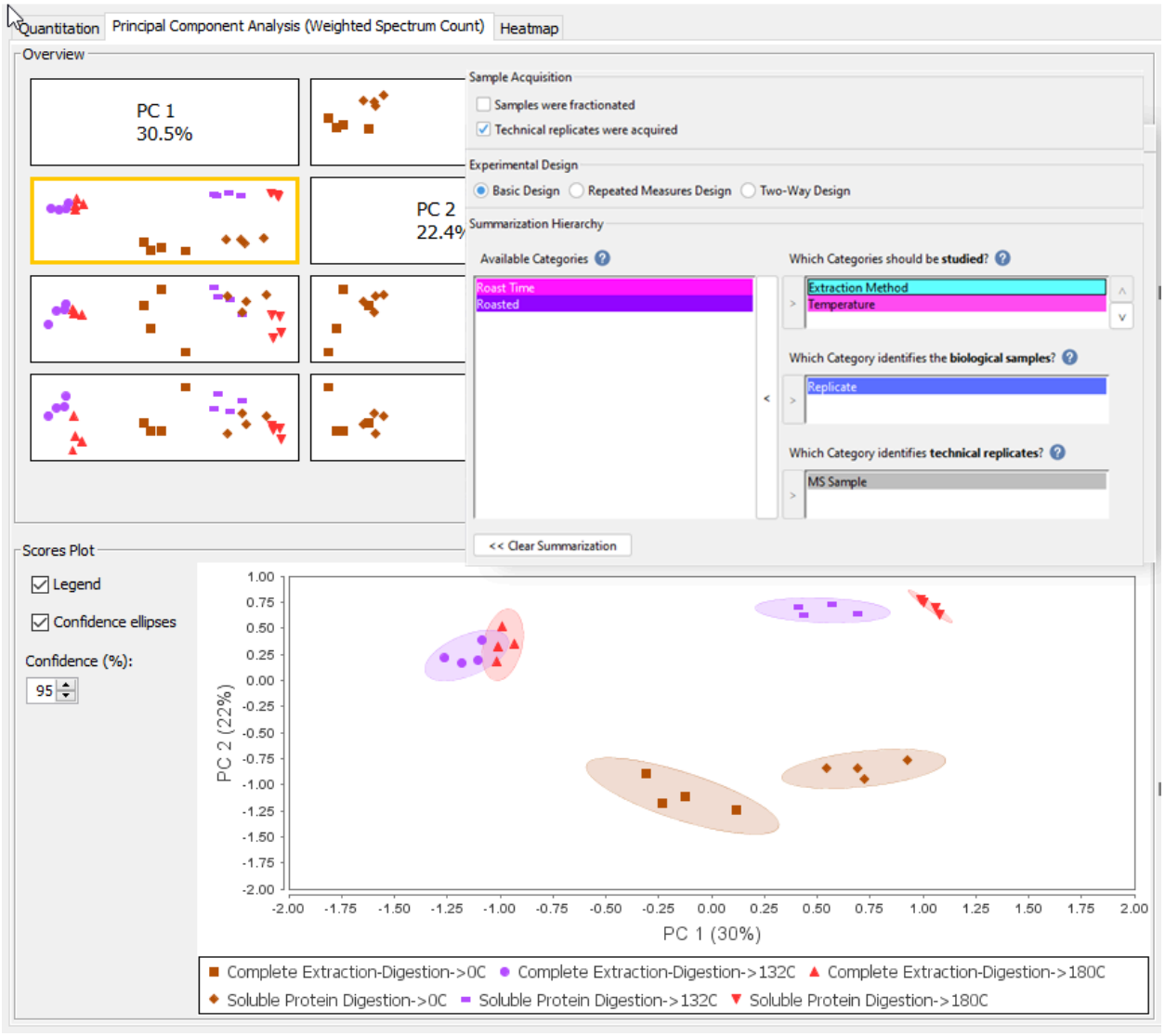
As can be seen below, Extraction Method produces the clearest clustering of all in PC2 vs PC1:

Figure 16: 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 17: 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.

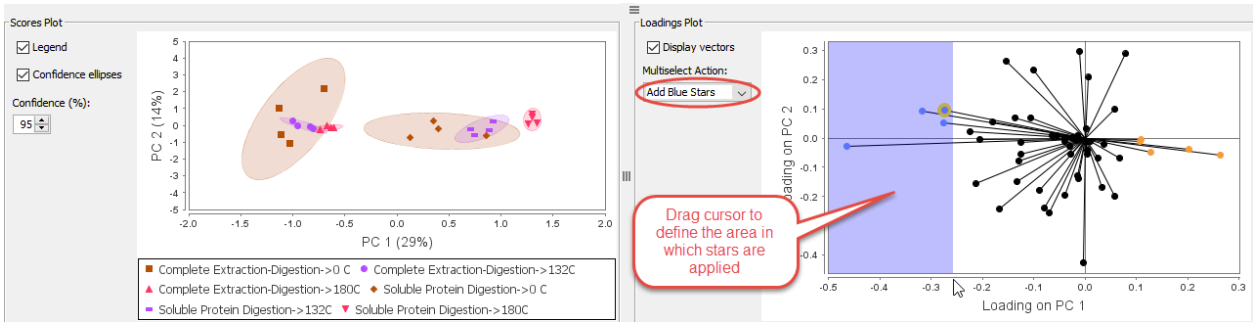
Loadings Plot

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.

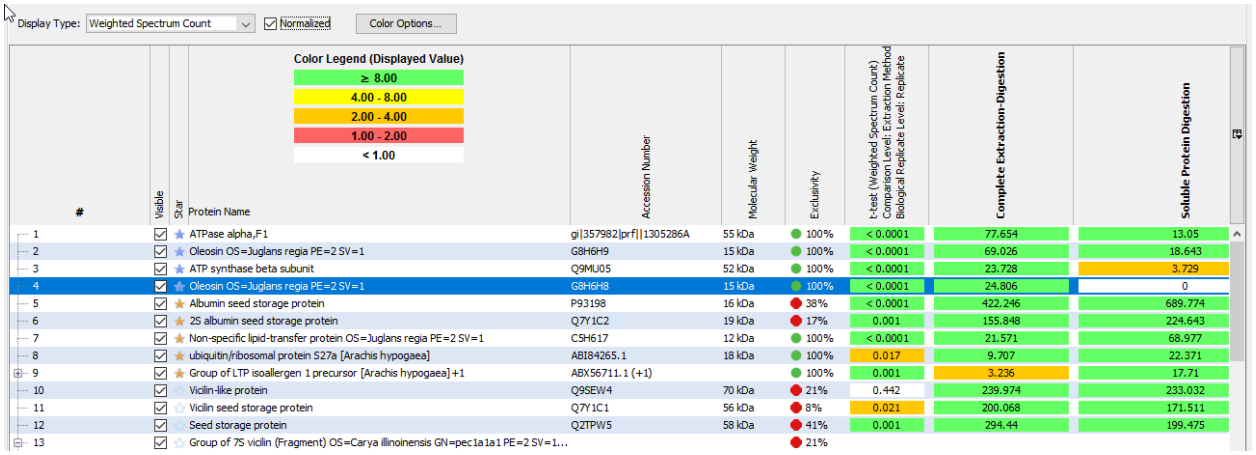
We can mark s through the Scores plot that may prove useful in identifying samples that are particularly effective at differentiating samples based on certain criteria. For example, we place orange stars on the s to the right in the Loadings Plot, and blue stars on proteins to the left:

Figure 18: Starring proteins 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 proteins are significantly higher in the complete extraction.

Figure 19: Viewing the starred proteins in the Samples View



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 I: How PCA is Performed in Scaffold DIA

Principal Component Analysis (PCA) is a classical dimension-reduction technique based on linear algebra. The idea is to find the "underlying processes" that explain the variance in the data. In PCA, these "underlying processes" consist of linear combinations of the original variables.

The principal basis vectors are chosen one at a time in such a way that each vector chosen

- is perpendicular to all the previously chosen principal basis vectors
- is one unit long
- points in the direction that explains the most variation of the data (given the constraints)

Dimension reduction is achieved by projecting the original vectors into the space spanned by some subset of the principal basis vectors.

For details, see Appendix E in the Scaffold DIA User's Guide, available through Help>Scaffold DIA User's Guide.

In Scaffold DIA, 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[17] which is also used

[17] Durbin BP, Hardin JS, Hawkins DM, Rocke DM. A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics. 2002;18(Suppl. 1):S105–S110.

by MetaboAnalyst[18].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,

$$a_{ij} = \mathrm{glog}(I_{ij}),$$

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).

[18]Xia, J., Sinelnikov, I., Han, B., and Wishart, D.S. (2015) MetaboAnalyst 3.0 - making metabolomics more meaningful . Nucl. Acids Res. 43, W251-257.

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[19] 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 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

[19]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.

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 DIA

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 to the attribute to which they correspond in the Comparison Level.

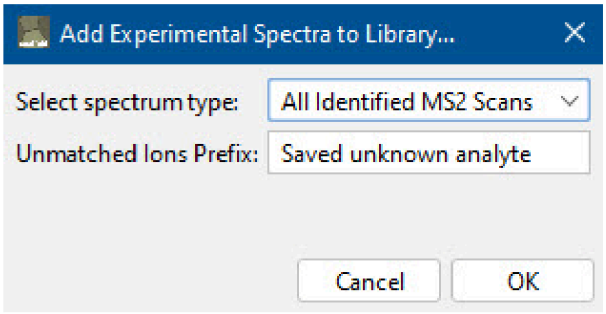
The scores plot also shows the 95%-confidence ellipses for each attribute. (Actually the p%-confidence ellipses where p can be specified by the user.) These ellipses show the region where 95% of the data points will lie assuming their distributions are independent and normally distributed in the dimensions 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 J: Description of Mouse Right Click Context Menu Commands

- **Select All** - Select all rows in the tree table.
- **Set as internal standard** - Set the selected analytes as the standard set for normalization.
- **Add to library...** - Opens a dialog to allow the user to add the spectra for the selected analytes to the current library.

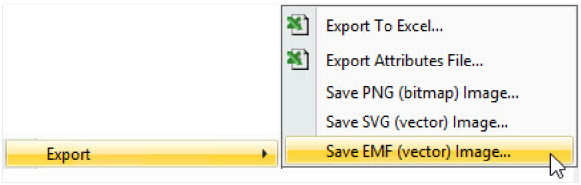


- **Stars** - Provides a mechanism to allow flagging of selected samples with colored stars
  - Add Orange - Adds an orange star to selected row(s). If a sample already has a blue star, it becomes a blue and orange star.
  - Add Blue - Adds a blue star to selected row(s). If a sample already has an orange 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 options for copying data from a table

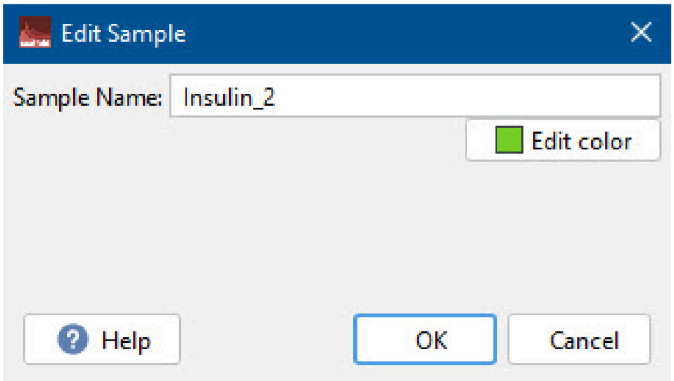
  - *Copy Image* - Copies the image of the current table.
  - *Copy Selected Cell* - Copies data contained in the selected cell of the current table.
  - *Copy Selected Row* - Copies data contained in the selected row of the current table.
  - *Copy All Data* - Copies data contained in the current table.
- **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 an 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

Agilent, 30
Apply GO terms, 59
Associate a FASTA to a library, 32
Assumptions for the manual, 4
Automatically add selected files to Library Manager, 33
Available Categories, 112
Basic Design, 108
BLIB, 38
Block Effect, 122
Bulk Edit Sample Names, 104
Choosing a search type option, 48
Clearing the peptide filter(s), 125
Cluster Columns for Heatmap, 152
Cluster Value Suppression, 58
Colored Bars in Sample Column Headers, 114
combined FASTA, 33
Comparison Groups in Statistical Tests, 114
Conventions used in the manual, 3
Convert Raw Files to DIA Files, 53
Copyright, 2
d directory, 30
Data value tags
  Missing Ref., 92
  Missing Values, 92
  No Values, 92
Digestion Enzyme, 43
Display Pane, 74
  column sorting feature, 74
  moving columns around, 74
  multi selection of rows, 74
  resizing of columns and panes, 74
  tables column control, 74, 75
Display pane in the Samples View
  display options, 91
Displayed GO terms, 61
DLIB, 21, 25
Enzyme, 43
Exclusive Peptide Quantification, 84
Export Attributes File..., 176
Export MzTab Report for PRIDE, 177
Export Peptide Match Report to Excel..., 177
Export Workflow, 177
FDR
  dialog box;, 89
FDR Info Box, 72
FDR Level q*, 121
Filter Retention Time Range, 46
Filtering Control Bar, 67
Filtering Pane
  advanced filter icon; Advanced Filter icon;, 68
Fractions, 111
Fragment Consistency Score, 126
Fragment intensity quantitation
  Calculations, 162
GO Annotations, 59
Global Probability, 127
Global q-value, 127
Heatmap, 58
How to Choose the Appropriate Data
  Acquisition Mode, 49
Identified Peptide Count, 84
Installing Scaffold Variants, 7
Interaction Effect, 122
Key Types, 9
Labeled Quant Key, 9
Library Manager - Automatically add selected
  files, 57
Licensing the Program, 9
Loading data into Scaffold DIA, 19, 29
Manual registration, 12
Memory Usage, 57
Menu commands, 53
  export, 55
  files, 53
  redo and undo commands, 56
  view, 54
Modifications, 39
Moving the Program to a different computer,
  15
MSP, 21, 25, 38
Multiple Test Correction, 121
mzTAB, 177
Navigation Pane, 71

Non-overlapping Windows, 49
Nonparametric tests, 167
Offline activation, 12
Open Library Manager, 53
Organize Samples window
  using;, 98
Overlapping Margins, 49
Parametric tests, 167
PCA, 202
Percolator Settings, 46
pi0, 72
Precursor intensity quantitation
  Mascot Distiller, 163
  MaxQuant, 163
  Preparing data for, 163
  Proteome Discoverer, 163
  Spectrum Mill, 163
Preferences Dialog box, 57
Preferences Dialog Box
  Analysis Settings, 57
  Colors, 57
  General Settings, 57
  Internet, 57
  Processors Settings, 57
PRIDE submission, 177
Primary Factor Effect, 122
Prosit Library, 21
Protein List
  Confidence Thresholds, 89
  how to apply filters
    Applying filters to the Protein List, 90
protein quantification, 17
Publish Report, 157
Quantified Checkbox, 126
Quantified Peptide Count, 92
Quantitative CVs Chart, 145
quantitative sample
  see also Organize Samples
    window[quantitative sample
    aaa], 98
quantitative sample category
  see also Organize Samples window
    [quantitative sample category
    aaa], 98
Randomized Block, 110
RAW, 30
Reference Samples, 170
Release Information, 2
Remove Library, 32
Reorder Libraries in the Library Manager table, 33
RT Filtering, 46
Sample Hierarchy Tab, 164
Sample Probability, 128
Sample q-value, 128
Sample q-value thresholds, 84
Samples Table
  Color Legend, 87
  Initial Sorting of Columns, 83
Samples View
  protein list, 87
  hiding proteins from
    hiding in the Samples View, 88
  proteins of interest
    identifying in the Samples View starring
      proteins, 88
  sorting feature
    Samples View, 83
Secondary Factor Effect, 122
Selecting a library from the Library Manager, 33
Selecting a Pending Library, 33
Selecting Multiple Libraries, 33
Show All Peptides, 129
Special information about the manual, 3
SPTXT, 21, 25, 38
Staggered Windows, 50
Statistical Analysis Tab, 164
Statistics Table, 133
Summarization Pane, 73
Supplementary Data, 157
System Requirements, 7
Table Tab Display pane
  column crdering selection menu, 91
  display options button, 91
Technical Replicates, 111
The Library Manager, 30
The Multi-Select Action, 140
Thermo, 30
TMTpro, 164
Tool-bar, 52
Total Unique Peptide Count, 85
TraML, 21, 25, 38
Treatment Effect, 122
Two-Way ANOVA, 122
Use all reference samples, 170
Use matched reference samples, 171

User Interface, 57
Using the manual, 3
Violin Plot, 132
Warning Icon, 34
WIFF, 29
WIFF.SCAN, 29
yellow triangle, 34