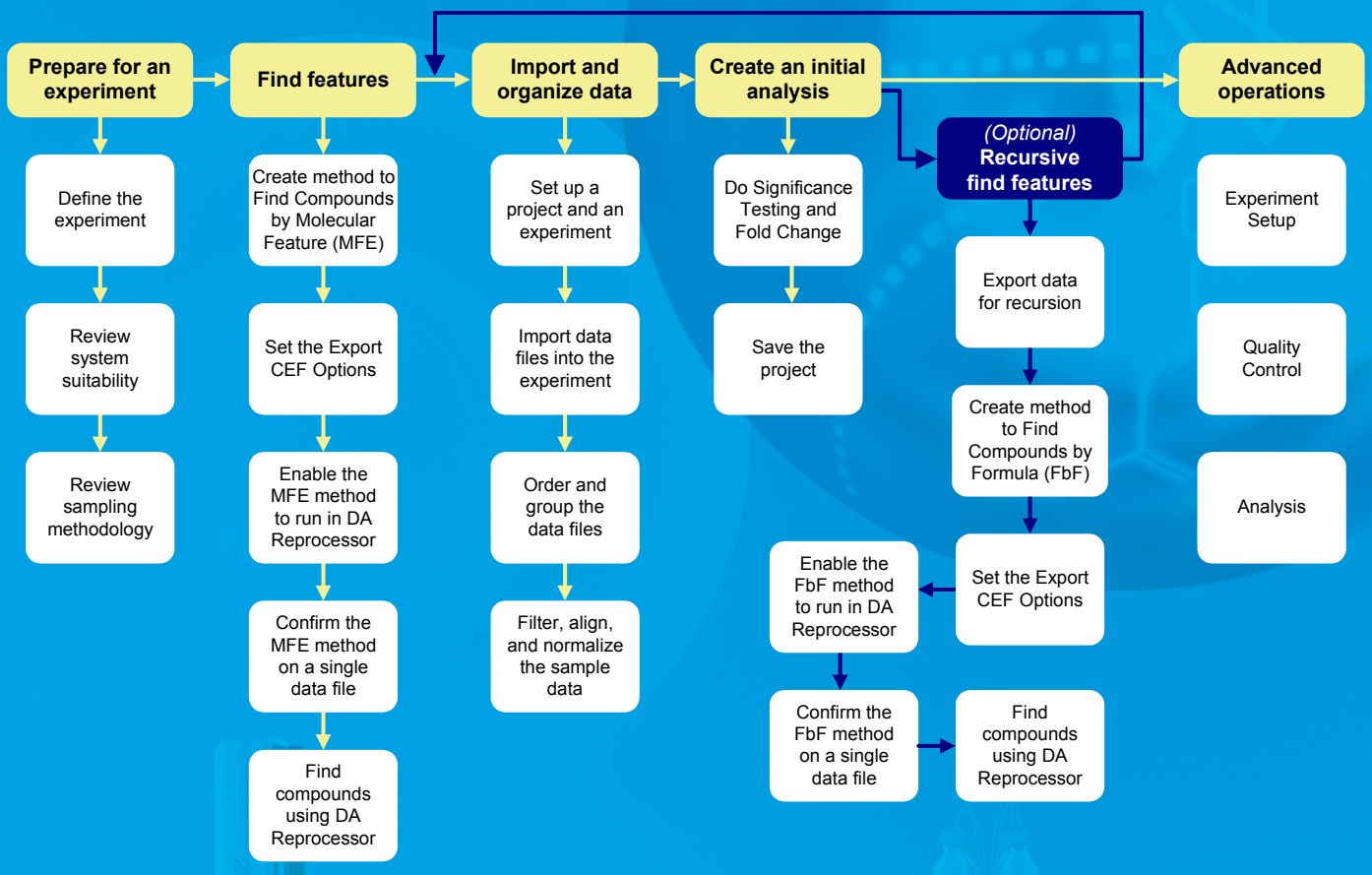
Agilent Metabolomics Workflow

Discovery Workflow Guide



Agilent Technologies

Notices

Manual Part Number

5990-7067EN

Edition

Revision B, October 2012

Printed in USA

Agilent Technologies, Inc.
5301 Stevens Creek Blvd.
Santa Clara, CA 95051

Acknowledgements

Warranty

Technology Licenses

Restricted Rights

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.

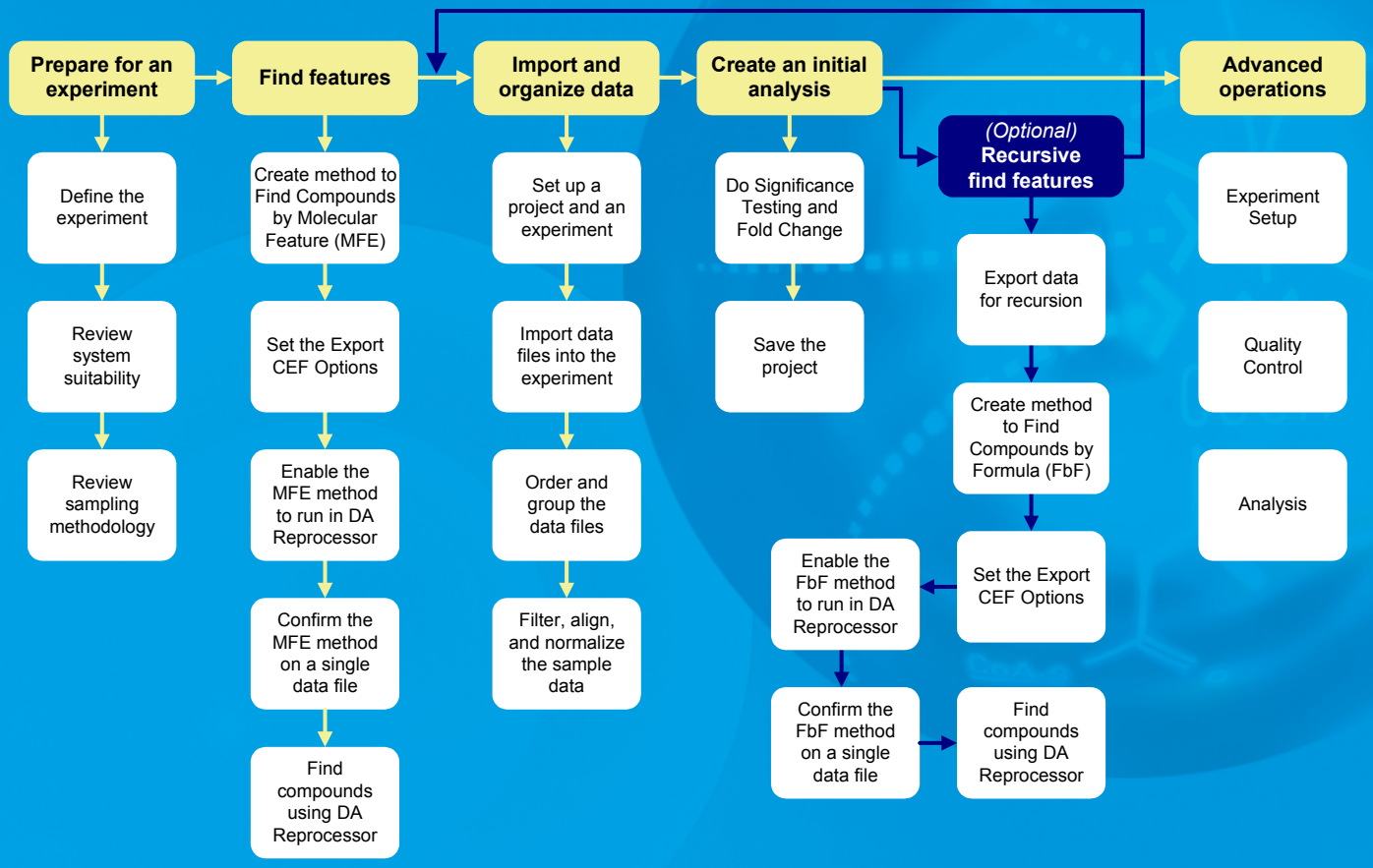
Contents

What's new in Revision B

- The Mass Profiler Professional wizard and workflow images are based on version 12.01 or later.
- Formatting of text that appears in the left-hand margin is improved for guiding you through the main processes and operations.
- The Metabolomics Experiment chapter is improved for readability and terminology.
- Each step of the five-step workflow includes an overview of the procedures performed in the step and a brief guide to the next step.
- The Advanced Analysis section guides you through the operations under Experiment Setup, Quality Control, and Analysis.
- Statistical Analysis operations include flow charts that show you how the wizards are navigated based on your experiment and selections.

Before You Begin

Make sure you read and understand the information in this chapter and have the necessary computer equipment, software, experiment design, and data before you start your analysis.



Agilent Technologies

Introduction

Metabolomics is an emerging field of 'omics' research that is concerned with the characterization and identification of metabolites, the end products of cellular metabolism. Metabolomics research leads to complex data sets involving hundreds to thousands of metabolites. Comprehensive analysis of metabolomics data uses a strategy that is often unique and requires specialized data analysis software that enables cheminformatics analysis, bioinformatics, and statistics. Agilent MassHunter Qualitative Analysis and Agilent Mass Profiler Professional together enable metabolomics data analysis.

This Discovery Workflow Guide is used to identify differences in your samples based on their metabolites. The workflow is applied after you collect untargeted GC/MS and/or LC/MS data. The first step in the metabolomics workflow is to "find" the signals in your data based on retention time, mass, and abundance. These signals are called features. Feature finding is accomplished using Agilent MassHunter Qualitative Analysis. Agilent Mass Profiler Professional identifies the most significant features, performs statistical analyses, and provides interpretation through differentiation based on relative metabolomic profiles and your sample grouping.

Metabolomic studies involve the process of identification and quantification of the endogenous components that form a chemical fingerprint of an organism, or situation under study, and may involve the process of identifying correlations related to changes in the fingerprint as affected by external parameters (metabonomics). Mass Profiler Professional may be used in the study of metabolomics and metabonomics for small molecule studies, proteomics for protein biomarker studies, and general differential analysis. Regardless of the specific study and molecular class, the process is referred to as "metabolomics" throughout this workflow.

To increase your confidence in obtaining reliable and statistically significant results, the metabolomics analysis must include a carefully thought-out experimental design that includes the collection of replicate samples. Replicate data collection and proper experimental definition establish the quality and significance of the Mass Profiler Professional analysis.

More information

The metabolomics discovery workflow is part of the collection of Agilent manuals, help, application notes, and training videos. The current collection of manuals and help are valuable to users who understand the metabolomics workflow and who may require familiarization with the Agilent software tools. Training videos provide step-by-step instructions for using the software tools to reduce example GC/MS and LC/MS data but require a significant time investment and ability to extrapolate the example processes. This workflow provides a step-by-step overview of performing metabolomics data analysis using Agilent MassHunter Qualitative Analysis and Agilent Mass Profiler Professional.

The following selection of publications provides materials related to metabolomics and Agilent MassHunter Mass Profiler Professional software:

- *Manual:* Agilent Metabolomics Discovery - Discovery Workflow Overview (5990-7069EN, Revision B, October 2012)
- *Manual:* Agilent G3835AA MassHunter Mass Profiler Professional - Quick Start Guide (G3835-90009, Revision A, November 2012)

- *Manual:* Agilent G3835AA MassHunter Mass Profiler Professional - Familiarization Guide (G3835-90010, Revision A, November 2012)
- *Manual:* Agilent G3835AA MassHunter Mass Profiler Professional - Application Guide (G3835-90011, Revision A, November 2012)
- *Brochure:* Agilent Solutions for Metabolomics (5990-6048EN, April 30, 2012)
- *Brochure:* Agilent Mass Profiler Professional Software (5990-4164EN, April 27, 2012)
- *Application:* Mass Profiler Professional and Personal Compound Database and Library Software Facilitate Compound Identification for Profiling of the Yeast Metabolome (5990-9858EN, April 25, 2012)
- *Application:* Multi-omic Analysis with Agilent's GeneSpring 11.5 Analysis Platform (5990-7505EN, March 25, 2011)
- *Presentation:* Multi-omic Analysis Software for Targeted Identification of Key Biological Pathways (USHUPO_IB_March_2012.pdf, March 2012)
- *Application:* An LC/MS Metabolomics Discovery Workflow for Malaria-Infected Red Blood Cells Using Mass Profiler Professional Software and LC-Triple Quadrupole MRM Confirmation (5990-6790EN, November 19, 2010)
- *Brochure:* Integrated Biology from Agilent: The Future is Emerging (5990-6047EN, September 1, 2010)
- *Primer:* Metabolomics: Approaches Using Mass Spectrometry (5990-4314EN, October 27, 2009)

A complete list of references may be found in "References" on page 190.

NOTE

This manual gives links to most references. If you have an electronic copy of this manual, you can easily download the documents from the Agilent literature library. Look for and click the blue hypertext; for example, you can click the "Agilent literature library" link in the previous sentence.

If you have a printed copy, go to the Agilent literature library at www.agilent.com/chem/library and type the publication number in the **Keywords** or **Part Number** box. Then click **Search**. (Note: If you type the publication number into the **Keywords** box, you find the publication number *and* additional publications that reference the publication number.)

"Definitions" on page 180 contains a list of terms and their definitions as used in this workflow.

Overview of the workflow

The metabolomics workflow describes the basics of metabolomics data analysis using Agilent MassHunter Qualitative Analysis and Agilent Mass Profiler Professional to formulate statistically significant answers to simple questions presented to complex data sets. Mass Profiler Professional is designed to let you take full advantage of the unique analytical capabilities of any Agilent mass spectrometer.

The specific goals of the metabolomics workflow are:

1. Present general definitions of metabolomics and the metabolomics workflow to highlight the analytical applicability of MassHunter Qualitative Analysis and Mass Profiler Professional to a wide variety of samples: such as metabolomics, proteomics, food safety, environmental, forensics, toxicology, petrochemical, and biofuels.
2. Introduce the concept of formulating a hypothesis, a question that proposes a possible correlation observed in the data, that differentiation identified within the data answers.
3. Relate the hypothesis to the sample collection and preparation so that a reasonable expectation is established for the outcome.
4. Present an effective step-by-step process using MassHunter Qualitative Analysis and Mass Profiler Professional to serve as a guide for you to perform metabolomics discovery using any data set.
5. Present an effective step-by-step process for employing targeted analyses using quantitative data imported into Mass Profiler Professional.

This workflow describes how to use the following Agilent software programs together: MassHunter Qualitative Analysis, MassHunter DA Reprocessor, Mass Profiler Professional, and ID Browser.

You can use this workflow as a road map for any analysis that requires statistically significant answers to simple questions presented to complex data sets.

Advantages of this workflow

This workflow automates many parts of metabolomics data analysis and provides powerful visual tools for statistically answering the question proposed to the experiment. The results may be used to create models for automated class prediction.

What you cannot do with the workflow

Statistically meaningful results require that the experiment be performed with a clear understanding of the variables and that the data is collected with sufficient replicates.

Safety notes

This workflow does not involve the collection of data or operation of any instrument.

Required items

The Metabolomics Workflow performs best when using the hardware and software described in the "required" sections below. The required hardware and software is used to perform the data analysis tasks shown in Figure 1.

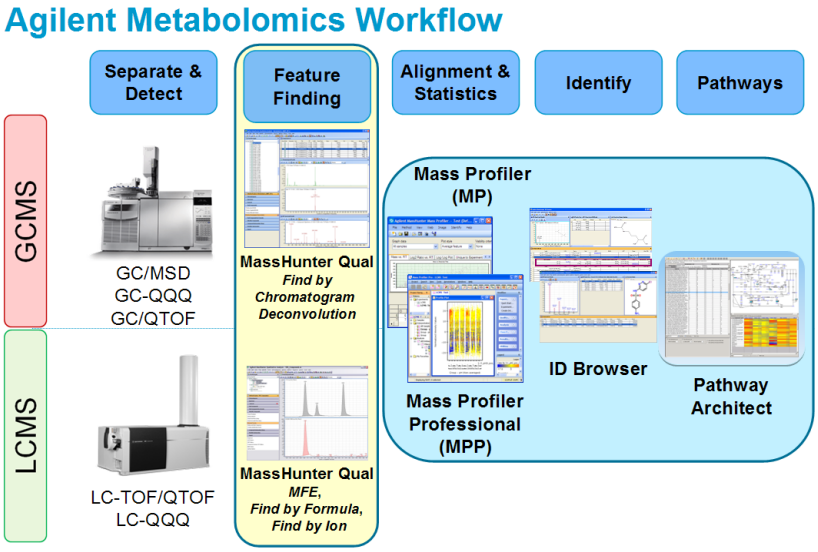


Figure 1 *Agilent hardware and software used in performing metabolomics.*

Required hardware

- PC running Windows
  - *Minimum:* XP SP3 (32-bit) or Windows 7 (32-bit or 64-bit) with 4 GB of RAM
  - *Recommended:* Windows 7 (64-bit) with 8 GB or more of RAM
- At least 50 GB of free space on the C Partition of the hard drive
- Data from an Agilent GC/MS, LC/MS, CE/MS and/or ICP-MS system or data that may be imported from another instrument.

Required software

- Agilent Mass Profiler Professional Software B.12.00 or later

  Agilent Mass Profiler Professional software is a chemometrics software package designed to exploit the high information content of mass spectrometry data. Researchers can easily import, analyze and visualize GC/MS, LC/MS, CE/MS and ICP-MS data from large sample sets and complex MS data sets.

  Mass Profiler Professional integrates smoothly with Agilent MassHunter Workstation and ChemStation software, and is ideal for any MS-based application where you need to determine relationships among sample group and variables, including metabolomics, proteomics, food safety, environmental, forensics and toxicology.

  For metabolomics and proteomics studies, the optional Agilent Pathway Architect software helps you evaluate MS data in biological context.

- Agilent MassHunter Qualitative Analysis software, Version B.03.01, B.04.00, B.05.00 SP1 or later

  The Agilent MassHunter Qualitative Analysis software lets you automatically find and extract all spectral and chromatographic information from a sample, even when the components are not fully resolved. Powerful data navigation capabilities permit you to browse through compound-specific information in a single sample and compare chromatograms and spectra among multiple samples. The software also includes a customizable user interface and the capability to save, export or copy results into other applications.

- Agilent MassHunter Data Acquisition software, Version B.03.02, B.04.00, B.05.00 or later

  The DA Reprocessor program is a utility that is shipped with the Agilent MassHunter Data Acquisition software. It is included on the Data Acquisition Utilities disk. See the *Data Acquisition Installation Guide* for information on installing this program. The version B.0X of the MassHunter Data Acquisition software must match the version of MassHunter Qualitative Analysis software (for example, B.03.02 MassHunter Data Acquisition must be used with B.03.01 MassHunter Qualitative Analysis and B.04.00 MassHunter Data Acquisition must be used with B.04.00 MassHunter Qualitative Analysis).

- Agilent MassHunter Quantitative Analysis software, Version B.03.02 or later

  The MassHunter Quantitative Analysis software supports simple and efficient review of large multi-compound quantitation batches. A graphical "Batch-at-a-Glance" interface lets you navigate results by compound or sample, or switch between the two approaches. A sophisticated quantitation engine lets you set up over 20 different outlier criteria, and a parameter-less integrator facilitates reliable unsupervised quantitation. The ability to filter results and focus on outliers or questionable peak integrations significantly reduces the data review time for large multi-compound batches. A method task editor and "Curve-Fit Assistant" provide for simple method and multi-level calibration setup.

Optional software

- Agilent ChemStation software

  Agilent ChemStation handles a wide variety of separation techniques such as GC, LC, LC/MS, CE and CE/MS. It is a scalable data system ideally suited for applications in all industries ranging from early product development to quality control. Extensive customization capabilities as well as configurable regulatory compliance provide the flexibility to support different workflows. Sophisticated level-5 control and monitoring of LAN-based instruments ensures fast and flexible data acquisition, which is complemented by advanced data analysis and reporting capabilities for highest productivity.

- AMDIS

  AMDIS is an acronym for the automated mass spectral deconvolution and identification system developed by NIST. (http://www.amdis.net) AMDIS helps analyze GC-MS data of complex mixtures, even data with strong background ions and coeluting peaks. AMDIS is not for use with data collected in SIM mode. AMDIS automatically extracts pure (background free) component mass spectra from highly complex GC-MS data files and uses these purified spectra when searching a mass spectral library.

- MassHunter ID Browser B.03.01 or later

  ID Browser, which is built into Mass Profiler Professional and Qualitative Analysis, performs compound identification using:
  - LC/MS Personal Compound Database (METLIN, pesticides, forensics)
  - GC/MS libraries (NIST and Fiehn library)
  - Empirical Formula Calculation using Agilent's Molecular Formula Generator (MFG) algorithm

  Compounds may be quickly and easily identified within the Mass Profiler Professional environment. ID Browser automatically annotates the entity list and puts the compound names onto any of the various visualization and pathway analysis tools.

- METLIN Personal Compound Database and Library

  METLIN personal compound database and library (PCDL) contains over 25,000 compounds, including 8,000 lipids with retention times for about 700 standards. Used with TOF and Q-TOF data, identification is enabled using accurate mass and/or retention time database searching. Searching the MS/MS spectral library with more than 2,200 compounds enables more confident identification. PCDL represents a data management system designed to assist in a broad array of metabolite research and metabolite identification by providing public access to its repository of current and comprehensive mass spectral metabolite data. (http://metlin.scripps.edu/)

- Agilent Fiehn GC/MS Metabolomics Library

  The Agilent Fiehn GC/MS Metabolomics RTL Library, developed in cooperation with Dr. Oliver Fiehn, is a growing metabolomics-specific library that contains searchable EI spectra and retention-time indexes for approximately 700 common metabolites. The Fiehn library integrates with Agilent's other software tools for GC/MS metabolomics to deliver metabolite identities faster and expand knowledge of metabolomic samples.

Compliance

21 CFR Part 11 is a result of the efforts of the US Food and Drug Administration (FDA) and members of the pharmaceutical industry to establish a uniform and enforceable standard by which the FDA considers electronic records equivalent to paper records and electronic signatures equivalent to traditional handwritten signatures. For more information, see

http://www.fda.gov/RegulatoryInformation/Guidances/ucm125067.htm

MassHunter Data Acquisition Compliance Software includes the following features which support 21 CFR Part 11 compliance:

- Hash Signature for data files let you check the integrity of files during a compliance audit
- Roles that restrict actions to certain users
- Method Audit Trail Viewer

MassHunter Quantitative Analysis Compliance Software includes the following features which support 21 CFR Part 11 compliance:

- Security measures ensuring the integrity of acquired data, analysis, and report results
- Comprehensive audit-trail features for quantitative analysis, using a flexible and configurable audit-trail map
- Customizable user roles and groups let an administrator individualize user access to processing tasks

Before you begin creating methods and submitting studies, you may decide to install MassHunter Data Acquisition Compliance Software and MassHunter Quantitative Analysis Compliance Software.

The Quantitative Analysis Compliance program is installed separately from the Quantitative Analysis program. See *Agilent MassHunter Quantitative Analysis Compliance Software Quick Start Guide* (Agilent publication G3335-90099, Revision A, February 2011) for instructions on installing the Compliance program.

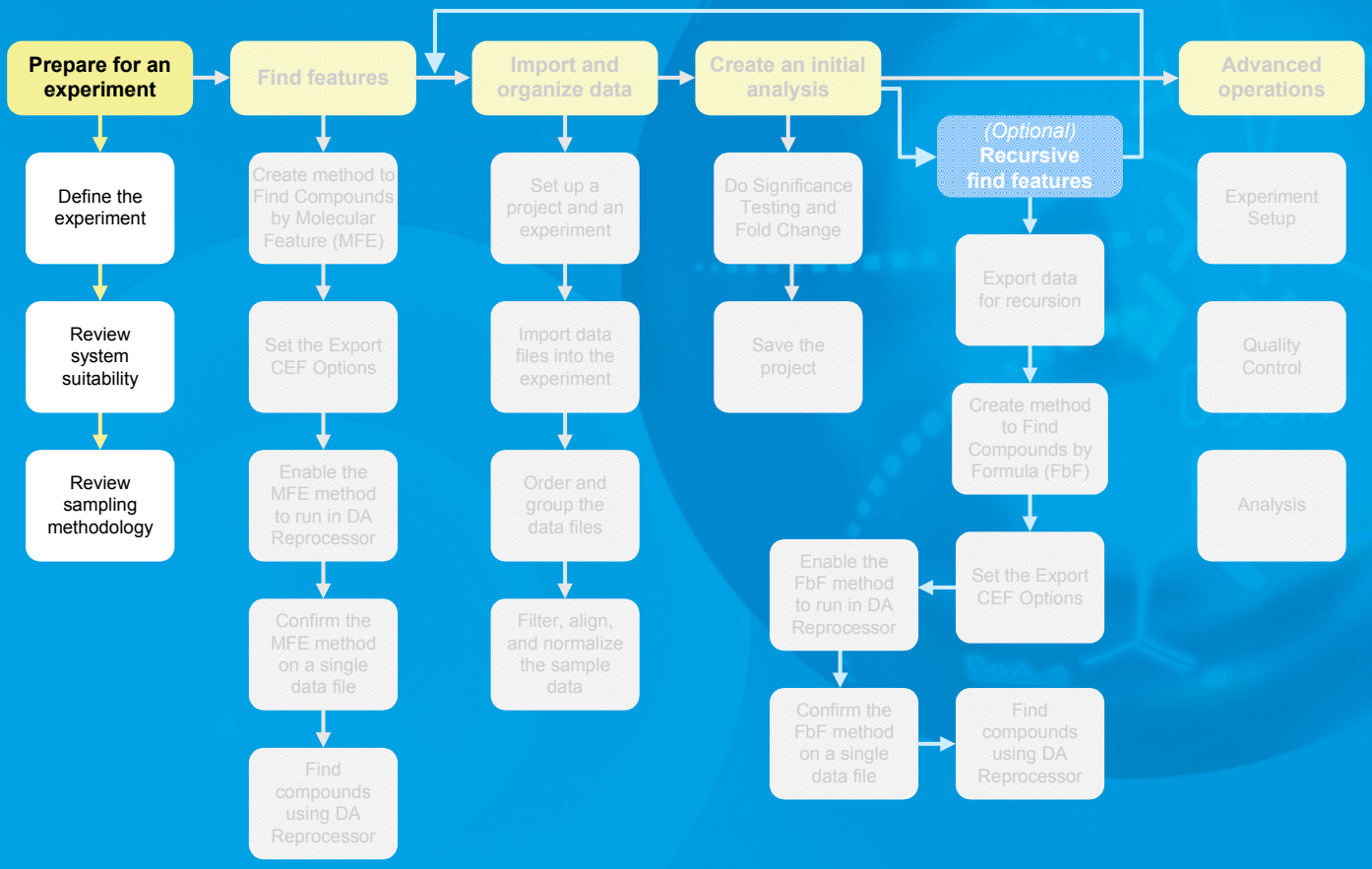
The Data Acquisition Compliance program is installed automatically with the MassHunter Data Acquisition software. See *Agilent MassHunter Data Acquisition Compliance Software Quick Start Guide* (Agilent publication G3335-90098, Revision A, February 2011) for instructions on enabling and using the MassHunter Compliance Software.

Roles

When Compliance is enabled, only certain users can perform certain actions. For example, the user that logs on to the system to submit a study needs to have certain Quantitative Analysis privileges to automatically build the quantitative analysis method.

Prepare for an experiment

Formulate your hypothesis, the question of correlation that is answered by the analysis. Your preparation includes an experiment definition that considers natural variability and replicate sampling, and reviews system suitability and sampling methodology to improve the significance of your analysis.



Agilent Technologies

What is metabolomics?

Metabolomics is the study of the metabolite content of a cell or whole organism. Metabolomics studies let researchers view biological systems in a way that is different from but complementary to genomics, transcriptomics, and proteomics studies. Discovery metabolomics experiments involve examining an untargeted suite of metabolites, finding the metabolites with statistically significant variations in abundance within a set of experimental versus control samples, and answering questions related to causality and relationships. Metabolomics is a powerful, emerging discipline with a broad range of applications, including basic research, clinical research, drug development, environmental toxicology, crop optimization, and food science.

Agilent provides you with tools to perform metabolomics research involving the collection and analysis of complex data sets containing hundreds to thousands of metabolites. Comprehensive analysis of metabolomics data requires an analytical approach and data analysis strategy that is unique and requires specialized data analysis software that enables cheminformatics analysis, bioinformatics, and statistics. Agilent MassHunter Qualitative Analysis and Agilent Mass Profiler Professional together enable metabolomics data analysis.

Experiment variables are derived from your experiment. When one or more of the attributes of the state of the organism are manipulated those attributes are referred to as independent variables. The biological response to the change in the attributes may manifest in a change in the metabolic profile. Each metabolite that undergoes a change in expressed concentration is referred to as a dependent variable. Metabolites that do not show any change with respect to the independent variable may be valuable as control or reference signals.

The metabolites in a sample may be individually referred to as a **compound**, **feature**, **element**, or **entity** during the various steps of the metabolomic data analysis. When hundreds to thousands of dependent variables (e.g., metabolites) are available, chemometric data analyses is employed to reveal accurate and statistically meaningful correlations between the attributes (independent variables) and the metabolic profile (dependent variables). Meaningful information learned from the metabolite responses can subsequently be used for clinical diagnostics, for understanding the onset and progression of human diseases, and for treatment assessment. Therefore, metabolomic analyses are poised to answer questions related to causality and relationship as applied to chemically complex systems, such as organisms.

The metabolomics workflow may be used to perform the following analyses:

- Compare two or more biological groups
- Find and identify potential biomarkers
- Look for biomarkers of toxicology
- Understand biological pathways
- Discover new metabolites
- Develop data mining and data processing procedures that produce characteristic markers for a set of samples
- Construct statistical models for sample classification.

Agilent enables metabolomics research for a variety of applications:

Basic and clinical research - Identify and validate metabolite biomarkers that correlate with disease states as well as provide fundamental insights into biology

Pharmaceutical - Identify metabolites and markers of toxicity for drug discovery and development

Agriculture - Identify and understand metabolic pathways to optimize crop development, yield improvement, and pesticide/herbicide resistance

Environmental - Identify metabolites that relate to the effects of chemicals and other stressors in the environment on a biological system

Biofuels - Identify metabolite profiles to optimize fermentation processes and biofuel production

Food / Nutrition - Identify the presence or absence of metabolites that correlate with major traits such as food quality, authenticity, taste, and nutritional value, and aid in the development of nutraceuticals

What are metabolites?

Metabolites - small organic molecules - are important modulators, substrates, byproducts, and building blocks of many different biological processes. Because of their importance, the presence or absence of specific metabolites in a cell or sample provides important information about the physiological and functional status of the biological system or test sample. The accumulation of a specific metabolite may signal a defect in a pathway, activation of a signal response pathway, or optimization of a biosynthetic pathway.

What is discovery metabolomics?

Discovery metabolomics experiments involve examining an untargeted suite of metabolites, finding the metabolites with statistically significant variations in abundance within a set of experimental versus control samples, and determining their chemical structure. Pathway analysis lets you connect the metabolite with the biological process or condition.

Small molecule tuning (LC/MS)

Metabolomics involves the analysis of small molecules, molecules nominally with a molecular weight from 50 to 6000 amu. Since the best results for metabolomics studies involve the identification of the exact mass of the molecular ion, LC/MS instrument tunes should be adjusted to (1) improve the sensitivity for intact molecular ions and (2) improve the overall sensitivity for small, low-mass, compounds. A typical automated instrument tune optimizes the instrument sensitivity across the entire mass range and may result in lower sensitivity for small molecules combined with higher fragmentation than desired for metabolomics.

Examples

Agilent Tools

A typical Agilent metabolomics workflow is illustrated in Figure 2 on page 16 starting with data acquisition through to analysis involving both untargeted (discovery) LC/MS and targeted (confirmation) LC/MS/MS analyses. Molecular feature extraction (MFE) and Find by Formula (FbF) are two different algorithms used by Mass-

Hunter Qualitative Analysis for finding compounds. All results files generated by Agilent analytical platforms can be imported into Mass Profiler Professional for quality control, statistical analysis and visualization, and interpretation.

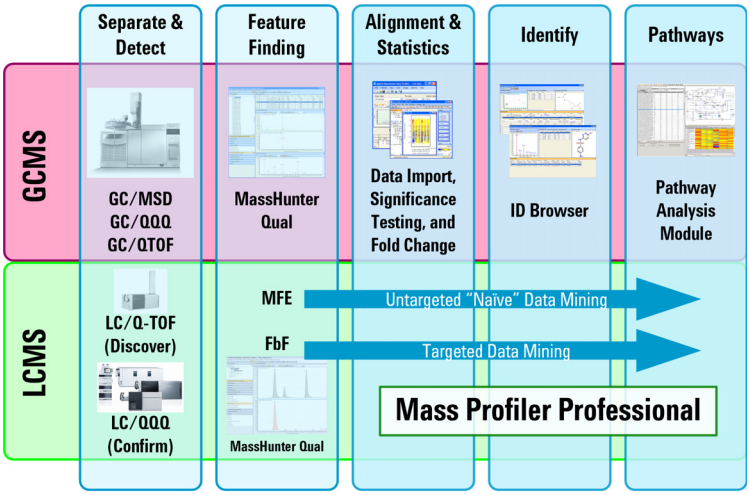


Figure 2 *An Agilent metabolomics workflow from separation to pathway analysis typically involving either or both GC/MS and LC/MS analyses.*

Principal Component Analysis

A principal component analysis (PCA) of metabolite data from replicate samples shown in Figure 3 highlights variability in metabolite abundance profiles between Infected versus Control samples extracted at pH 7. A significant amount of the variability at pH 7 is contained in the z component plotted on the Z-Axis. However, the z component of this PCA does not comprise the significant variability between the infected versus the control samples at pH 2 or pH 9.

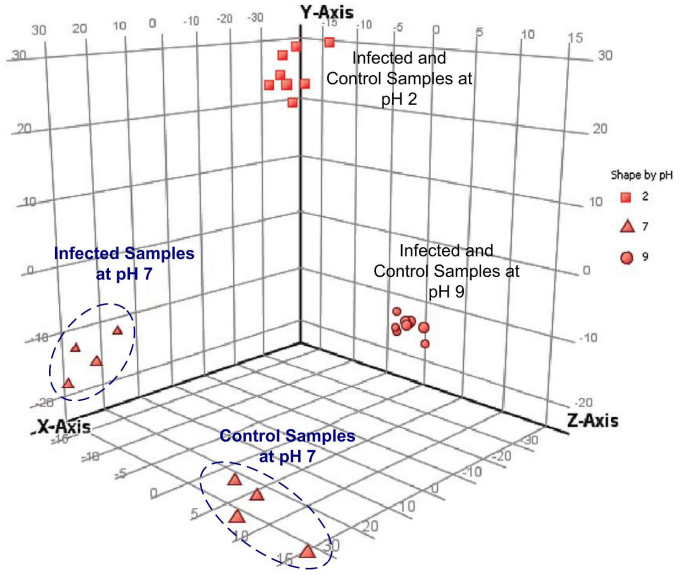


Figure 3 *Principal component analysis of metabolite data from replicate samples highlights variability in metabolite abundance profiles between Infected versus Control samples extracted at pH 7.*

Statistical Treatment

PCA may be combined with common statistical treatments to find improved discrimination among the independent variables and find the effect of the independent variables on the species under evaluation. Figure 4 shows how the effects of an experiment involving two attributes (Species and Treatment) are able to be statistically and visually discerned when processed using PCA and analysis of variance (ANOVA). PCA analysis only differentiated between the species. When PCA was combined with ANOVA, differentiation was obtained not only for both species but also for the affected status regardless of the species. The statistical results generated by using Mass Profiler Professional can be used to generate a sample class prediction (SCP) data model. The SCP model can subsequently be used to automate future identification of affected versus unaffected samples from the target species and thereby provide an indication for the application of the appropriate treatment.

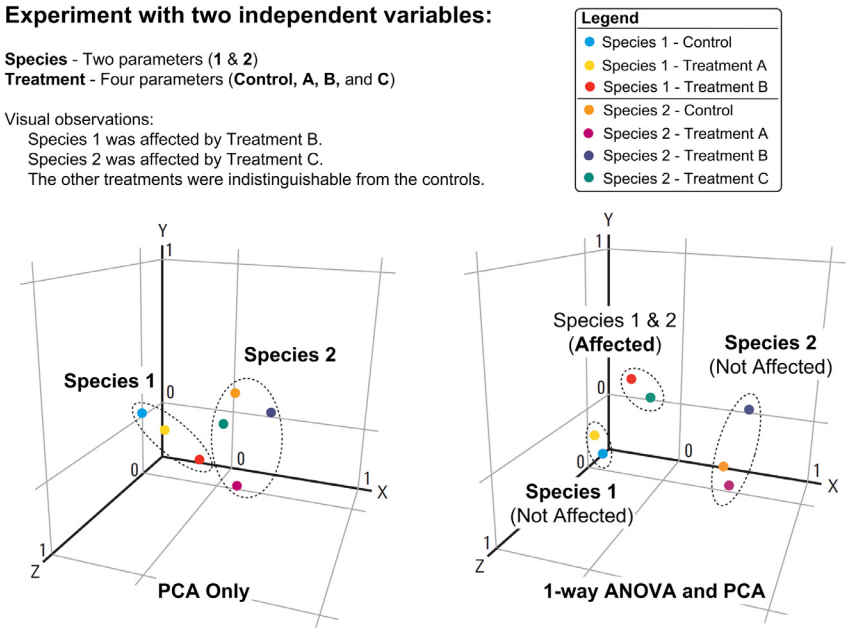


Figure 4 *Principal component analysis without prefiltering of the data and combined with 1-way analysis of variance (ANOVA).*

Summary

When you have hundreds, thousands, and more dependent variables (e.g., a metabolic profile), conventional target data analysis and correlation becomes difficult and resource prohibitive. Chemometric data analyses using Agilent Mass Profiler Professional lets you obtain accurate and statistically meaningful information correlating changes within the metabolic profile to established independent variables. The results may be used to understand the effects of your experiment in biological context using the Pathways analysis module and to develop a class prediction model that may be applied to new samples.

Capabilities of the metabolomics workflow

Mass spectrometry has been utilized in metabolomics research due to its wide dynamic range, reproducible quantitative analysis, and the ability to analyze very complex biological matrices. Due to the complex nature of these samples, separation (gas chromatography, liquid chromatography, or capillary electrophoresis) is often performed before mass analysis to facilitate the detection of as many metabolites as possible. The most compatible separation and analysis techniques for common classes of compounds are shown in Figure 5.

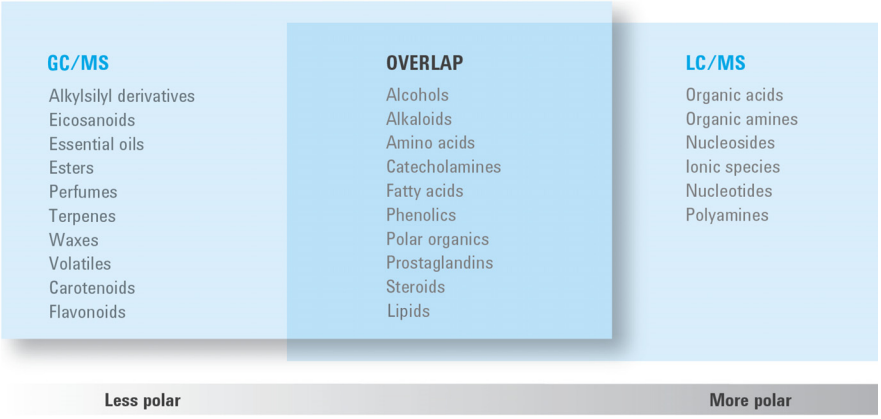


Figure 5 *Chemical classes suitable for GC/MS versus LC/MS.*

Discovery metabolomics using Mass Profiler Professional involves the comparison of metabolomes (the full metabolite complement of an organism) between control and test groups to find differences in their profiles. Usually, several steps are involved in discovery metabolomics analyses: profiling, identification, and interpretation.

Profiling

Profiling (also known as differential expression analysis) involves finding the interesting metabolites with statistically significant variations in abundance within a set of experimental and control samples. Profiling combines the targeted profiling of known metabolites with comprehensive feature extraction to find unexpected metabolites. A metabolite is represented by its molecular features, which are defined by the combination of retention time, mass or mass spectra and abundance.

The steps involved in profiling are

1. **Analysis:** Analysis of samples by GC/MS or LC/MS is required to separate and detect all of of the metabolites in the sample. Because metabolomes are large and exhibit significant natural variation, even in normal organisms, real differences can only be seen by analyzing large numbers of samples containing large numbers of compounds. Analytical instruments that have low error rates and facilitate high throughput are necessary.

2. **Feature Finding:** Feature finding specifically identifies all of the metabolites by mass and retention time. An undetected metabolite is a lost opportunity; it is essential to find as many of the metabolites in a sample as possible. This goes beyond simple chromatographic peak finding. Even with the best separation, a peak can contain multiple components. You can find multiple components in the same peak using the MassHunter Qualitative Analysis software as shown in Figure 6 on page 19.

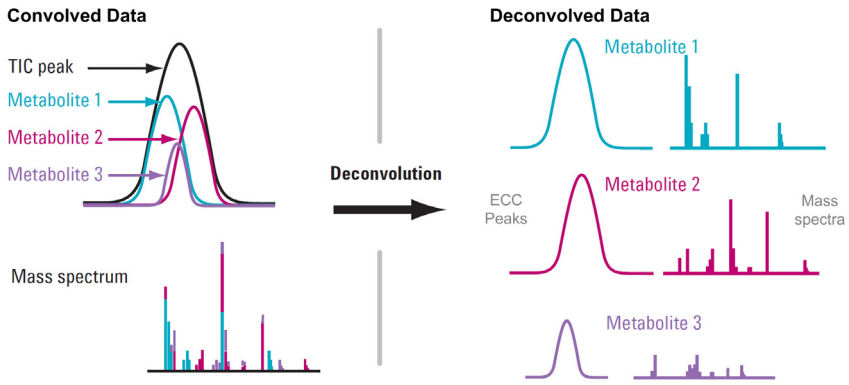


Figure 6 *Deconvolution using MassHunter Qualitative Analysis finds metabolites that are chromatographically unresolved or poorly resolved. Deconvolution generates an extracted ion chromatogram and a reconstructed, single component spectrum for each metabolite.*

3. **Data Normalization:** Normalization lets data collected over a period of time be corrected for changes in retention time and/or response so that a single feature common to several samples is not treated as a unique feature being separately sought in each sample.

4. **Statistical Analysis:** Statistical data analysis is used to discover significant differences between the sample sets.

Identification

Identification is the determination of the chemical structure of these metabolites after profiling. EI spectra from GC/MS are well suited to spectral library searching. LC/MS spectral data are evaluated using searches performed with a database of metabolite information to help narrow the list of possible candidates. Accurate mass data makes this database searching more effective by narrowing the mass window that needs to be searched and thus reducing the number of possible identities.

Interpretation

Interpretation, the last step in the workflow, makes connections between the metabolites discovered and the biological processes or conditions. Once the metabolites are identified, it is necessary to understand their relation to biological pathways in metabolism by interpreting the results of the experiment. Pathway analysis makes connections between the metabolite markers discovered and the biological processes or conditions being studied, and helps to elucidate the biological relevance of metabolomics data in a systems context. Pathway analysis typically requires integration of metabolomics data with genomics and proteomics data.

Introduction to the workflow

The Agilent metabolomics workflow consists of six steps. You are encouraged to follow the steps using your own data.

Step 1

Prepare for an Experiment

The first and most important step in the metabolomics workflow is to formulate your hypothesis, the question of correlation that is answered by the analysis. Your preparation includes an experiment definition that considers natural variability and replicate sampling, and reviews system suitability and sampling methodology to improve the significance of your analysis.

Step 2

Finding Features

Find Compounds by Molecular Feature (MassHunter Qualitative Analysis)
Compounds, referred to as molecular features, are extracted from your data based on mass spectral and chromatographic characteristics. The process is referred to as Molecular Feature Extraction (MFE). Molecular feature extraction quickly and automatically generates a complete, accurate list of your compounds which include molecular weight, retention time, m/z, and abundance.

Step 3

Import and Organize Data

Organize, import, and prepare your data (Mass Profiler Professional)
After you create a project and an experiment, the "MS Experiment Creation" guides you through the necessary steps to organize your experiment, import your data, define your experimental variables, and prepare your data for analysis. The data preparation includes filtering, alignment, normalization, and baselining.

Step 4

Create an Initial Analysis

Quality control and initial differential expression (Mass Profiler Professional)
The "Significance Testing and Fold Change Wizard" guides you through the necessary steps to enter parameters and values that improve the quality of your results and produce an initial differential expression for your analysis.

Step 5

Recursive Find Features *(Optional)*

Find Compounds by Formula (MassHunter Qualitative Analysis)
Importing the most significant features back into MassHunter Qualitative Analysis as targeted features improves finding the features in your samples. This repeated feature finding is referred to as recursion. Improved reliability in finding your features leads to improvement in the accuracy of your analysis.

Step 6

Advanced Operations

Customize your analysis and interpret the results (Mass Profiler Professional)
The most significant features in your data are processed by Mass Profiler Professional into a final statistical analysis and interpretation. The results from the final interpretation may be used to prove or disprove your hypothesis and may be used to create a sample class prediction model.

Features of example experiments

The metabolomics workflow is illustrated using an experiment with two independent variables. In "Advanced operations" on page 117 some of the capabilities of Mass Professional are illustrated with an experiment with a single independent variable. The processes used to generate results from these experiments helps you use MassHunter Qualitative Analysis, Mass Profiler Professional, and ID Browser with your experiments.

Definitions

Terms and definitions used in metabolomics and metabolomic analyses vary. It is recommended that you refer to the "Definitions" on page 180 for a list of terms and their definitions as used in Mass Profiler Professional and in this workflow.

One-variable experiment

The one-variable experiment presents an analysis of a metabolomic response to changes in a single independent variable, also referred to as a parameter. The data was acquired using four (4) parameter values for the independent variable. The parameter values consist of a single control data set that represents the organism without perturbation and data sets from three variations where the organism is subject to one of three conditions established by the experiment design. In summary, the one-variable experiment contains a single parameter with four parameter values and ten replicate samples for each parameter value.

An ideal experiment involves at least ten (10) replicates for each parameter value. Thus an ideal experiment with a single parameter and four parameter values has a data sample size of at least forty (40) samples. In this example the minimum sampling conditions are met.

Two-variable experiment

The two-variable experiment presents an analysis of a metabolomic response to changes in two independent variables (parameters), each with two parameter values. The parameter values of the first parameter represent a control data set associated with the organism without perturbation and when the organism was subject to a known perturbation. The parameter values of the second parameter represent a pair of metabolite extraction techniques where the first parameter value represents the current state-of-the-art extraction process and the second parameter value represent the addition of a step designed to improve metabolite extraction. In summary, the two-variable experiment contains two parameters with two parameter values, for a total of four permutations, and four replicate samples were obtained for each permutation.

An ideal experiment involves at least ten (10) replicates for each parameter value. Thus an ideal experiment with two parameters, each with two parameter values, has a data sample size of forty (40) samples (See "Replicate data" on page 36). The ideal sample size is calculated by multiplying 2 parameters by 2 parameter values for each parameter and then multiplying by 10 replicates for an ideal minimum sample size of forty (2 x 2 x 10 = 40) samples. In this example the minimum sampling conditions are not met; four replicates exist for each permutation for a total of sixteen (16) samples. While the sampling falls short of the minimum sampling recommendation, the strong correlation of cause and effect in this experiment overcomes the sampling deficiency and provides support for further investment in the metabolomics question being studied.

Define the experiment

A complete metabolic understanding of a living organism involves the identification and quantitation of all of the metabolites of the organism's cellular makeup in a given state at a given point in time. Commensurately, the complete metabolic understanding includes an understanding of the metabolic response and the change in the metabolic composition from the organism's normal function with respect to altered function(s) caused by specific disease or external influences.

Metabolic pathways are complex and interconnected (see Figure 7) making exact correlation of any change in a *single metabolite* with the organism's response to disease or other influence nearly impossible. However studies show that the organism's response to disease or other influences does correlate with a change in the *metabolic profile* of the organism.

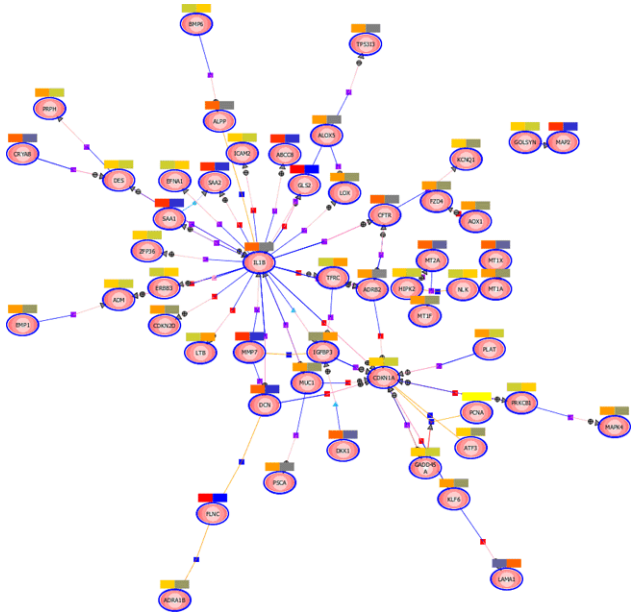


Figure 7 *Metabolic pathways are complex and interconnected, reducing the effectiveness of correlating an organism's response to a perturbation with changes in a single metabolite compared to correlating the organism's response to a change in the metabolic profile.*

Agilent's suite of tools helps medical and research personnel correlate changes in the metabolic profile of empirical samples with changes made in a well defined set of conditions. After subjecting replicate research and diagnostic samples to well defined treatments and exposures (independent variables), as well as maintaining a set of controls (control samples), researchers use Agilent tools to facilitate the interpretation of acquired metabolomics profile data. These tools perform data extraction, data processing, and statistical analysis, which leads to answers regarding critical questions of cause and effect.

Whether the organism is exposed to disease or chemical influence, a change in the organism's state and subsequent metabolic profile occur. Agilent MassHunter Qualitative Analysis and Agilent Mass Profiler Professional make possible the analysis of the metabolic profile and correlation of the changes in the metabolic profile to the organism's changed state. They enable the process by which the vast data gener-

ated in metabolomic studies is reduced to the significant information from which cause and effect correlations may be made.

What is an experiment?

An experiment consists of the analysis of a set of replicate samples collected from the organism as it was exposed to well defined treatments and exposures (independent variables), as well as when the organism was exposed to a set of controls representing minimal or normal perturbations (control samples). The results of changes observed in the samples is designed to provide an answer to your hypothesis. The hypothesis may be proved or disproved by analyzing the correlation of the independent variables on the resulting expression of a large number of dependent variables - the metabolite constituents of the samples. The results must be significant beyond natural variability.

An experiment involves the following steps:

1. Ask a question.
2. Formulate your hypothesis.
3. Assess the chemical makeup.
4. Identify your variables.
5. Verify that metabolomics simplifies your analysis.
6. Design your experimental plan including controls and sampling.
7. Execute your plan.

After you obtain your samples and acquire your data using hyphenated mass spectrometry techniques, the metabolomics workflow takes you through data extraction, processing, and statistical analysis so that you can prove or disprove your hypothesis.

Using questions to formulate a hypothesis

A hypothesis may be derived through proposing a number of more or less specific questions that are important to your research. By evaluating these questions and evaluating their relationship with the available analytical approach to solve the problem, you may identify the key variables, the independent variables, and the dependent variables.

Formulating the hypothesis

The most important step in designing your experiment is to formulate a proposition that explains certain facts of correlation among your variables that can be answered by the analysis. The proposition, expressed as a question referred to as the hypothesis, puts forth a possible correlation, for example a cause and effect, between a set of defined attributes (independent variables) and the resulting metabolic profile (dependent variables). Tentatively accepting your hypothesis provides the basis of the statistical parameters employed while following the metabolomics workflow. The metabolomics workflow is used to prove or disprove your hypothesis.

Key elements in the sample specimens

A successful experiment requires several key elements to be present in the samples that are analyzed. The presence of these elements is a reasonable indicator that data acquired from the samples is suitable for metabolomic analysis to prove or disprove your hypothesis. The samples are taken from a specimen, an individual organism (e.g., a person, animal, plant, or other organism) of a class or group that is used as a representative of a whole class or group. For clarification on the terms used throughout the workflow, see "Definitions" on page 180. There are six key elements.

Many chemical constituents: Samples taken from a specimen should contain many chemical constituents (metabolites) that are naturally expressed by the specimen or are a product of the specimen during its interaction with controllable externalities.

Complex relationship among the chemical constituents: The relative composition of the chemical constituents within each sample is the result of complex relationships, such as the biological functions within the specimen as shown in Figure 7 on page 22.

Subject specimens: The specimens selected for evaluation represent a random set of specimens selected from a larger population. If the specimens come from organisms subject to specific and known perturbations in their externalities (independent variables) they are known as *subject specimens*. If the specimens come from organisms that are not perturbed in their externalities they are known as *control specimens*.

Control specimens: A sub-group existing within the sample specimen population that represents the "normal" attributes of the organism under natural, unperturbed, externalities. The inclusion of control specimens reduces the occurrence of false positive and false negative correlations and provides a point of reference for assigning a polarity to any observed effect.

Independent variables: The specific and known perturbations that may be made to a specimen, or quantifiable externalities that affect the specimen, and which are expected to induce a measurable change in the metabolic profile. The relative concentrations of the chemical constituents may be influenced by attributes, or parameters, that are independently controlled through either, (a) informed selection of the specimens (controlled sampling), or (b) application of a set of experimentally controlled externalities or attributes (experiments).

Dependent variables: The measurable response of the specimen to the specific and known perturbations made through variation of the independent variables. The specimen response is manifest in the relative concentration of the chemical constituents (metabolic profile) with respect to each varied attribute. The data is collected using hyphenated mass spectrometry techniques. If the response of the metabolic profile to the independent variables is known and finite, then traditional quantitation techniques may be applicable (for example, GC/MS may be used to study how the abundance of marker metabolites change when organisms are treated versus not treated for a condition). Otherwise, if the response is unknown or complex, and spans a large number of chemical constituents, the data is suitable for analysis using the metabolomic workflow.

Using restatements of the hypothesis to identify independent variables

Evaluating the hypothesis and various restatements of the hypothesis helps you identify the independent variables. The independent variable is something that you can influence through sampling or an experiment. The dependent variable is always the same, the myriad of chemical constituents expressed by the specimen that may or may not change in relation to variations selected as independent variables. As you analyze your data you reduce the dimensions of the data (the relationships among the independent and dependent variables) while retaining those dimensions of the data that contribute most to the characteristic variability of the data. Retaining only the dimensionality of the data that best shows correlations between the specimen and the independent variables facilitates sample classification.

Example - grape wine

A hypothetical example involving grape wine is described to help illustrate application of the key elements to the formulation of an experiment's hypothesis.

1. Ask a question: "Can grape wine be correlated to the country of origin?" Proper characterization of the geographical origin of wine is important to quality control and important to revenue taxation across political boundaries since mislabeling the country of origin may lead to financial gain. The question is directed to finding a correlation between a wine and country of origin of the wine that is independent of wine variety for the single purpose of political taxation.

2. Formulate a hypothesis: A more precise question stated as a hypothesis is "Does the chemical makeup of grape wine provide a signature (metabolic profile) that correlates to the country of origin?"

3. Assess the chemical makeup: Grape wines are known to have a very complex chemical makeup. The presence of a complex relationship among many chemical constituents, derived from grapes and grape fermentation, and the ability to affect specific and known perturbations to the sample specimens through controlled sampling leads to a question suitable for answering using the metabolomics workflow.

Since the correlation of the independent and dependent variables involves the analysis of thousands of chemicals and potentially hundreds of samples, the question "Can grape wine be correlated to the country of origin?" is an ideal question to address using metabolomics.

4. Identify the variables: The potential independent variables are the county of origin of the grape wine production and the wine variety. The dependent variables are manifest in the several thousand chemical constituents that are produced by the grapes and the wine making process.

Even though the question presented refers to only one variable, two key variables may affect the chemical makeup of the wine: (1) country of origin of the grape and wine production and (2) wine variety. While wine variety is not necessarily germane to the question, if this information is known and entered with the sample data, it may provide a beneficial statistical parameter to the metabolomic analysis and may even provide for the future analysis of another hypothesis with minimal additional effort.

5. Verify that metabolomics simplifies a traditionally complex analysis: Traditional correlative analysis of the question would require the identification and quantification of thousands of chemical constituents and studying how the expression of each constituent is affected by the change in geographical location, while ruling out false effects that may be manifest as a result of the various grape species and wine varieties that are also involved in the study. Without going into any calculations, a traditional correlative analysis of this hypothetical example should be readily apparent as a costly undertaking. The statistical approach employed by metabolomics makes this unwieldy analysis manageable.

If we continued developing this example experiment, the final steps *6* and *7* are to design the experimental plan including controls and sampling and executing the plan, respectively.

Natural variability

Before any statistical analysis is begun, it is important to understand how a sample taken from any one specimen represents the population as a whole and how increasing the sample size improves the accuracy of the sample set in describing characteristics of the population.

Under identical conditions, all life systems produce a range of results. Specimens taken from the population may show one of the following characteristics:

(1) Results comparable to the mean of the population (i.e., characteristics shown by the majority of the population), for example results within ±1 standard deviation (~68%) from the mean.

(2) Results that differ significantly from those shown by the majority of the population (i.e., characteristics that are not shown by the majority of the population), for example results beyond ±3 standard deviations (~99.7%) from the mean.

(3) Results anywhere in between ±1 to ±3 standard deviations from the mean.

In many biological and biochemical systems characteristics are found to show a probability of variation referred to as a normal distribution. Figure 8 on page 27, for instance, shows a normal distribution of a characteristic within a population where 68% of the sampled population would be shown to have the mean characteristic plus or minus one standard deviation (σ). This natural variation of the population response to identical conditions is referred to as natural variability. Natural variability thus means that any single sample specimen taken from a population is not guaranteed to reflect the mean characteristics of the population.

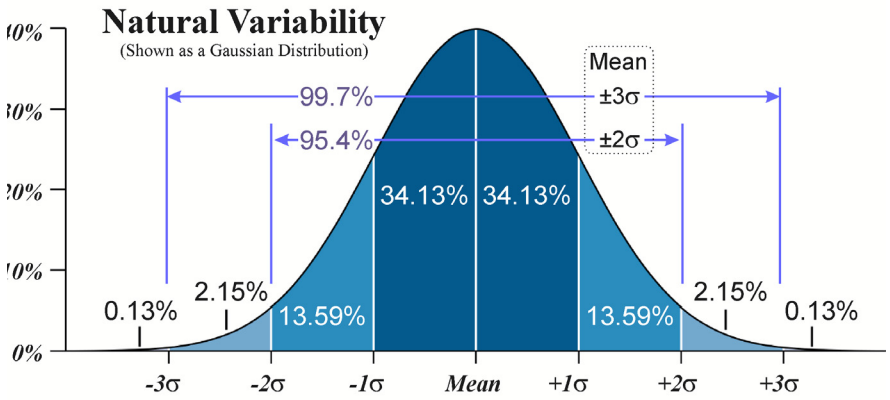


Figure 8 *Natural variability shown as a Gaussian (normal) distribution. Depending on the predefined requirement for significance, if the mean of a sample set is beyond ±2σ from the natural variation there may be a significant effect. Similarly, if a particular observation routinely falls beyond ±2σ from the natural variability of the data the change producing the effect may be considered significant.*

Natural variability is found to occur from inherent randomness and unpredictability in the natural world. Natural variability is found in all life and natural sciences and in all forms of engineering. For example, a population of plants grown under identical conditions of illumination, precipitation, and nutrient availability shows a range in growth mass. This range of variable growth mass may be expressed as a mean where 95% of the population is expected to show a natural range of variability within two standard deviations of the mean (see Figure 8).

In other words, for a set of fixed attributes (independent variables), a representative set of samples taken from the population of plants shows a natural variability in the dependent variable growth mass. When an experiment is undertaken where plants from the same population are subject to variations in the fixed attributes, the plants response shows a change in growth mass in addition to their natural variability in growth mass. Thus if the entire population is sampled, we see two adjacent normal distributions with means reflecting the plant growth mass under the two conditions (see Figure 9 on page 28).

Likewise, an animal population subject to any controlled exposure shows a naturally varying effect of that exposure as expressed through the chemical makeup of serum samples taken from the population. Such unpredictability in the measurable variability of any biochemical expression must not be mistakenly correlated with deliberate variations of an independent variable.

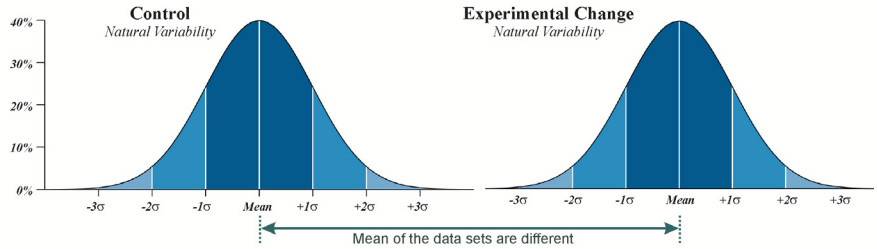


Figure 9 *Natural variability of populations with subject to two different experimental conditions where the means of each data set falls outside of the mean ±3 of the other data set.*

During your experiment the natural variability of the data representing a population must be understood in order to confidently express any experimental correlation. The investment of time and resources in performing a statistical analysis requires that the natural variability of the subject specimen be known or reasonably estimated so that the results of the analysis may be conclusively shown to be either within the natural variability (no correlation) or outside of the natural variability and therefore provide for a degree of correlation with the independent variable(s).

Experimental data collection that does not incorporate consideration of the natural variability of the data does not yield meaningful results. Thus, crucial to the metabolomics workflow, as with all statistical data treatments, is an understanding and well planned collection of the data; without that, the results follow the adage "garbage in, garbage out."

Replicate data

Replicate sampling and measurement of many specimens from the population is the only way to estimate the natural variability of your data. No guarantee exists that a single sample specimen from a population represents the mean of the population. Any single sample from a population with a natural variability shown in Figure 8 on page 27 has a 99.99% chance that it lies within four standard deviations (±4) of the mean of the true population, but in fact that single sample may on a rare occasion fall even further from the population mean.

However, if ten (10) samples are taken from the population, the mean of these samples produces a statistically more accurate approximation of the true mean of the population. The accuracy of the approximation of the true population mean proportionally improves with more samples. The true value of the population mean is achieved only if the entire population is sampled. However, sampling the entire population is not typically feasible because of constraints imposed by time, resources, and finances. On the other hand, evaluating fewer samples increases the chance of false negative and false positive correlations from your experiment.

Too few samples may lead to an incorrect conclusion. Figure 10 on page 29 shows that if too few samples are evaluated, and if these samples just happen to be samples lying far from the mean because of natural variability, an incorrect conclusion may be drawn that the change in the independent variable produced no significant change in the response. The estimate of the standard deviation of the sample mean

estimate of a population mean (standard error) is equal to the standard deviation of the samples divided by the square root of the number of samples (Equation 1).

Equation 1 $$SE = \sigma/(\sqrt{N})$$

where *SE* = standard error of the population; σ = standard deviation; and *N* = number of samples.

Large sample sizes lead to more confident conclusions. As the sample size increases, the likelihood that the data approximates the true response of the population increases. The standard deviation of the sample may become smaller, and the likelihood of making a correct correlation between cause and effect is improved (Equation 2).

Equation 2 $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})}$$

where σ = standard deviation; *N* = number of samples; x_i = the value of an individual sample; and $\bar{X}$ = mean (or average) of all *N* samples.

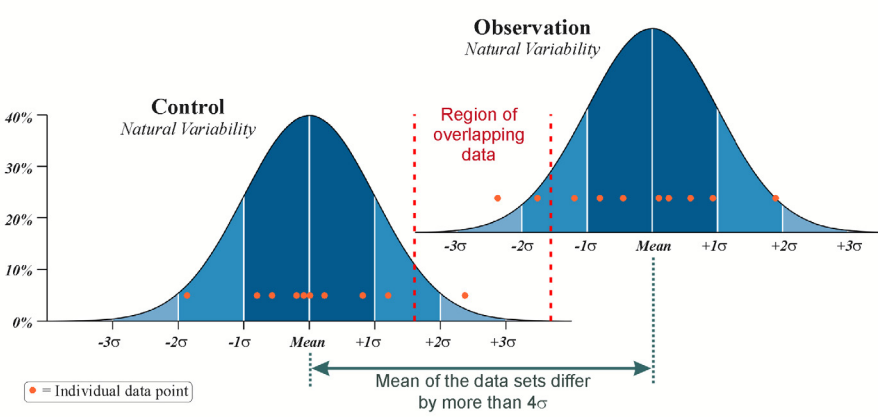


Figure 10 *Replicate data are necessary to distinguish whether the represented populations actually show significant differences. The three data points in the region of overlap of the natural variability may, if too few replicates are selected, lead to a result suggesting a less significant difference in the populations.*

Successful application of metabolomics analyses depends on the availability of sufficient replicate samples and specimens. Coupled with an understanding of the systems under study and a well planned collection of the samples and concomitant data, the statistical data treatment of the replicate samples is the backbone of the metabolomics workflow. A sufficient set of replicate data, ten (10) or more repli-

cates, may provide a significant answer to the hypothesis and prevent a total loss of time and resources invested in performing the described statistical analyses.

Summary

An experiment involves the formulation of a hypothesis that may be answered by analyzing the potential correlation of independent variables on the resulting expression of a large number of dependent variables – potentially interrelated chemical constituents – through the analysis of replicate samples whose results are significant beyond natural variability. Metabolomics analyzes the data acquired in your experiment to answer questions related to causality and relationships from among the chemically complex systems where traditional quantitation techniques cannot handle the vast amount of data produced.

Review system suitability

In order for data to be meaningfully, statistically processed, the spectral data acquired over the time scale of the experiment must have a consistent methodology and means to adjust for experimental variability. Typical GC/MS and LC/MS acquisitions provide three dimensional data: chromatographic retention time, ion mass to charge ratio, and ion signal intensity (Figure 11). Individual data collected at various periods of time over days, weeks, and months may show variations in any or all three of these dimensions. Such variations are referred to as system drift and may occur even when the same method and instrument are used for all of the data collection. While regular instrument tuning and maintenance can help reduce system drift, the means to adjust for this experimental variability is called system suitability.

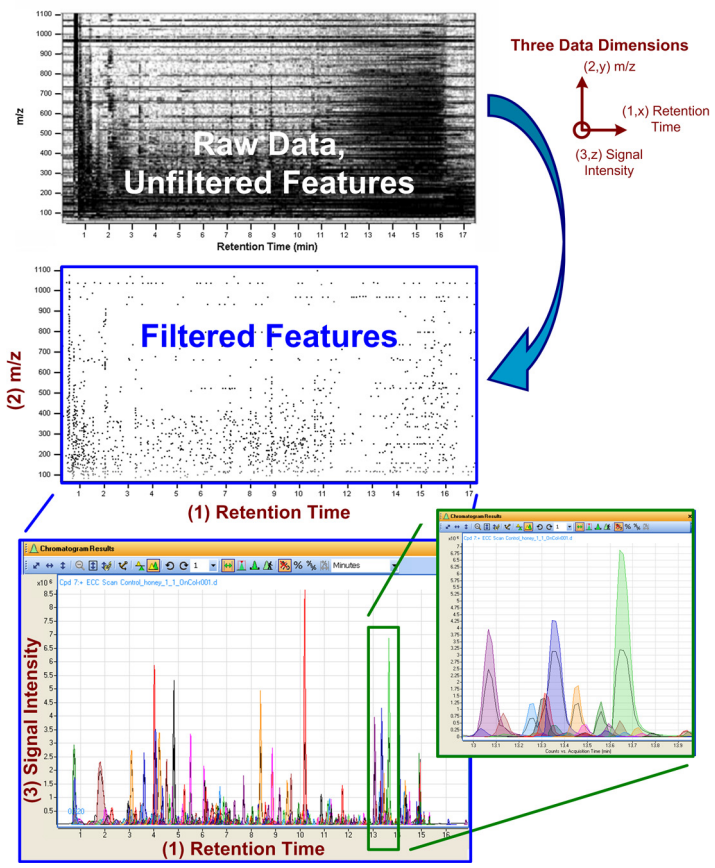


Figure 11 *GC/MS and LC/MS data provide three dimensions of information: retention time, m/z, and signal intensity.*

System suitability includes regular instrument tuning and maintenance to reduce system drift. Additionally, the Agilent metabolomics workflow provides a means to minimize and remove instrumental variations that may affect each of the data dimensions (retention time, m/z, and signal intensity). The processes to assure quality include retention time alignment, intensity normalization, mass variation, and baselining. Agilent Mass Profiler Professional performs retention time alignment, intensity normalization, and baselining. Agilent MassHunter Qualitative Analysis performs mass variation.

Assuring data quality

Quality data that is the result of a well planned experimental design has stable dimensionality and internal markers, such as internal standards, that let the data be compared and correlated to data collected over different periods of time, collected by different operators, and collected from different instruments. Employing retention time alignment, intensity normalization, mass variation, and baselining provides both a quality check to evaluate system drift and a means to adjust for instrumental variations to assure the meaningful results from your analysis.

Retention time alignment

Retention time alignment is used to adjust the chromatographic retention time of the eluting components based on the elution of specific components that are (1) naturally present in each sample or (2) deliberately added to the sample as a known compound or set of compounds that do not interfere with the sample. Alignment of retention time removes temporal variations of the chromatographic separation improving the correlation of identical features among data sets collected under different conditions; for example, different separation columns or identical separation columns but with different conditioning or mobile phase composition.

Unidentified compounds from different samples are aligned if they are within a specified tolerance, retention time window, and provided that the compounds meet criteria for mass spectral similarity. GC/MS data alignment is performed as unidentified compounds.

Identified compounds from different samples are aligned based on the similarities of the assigned attributes of compound name and ionization mode. LC/MS data alignment may be performed as identified compounds, unidentified compounds, or both (referred to as combined in the Mass Profiler Professional software).

More information regarding the specific details of alignment may be found in section 3.1.7 Experiment Creation > Getting Started > Alignment Parameters of the *Mass Profiler Professional User Manual*.

Intensity normalization

Normalization of feature intensities is used to adjust the intensity value from the absolute value of the extracted ion chromatogram (EIC) signal measured at the detector to a relative intensity based on the signal provided by either an (1) internal standard, (2) an external scalar, or (3) both an internal standard and an external scalar.

The highest quality data is achieved using standards in the data collection. Standards provide a means to align retention time and evaluate and compensate for variations in the m/z and ion intensity of the analytical system. The standard may be (1) one or more known, identified constituents internal to the sample matrix, (2) fabricated from one or more chemical compounds specifically acquired and added (spiked) to each sample at a precise amount, or (3) a pooled sample from a representative biological matrix that is added to each sample. The use of pooled samples as a standard provides good value from existing materials and an ideal match to the rest of the analytical conditions.

An internal standard is selected from each sample data set as a means to minimize the system signal variability attributable to instrumentation or sample preparation. After performing an internal standard intensity normalization, Agilent Mass Profiler

Professional adjusts for system variability by normalizing the individual samples to a value that reflects the differences in the signal intensity of each compound across all the samples with respect to the internal standard.

More information regarding the specific details of normalization may be found in section 3.1.9 Experiment Creation > Getting Started > Normalization Criteria of the *Mass Profiler Professional User Manual*.

Mass variation

The adjustment of the mass to charge (m/z) resolution from unity in 1,000 m/z (one part per thousand, ppt) to 0.001 in 1,000 m/z (one part per million, ppm) facilitates the unique identification of compounds that may have nearly identical or identical chromatographic behavior. Since some methods may not separate all of the compounds, or the time involved for a complete separation may not necessarily be practical, the added dimension provided by adjusting the mass variation for extracting ion chromatograms facilitates unique identification of co-eluting compounds. The technique of using mass variation to identify co-eluting compounds is referred to as deconvolution.

The chromatographic separation thus need not separate all of the compounds as viewed in the total ion chromatogram (TIC) but does need to separate all compounds that have a unique molecular formula.

More information regarding the specific details of mass variation may be found in the MassHunter Qualitative Analysis on-line Help under Finding Compounds.

Baselining

Baselining is a means of changing the feature signal intensities (referred to as an abundance unit in Mass Profiler Professional) to a new signal intensity that relates the feature's strength in one sample with respect to the feature's strength among all of the sample data. The original abundance unit assigned to each feature is based on the instrumental parameters at the time the data was acquired.

For example, the variation of a feature's signal intensity may vary over several orders of magnitude across the sample data sets (i.e., variation from 10,000 to 10,000,000 with a mean value of 150,000) can be converted to variation that spans positive and negative values with respect to the mean. By adjusting the feature signal strength through baselining the feature strength may be statistically weighted to be more consistent with biological significance rather than with the feature's amenability for ionization and detection by the instrumental method.

More information regarding the specific details of standards may be found in section 3.1.10 Experiment Creation > Getting Started > Baselining Options of the *Mass Profiler Professional User Manual* and by reading statistical literature on baselining approaches.

Mass Profiler Professional employs four (4) options for baselining: none, Z-transform, baseline to median or mean of all samples, and baseline to median or mean of control samples.

None

None simply treats feature compounds with greater abundances as more significant than features with smaller abundances.

Z-transform

Z-transform works best with large data sets that in general do not have missing features among the data sets. Quantitative data is a good candidate for processing with this baselining approach.

Baseline to the median or mean of all samples

Baseline to the median or mean of all samples acts to treat the data sets with more even weight by substantially reducing the influence of features that have particularly low or very large signal intensities. In this approach the abundance of each feature is adjusted by subtracting the median or mean of the feature across all of the data sets: this results in a value of zero (0) abundance representing a feature that has a strength equal to the median or mean of the feature as present in all of the data sets. A negative or positive abundance is a feature present in abundance less than or greater than the median or mean of all of the samples, respectively.

Baseline to the median or mean of control samples

Baseline to the median or mean of control samples uses the control samples to calculate the median or mean intensity for each feature. In this approach the abundance of each feature is adjusted by subtracting the median or mean of the feature across the control data sets: this results in a value of zero (0) abundance representing a feature that has a strength equal to the median or mean of the feature as present on all of the control data sets. A negative or positive abundance is a feature present in abundance less than or greater than the median or mean of the control samples, respectively.

Summary

System suitability involves collecting data that provides a means for evaluating system drift and adjusting for instrumental variations to assure quality results. Four approaches are employed together in order to produce the highest quality results:

1. Retention time alignment
2. Intensity normalization by using internal standards in the data collection experiment
3. Chromatographic deconvolution using m/z variation
4. Baselining of the features across the data sets

Review sampling methodology

Obtaining analytical results from a limited number of samples to prove or disprove a hypothesis about a much larger population requires a sufficient number of samples and involves skills that resemble art in addition to science. The best results of a metabolomics statistical study involve using more than one method of sample collection. This inherently leads to the collection of more samples and more samples have a positive impact to the significance of the results. Four main methods of sample collection are census, survey, experiment, and observation.

Methodologies

The purpose of each of the sampling methodologies is to provide data that spans the desired range of the independent variables (variable states) which may include collecting data at various points of time (time points) during an experiment. Regardless of which methodology is used to collect the data, replicate data samples are critical to the successful metabolomics analysis.

Census

A census methodology collects samples from every member of a population. In most experimental designs a census is not practical because of the cost and resource commitment required.

Survey

A survey methodology obtains samples from a subset of the population in order to estimate the population attributes. Samples are taken from random members of the population. All, or a portion, of the survey samples may be used to provide an estimate of the natural variability of the population as the baseline for the experimental study. The same, or the remaining, survey samples may be subject to experiments and thus used to estimate the response of the population to the experimental conditions.

Experiment

All of the samples studied in metabolomics form part of an experiment designed to answer a hypothesis regarding cause and effect correlations. Experiment-based sampling may be in one of two forms:

(1) Controlled sampling: taking random samples from a population that meets experimentally defined criteria

(2) Experimental exposure: subjecting random samples taken from a total population to experimentally defined conditions

When the independent variables of an experiment exist as a natural part of the population, such as the geographical region a grape is grown, the experiment involves controlled sampling. When the hypothesis is in relation to the effect shown by an organism to controlled conditions, such as a response to proposed treatments, then a random sample of the population is subject to independent variables specified by the experimental exposure. In either case, the hypothesis is related to causality of one or more independent variables on the metabolic profile which may relate to observable effects. The experiment is controlled in the sense that (1) sample subjects are selected from defined populations and (2) known treatments are applied to each group of samples.

In the experiment analysis, results are evaluated by comparing the dependent variable (the metabolic profile) against the known conditions of the independent variables. A conclusion regarding the causality of the treatment on the organism's response is then made against the hypothesis.

Observation

Any attempt to understand causality where no ability exists to (1) control how subjects are sampled and/or (2) control the exposure each sample group receives is considered an observation study. The data obtained from an observation study may be useful as a control sample or, if sufficient minimal information about the sample exists, may be useful for quality control against an experimental correlation derived from the hypothesis.

Replicate data

Replicate sampling and measurement of many specimens from the population is the only way to estimate the natural variability of your data. Using more than one method of sample collection inherently leads to the collection of more samples and has a positive impact on the significance of the results by including more samples. The accuracy of the approximation of the true population mean proportionally improves with more samples. The true value of the population mean is only achieved if the entire population is sampled (census sampling).

It is recommended to use a sample set of at least ten (10) replicates for each parameter within an independent variable or within each permutation of parameters when more than one independent variable exists. Too few samples increases the chance for obtaining a false negative or false positive correlation.

Timing of sampling and analysis

Using the discussions regarding natural variability, replicate data, and sampling methodologies a plan for sampling and analysis to answer the hypothesis may be proposed. The plan should not only outline the sampling size, the sampling methods, and the statistical analyses to use but should also propose an orderly sequence, or timing, for the sampling and analysis so that the hypothesis may be evaluated at least twice during the course of the metabolomics workflow. By evaluating a set of replicate samples early in your experiment covering the most varied of the independent variables, you gain the following:

- Familiarization with the metabolomics workflow
- Familiarization with the Agilent tools
- Partial answers to the hypothesis

Experience from early evaluations provides you with feedback that improves use of your time, resources, and funds. By collecting and processing initial data that spans the range of independent variables, you may gain timely feedback regarding the quality and direction of the results. If either the quality or the results are not meeting your expectations you may adjust the remainder of the experiment and/or sampling.

The grape wine example following illustrates this element of designing the experiment.

Example of planning for sampling and analysis

In the grape wine example the hypothesis involves two independent variables: (1) county of origin and (2) wine variety. In the Agilent LC/MS example data set related to identifying red wines, the independent variable "Variety" has three parameters and the independent variable "Country of Origin" has twelve parameters. This data set is shown in Figure 12.

An ideal metabolomics analysis of the wine data involves 360 samples (3 varieties multiplied by 12 countries of origin multiplied by 10 replicates per permutation). Thirty (30) samples would be expected for a particular country of origin to be properly represented; ten samples of each wine variety from each country. However, practical limitations resulted in this example data set being limited to 97 samples. In the data table shown in Figure 12, France has the potential for the best statistical representation.

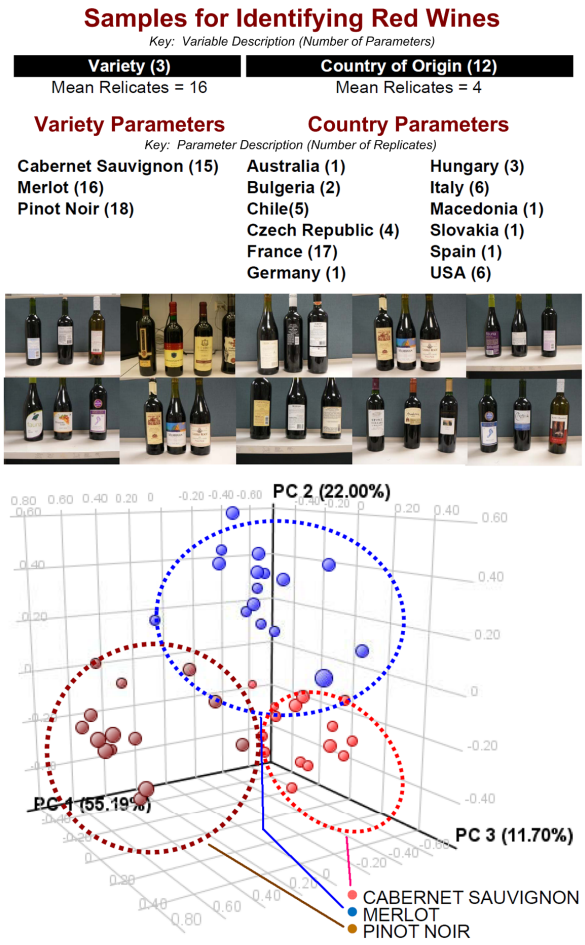


Figure 12 *Red wine samples and the resulting PCA analysis where the number of molecular features was reduced from 20,506 to 26 using metabolomics statistical treatments provided by Mass Profiler Professional.*

A further reflection on the data set shows that the sampling is more ideally suited for an initial analysis of the question "Can grape wine be correlated to variety?" If the answer to this question appears favorable at this early stage of sampling it may then be used to support continuing the sampling to answer both "Can grape wine be

correlated to the country of origin?" and "Can grape wine be correlated to variety?" with the appropriate sample size of at least 360 samples.

Summary

Improved data quality comes from matching the sampling methodology to the experiment design. Replicate data are collected that span the range of the parameters for the independent variables. Larger samples improve the statistics and improve the accuracy of your answer to the hypothesis. An understanding of the methodologies used in sampling and using more than one method of sample collection have a positive impact on the significance of your results.

Next step...

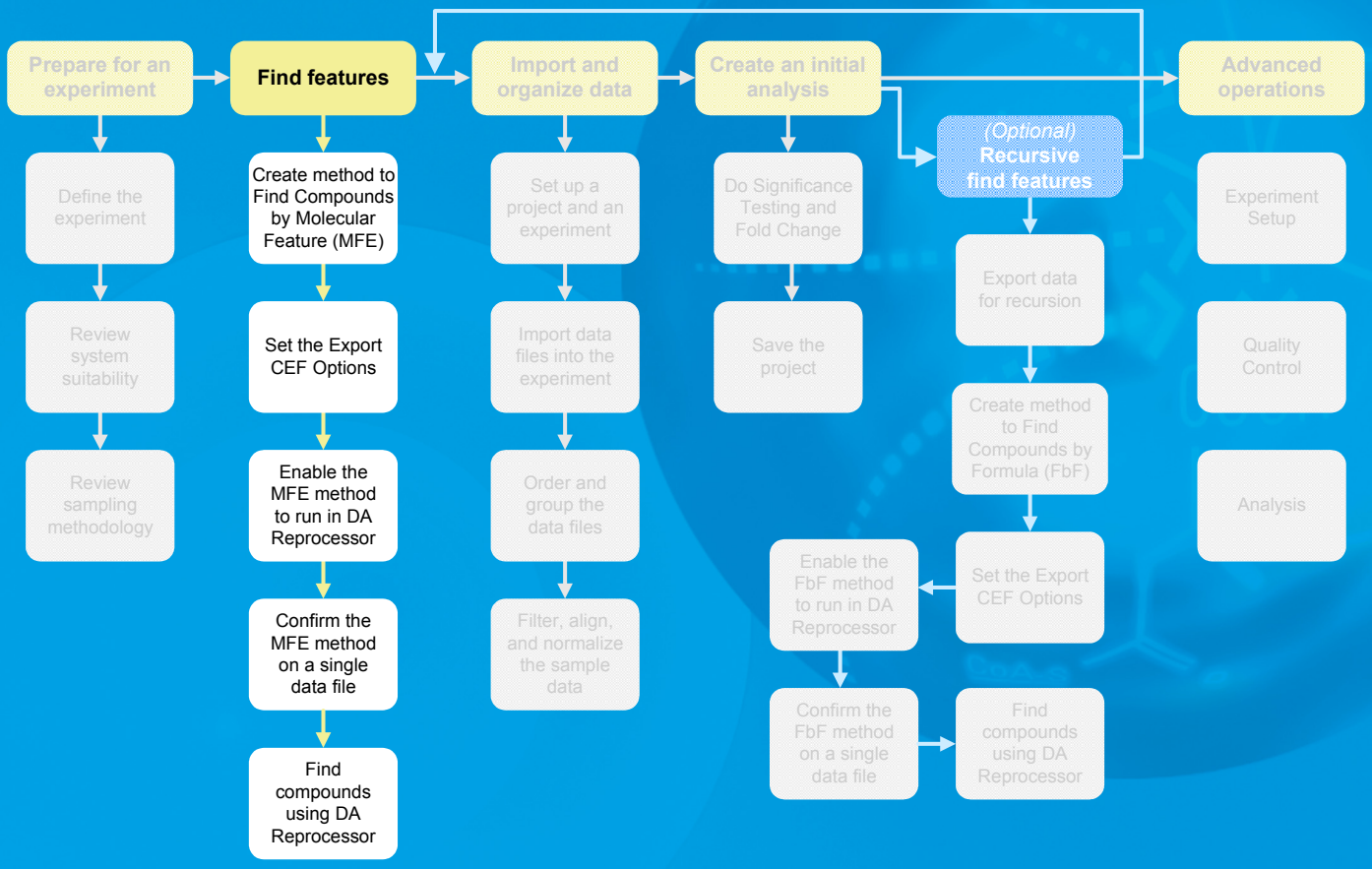
Apply the experiment definition process and methodologies to help you define and perform a metabolomics experiment:

1. Ask a question. By evaluating questions, and their relationship with the available analytical approach to solve the problem, you may identify the key variables, the independent variables, and the dependent variables.
2. Formulate a hypothesis that may be answered by analyzing the data for potential correlations among the independent variables, as well as correlations between the independent and dependent variables.
3. Assess the chemical makeup. The samples taken from specimen in your experiment should contain many chemical constituents (metabolites) that are naturally expressed through complex relationships by the specimen.
4. Identify your variables. Independent variables are the specific and known perturbations that may be made to a specimen, or quantifiable externalities that affect the specimen, and which are expected to induce a measurable change in the metabolic profile. Dependent variables are always the chemical constituents or the metabolic profile.
5. Verify that metabolomics simplifies your analysis. If your hypothesis is best proved or disproved using a correlative analysis on a large number of chemicals or metabolites the metabolomics process simplifies your analysis.
6. Design your experimental plan including controls and sampling. Determine how many replicate samples to use, the parameters and statistical analyses that are the most effective for proving or disproving your hypothesis, the sampling methodologies to employ, and how to minimize variations due to system drift over the time required to perform your experiment.
7. Execute your plan. While you are executing your experiment perform at least one early analysis on a set of replicate samples to assess the quality of your data, to become familiar with your data analysis, and develop a feel of the directions that the results are suggesting.

You have now completed the first step of the metabolomics workflow. In the next workflow step you begin organizing and analyzing your data.

Find features

Compounds, referred to as molecular features, are extracted from your data based on mass spectral and chromatographic characteristics. The process is referred to as Molecular Feature Extraction (MFE). Molecular feature extraction quickly and automatically generates a complete, accurate list of your compounds which include molecular weight, retention time, m/z, and abundance.



Agilent Technologies

Start Agilent MassHunter Qualitative Analysis

The following examples use Agilent MassHunter Qualitative Analysis B.05.00 running on 64-bit Windows 7 Professional.

User Interface Note: When you make a change to a parameter in MassHunter Qualitative Analysis, the software automatically places a change icon  (a blue triangle shape) in the Method Editor tab and next to the field containing the changed parameter. This icon indicates that you have unsaved changes in your method and helps you remember to save the changes you have made to the method. The original parameter value may be viewed by placing your pointer over the change icon. When you save your method the change icons disappears.

---

1. Start MassHunter Qualitative Analysis software.

MassHunter Qualitative Analysis is the software tool used to perform the function of finding molecular features in the raw and CEF data files. After the molecular features are found they are imported into Mass Profiler Professional for statistical analysis. Feature finding is an essential prerequisite to using Mass Profiler Professional.

Note: On-line help is available at any time within MassHunter Qualitative Analysis by pressing the **F1** key on the keyboard. The information presented specifically relates to the software fields and options available in the active display.

a Double-click the Qualitative Analysis icon  located on the desktop,

or (for Qualitative Analysis version B.05.00 or later on Windows 7)

Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.05.00**,

or (for Qualitative Analysis version B.03.01 on Windows XP)

Click **Start > Programs > Agilent > MassHunter Workstation > Qualitative Analysis.**

b Click **Cancel** in the **Open Data File** dialog box to start MassHunter Qualitative Analysis without opening any data files. To open data files later click **File > Open Data File.**

You do not need to open a data file at this time. You are prompted to open a data file in "Confirm the MFE method on a single data file" on page 53.

If you prefer to open one or more data files continue below, otherwise skip to the next step.

1. Select the data file or data files to open in the **Open Data File** dialog box.
   - An individual data file is selected by using a single click.
   - Select a continuous range of files with a click on a first file and Shift-click on a last file that includes the range of files you want to select.
   - Select discontinuous, individual files with a Ctrl-click on any file.
2. Click **Open** to start MassHunter Qualitative Analysis with the selected data file or data files.

---

2. Enable advanced parameters in the user interface.

Advanced parameters must be enabled in MassHunter Qualitative Analysis in order to show tabs labeled Advanced in the Method Editor and to enable compound importing for recursive finding of molecular features.

a Click **Configuration > User Interface Configuration**.

b Mark the **Show advanced parameters** check box under the Other group heading. See Figure 13.

If the files intended to be processed include GC/MS data, mark the **GC** check box under the Separation types group heading. If your analyzer is a quadrupole mark the **Unit Mass (Q, QQQ)** check box under the Mass accuracy group heading.

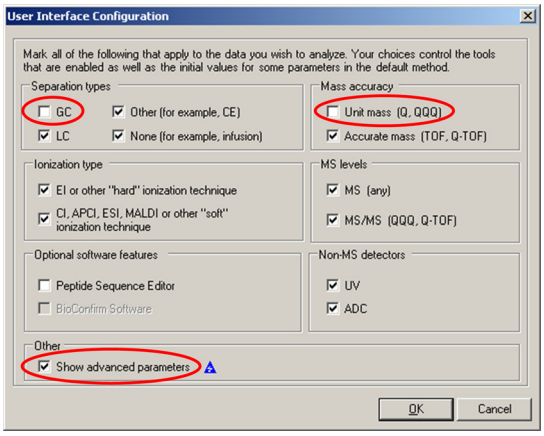


Figure 13 *MassHunter Qualitative Analysis user interface configuration dialog box*

c Click **OK**.

d Check to make sure that **File > Import Compound** is an available command. See Figure 14. This command is necessary to review CEF files before importing them into Mass Profiler Professional.

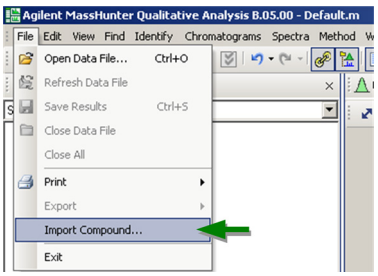


Figure 14 *MassHunter Qualitative Analysis import compound command location*

Create a method to Find Compounds by Molecular Feature

Find compounds by molecular feature is commonly referred to as molecular feature extraction (MFE). Molecular feature extraction involves chromatographic deconvolution as described in "Capabilities of the metabolomics workflow" on page 18 (see Figure 6 on page 19). Molecular feature extraction automatically finds related co-eluting ions, sums the related ion signals into single values, creates compound spectra, and reports results as a molecular feature, or compound.

Co-eluting ions: Molecular feature extraction finds co-eluting ions related to the same compound and creates an extracted compound chromatogram (ECC) including isotopes (13C, 15N, 2H, 18O), adducts (most commonly H+, Na+, K+ for positive ions and H- for negative ions), and dimers such as (2M + H)+.

Compound spectra: After creating extracted compound chromatograms, molecular feature extraction generates individual compound spectra for each molecular feature based on the co-eluting ions present.

Volume: The area of the ECC. The ECC is formed from the sum of the individual ion abundances within the compound spectrum at each retention time in the specified time window. The compound volume generated by molecular feature extraction is used by Mass Profiler Professional to make quantitative comparisons.

Composite spectrum: A compound spectrum that contains more than on co-eluting ion, more than just the (M+H) ion, within the molecular feature and is used by Mass Profiler Professional for recursive analysis and by ID Browser for compound identification.

Results: Each molecular feature, or compound, is uniquely identified by retention time, neutral mass, volume, and composite spectrum. An example of the relationships between some of the molecular features and the TIC is shown in Figure 15.

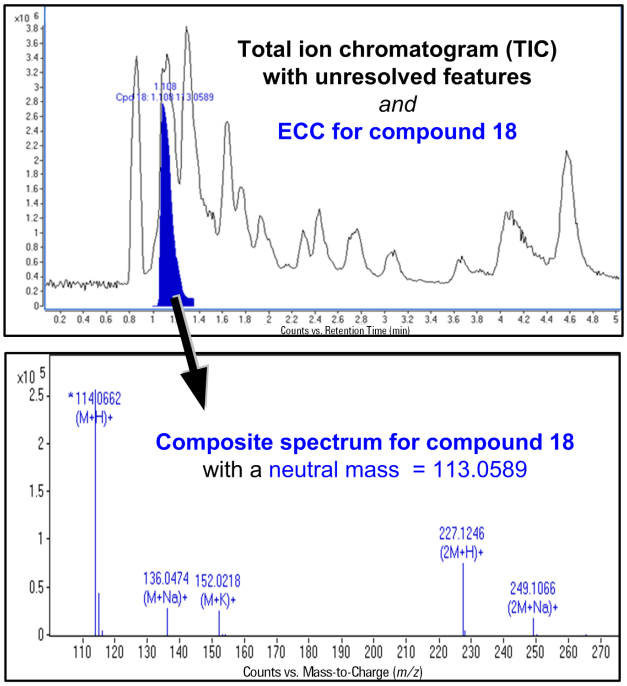


Figure 15 *Example of the relationship of some spectral components*

1. Open the Method Editor window for finding compounds by molecular feature.

a Open the **Method Editor: Find Compounds by Molecular Feature** section in the Method Editor window.

1. Click **Find Compounds** from within the **Method Explorer** window.
2. Click **Find by Molecular Feature**.

- All of the parameters involved in molecular feature extraction are accessed in the tabs presented in the **Method Editor: Find Compounds by Molecular Feature** section in the Method Editor window.
- To run molecular feature extraction after the parameters are entered, click **Find > Find Compounds by Molecular Feature** or click the Find Compounds by Molecular Feature button  in the **Method Editor: Find Compounds by Molecular Feature** section in the Method Editor window. Molecular feature extraction begins immediately and the progress is shown in an Operation in Progress status box.
- For more information, see the *Agilent MassHunter Workstation Software Qualitative Analysis - Familiarization Guide* (Agilent publication G3336-90018, Revision A, September 2011).

Note: Use the Extraction, Ion Species, and Charge State tabs to enter your parameters that control compound finding. Use the remaining tabs to enter parameters to filter the results and display the graphics.

2. Enter parameters for the tabs that control compound finding.

After the first time molecular feature extraction is run, any subsequent parameter changes you make within the "tabs that do not affect compound finding" (Figure 16) reprocess the data much more quickly because the find features algorithm is not repeated. The improved speed for reprocessing the molecular features lets you review several combinations of parameters to find the results that best suit your experiment.

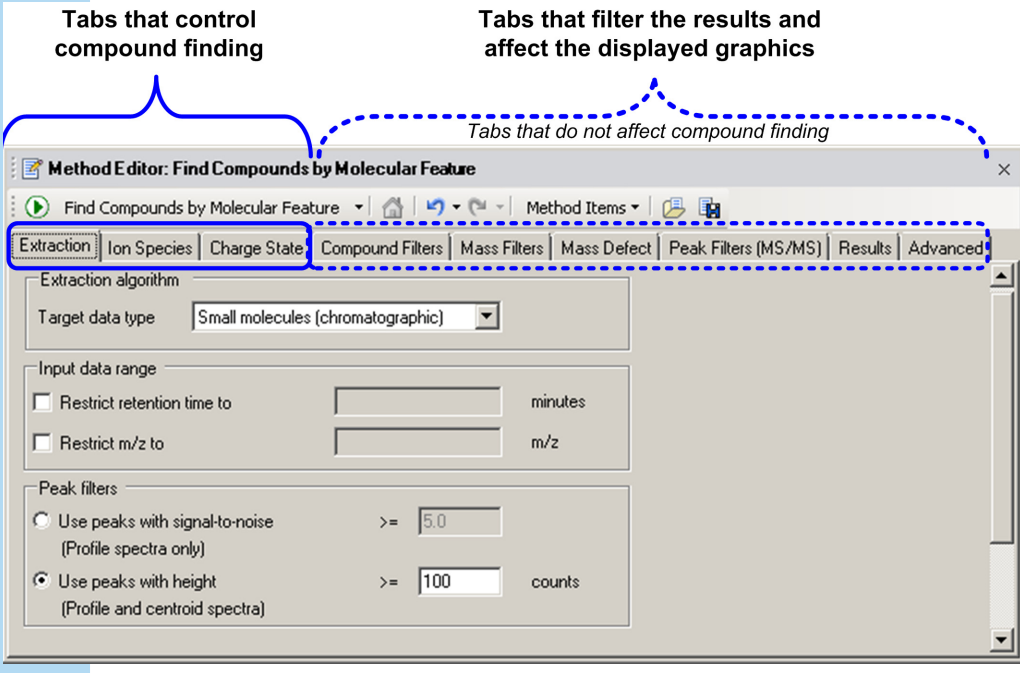


Figure 16 *Overview of the Find Compounds by Molecular Feature Method Editor tabs*

Extraction tab

a Edit the parameters on the **Extraction** tab.

The parameters in this tab let you specify features of the source data that enable the molecular feature extraction algorithm to perform more efficiently.

1. Click the **Extraction** tab.
2. Select **Small molecules (chromatographic)** in the **Target data type** box for working with metabolomic data.

   Note: The data must be collected in profile mode for the **Small molecules (infusion)** target data type. For **Large molecules (proteins, oligos)** the data must be collected in centroid mode or both modes.

Molecular feature extraction starts the data reduction process by creating a copy of the data file using the centroid of all of the ions. If you collect data in profile mode, it saves processing time if you also collect the data with centroid mode turned on. For all other data collection methods, it is recommended to save the data with centroid data.

3. Clear all of the check boxes under the Input data range group box.

Note: Marking the options under the Input data range group box is not necessary. Using **Restrict retention time to** and **Restrict m/z to** limits the location where the molecular feature extraction searches for features. It is recommended to let the molecular feature extraction algorithm find all of the features and then to use the filter parameters available in the "tabs that do not affect compound finding" to remove unwanted features.

4. Click **Use peaks with height** and type `300` for the counts. The counts value you enter represents a signal level at and above which actual ion signals are observed. 300 is typical if the background noise is approximately 100 counts.

   Note: The target **Use peaks with height** counts is three times (3x) the electronic noise in the mass spectrum, the signal measured by the detector that is not due to actual ions. The electronic noise is found by viewing the background signal level at the higher *m/z* range (around 1,000 *m/z*) of a single mass spectrum in the data set. Do not use an averaged or background subtracted mass spectrum. If the **Use peaks with height** value is set to a value too small, Find Compounds by Molecular Feature takes a very long time to run and finds features that are very small. If the **Use peaks with height** value is set to a value too large, then actual features may not be found.

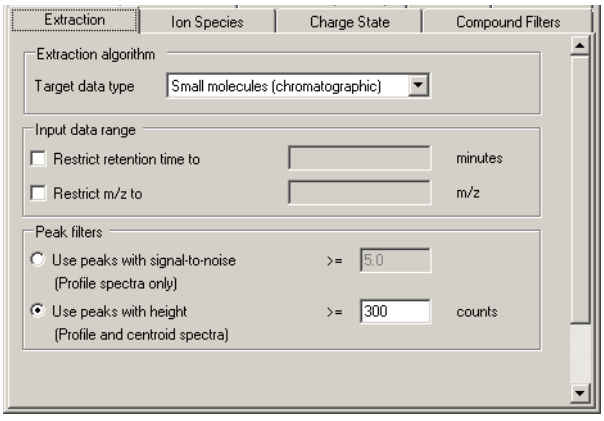


Figure 17 *Parameter values for the MFE Extraction tab*

Ion Species tab

b Edit the parameters on the **Ion Species** tab.

The parameters in this tab let you specify the ion adducts that the Find Compounds by Molecular Feature algorithm considers during the process of identifying molecular features.

1. Click the **Ion Species** tab.
2. Mark **+H**, **+Na**, and **+K** for positive ions and **-H** for negative ions.
3. Mark common **Neutral losses** if your mass spectrometer system is very energetic and thereby induces known neutral losses from the molecular ion.

   Note: Removal of possible ion adducts from the **Allowed ion species** group box increases the molecular feature extraction efficiency. Glass bottles leach sodium into the liquid introduction system. Thus, it is recommended to change the solvent and sample delivery bottles to bottles made from PTFE in place of bottles made from glass to reduce the background sodium levels.

   Note: Molecular feature extraction requires that the molecular feature involves at least the addition or loss of a proton unless or salt adduct.

4. Mark the **Salt dominated positive ions (M+H may be weak or missing)** check box to direct the molecular feature extraction algorithm to reduce the emphasis on identifying protonated ions (M+H) because they may be weak or missing in your data. For example, mark this check box when detecting sugars that are sodium adducted.

   Clear this check box to direct the molecular feature extraction algorithm to place a uniform weight across the allowed ion species in calculating the molecular weight for each feature.

   Note: If your sample contains an ion species that is not an available option in your method, add the ion species in the appropriate charge or neutral column.

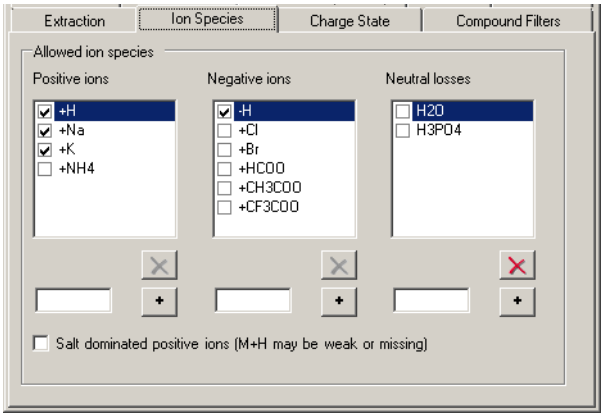


Figure 18 *Parameter values for the MFE Ion Species tab*

Charge State tab

c Edit the parameters on the **Charge State** tab.

The parameters on this tab let you set limits on the allowable ion charge states, and let you control how isotopes are identified and assigned to groups associated with each feature.

1. Click the **Charge State** tab.

2. Type `0.0025` for *m/z* and `7.0` for ppm into **Peak spacing tolerance**. These values are the tolerance that the molecular feature extraction algorithm uses to find isotope ions associated with each feature.
3. Select **Common organic molecules** for the **Isotope model**. For most metabolomics analysis, the **Common organic molecules** isotope model properly groups ions into the appropriate isotope clusters. Proper isotope clustering leads to an accurate assignment of charge state and mass for the molecular feature.

   Select **Unbiased** if metal containing molecules are expected.

   Note: Selecting **Unbiased** slows the molecular feature extraction calculations considerably because all isotope models are considered.

4. Mark the **Limit assigned charge states to a maximum of** check box and type `1`. You only type `2` if you have a very specific reason; otherwise, `1` is used for metabolomics. Increasing the value increases the chance of unwanted isotope grouping.
5. Clear the **Treat ions with unassigned charge as singly-charged** check box. When this check box is cleared, ions that cannot be assigned a charge state by the molecular feature extraction algorithm are ignored.

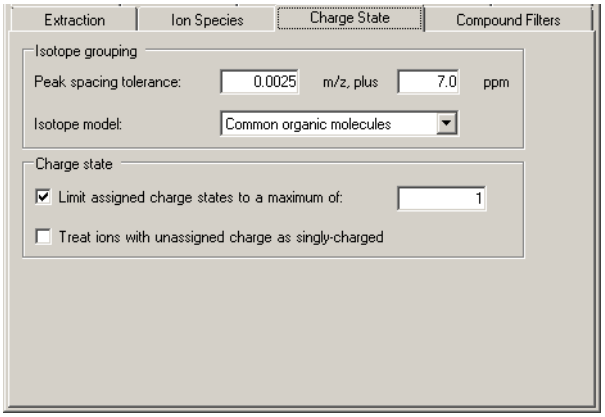


Figure 19 *Parameter values for the MFE Charge State tab*

3. Enter parameters for the tabs that filter results or affect the displayed graphics.

Compound Filters tab

a Edit the parameters on the **Compound Filters** tab.

1. Click the **Compound Filters** tab.
2. Clear the **Relative height** check box.
3. Mark the **Absolute height** check box and type in a value of `5000` counts.

   Note: In the final compound spectrum there must be at least one ion that is greater than or equal to the counts specified. The value typed is determined by empirically reviewing your mass spectral data. The absolute height in counts is different from the volume used to quantify the compound as a feature.

   Note: It is not recommended to mark **Relative height** or **Limit to the largest** when performing metabolomics.

4. Clear the **Restrict retention times to** check box. Only mark this parameter and type in the time range if you know the chromatographic void volume, solvent peak, or other region containing unwanted peaks in the data set.

5. Clear the **Restrict charge states** check box because the charge state was previously limited to `1` in the **Charge State** tab. If a larger number of charge states is allowed, then mark this parameter and enter a charge state value to filter the results.

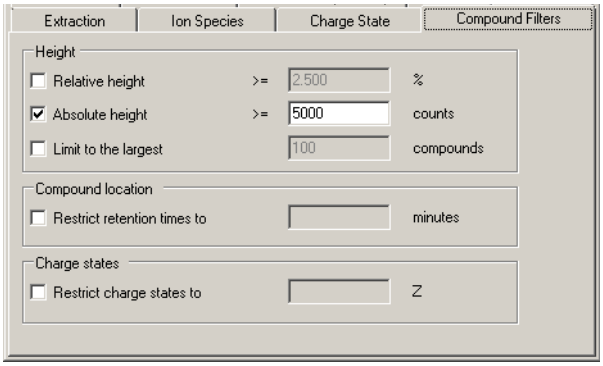


Figure 20 *Parameter values for the MFE Compound Filters tab*

Mass Filters tab

b Edit the parameters on the **Mass Filters** tab.

The parameters on this tab let you remove noise due to specific ions from the data without regard for retention time. Mass Profiler Professional is the preferred place to perform mass filtering. This filter feature can be unmarked and performed in Mass Profiler Professional more effectively with the addition of retention time.

1. Click the **Mass Filters** tab.
2. Clear the **Filter mass list** check box unless a specific list of neutral masses is known to be present in the data set that you wish to remove.
3. Select **Exclude these mass(es)** if you marked the **Filter mass list** check box.

   Note: It is not recommended to select **Include only theses mass(es)**.

4. If **Filter mass list** is marked, click the appropriate button in the Source of masses box indicating your source of the masses for the exclude filter. Two example masses to exclude are `120.0434` and `921.0013`.

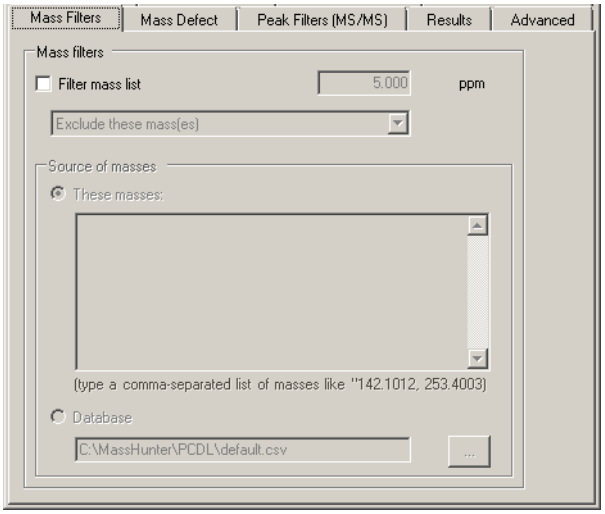


Figure 21 *Parameter values for the MFE Mass Filters tab*

Mass Defect tab

c Edit the parameters on the **Mass Defect** tab.

The parameters on this tab let you supply a range for the mass defect within which the identified mass may still be a metabolite. Since the range necessary for filtering by mass defect must be rather large it is not recommended to filter metabolomics results by mass defect.

1. Click the **Mass Defect** tab.
2. Clear the **Filter results on mass defects** check box.
3. If the **Filter results on mass defects** check box is marked, it is recommended to select **Variable** in the Expected mass defect box and type in values that work with your data set. If you select Variable, then the natural mass defect range increases with increasing mass.

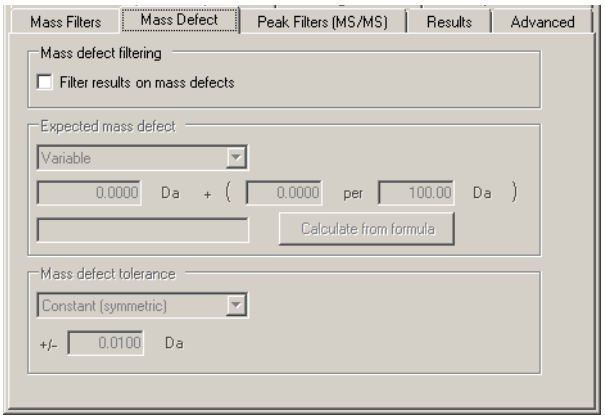


Figure 22 *Parameter values for the MFE Mass Defect tab*

Peak Filters (MS/MS) tab

d Edit the parameters on the **Peak Filters (MS/MS)** tab.

The parameters on this tab let you filter ions by height and quantity.

1. Click the **Peak Filters (MS/MS)** tab.
2. Clear the **Absolute height** check box. If this parameter is marked, do not type a counts value that is less than `10`. A typical counts value is around `100`. The best value to use is determined empirically.
3. Mark the **Relative height** check box and type `1` for the **% of largest peak**. The best analytical information is found using ions with an intensity at least within a factor of 100 of the base peak.
4. Clear the **Limit (by height) to the largest** check box.

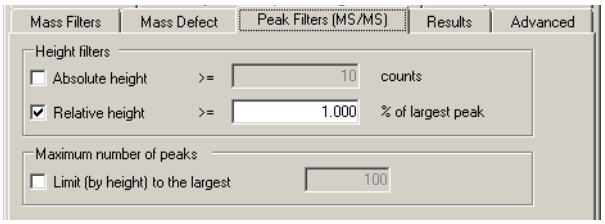


Figure 23 *Parameter values for the MFE Peak Filters (MS/MS) tab*

Results tab

e Edit the parameters on the **Results** tab.

The parameters on this tab let you customize the display of your results. To improve the speed of the extraction, do not draw graphics when running molecu-

lar feature extraction. If more information about a feature is desired it may be selectively obtained later instead of being generated for all of the features.

1. Click the **Results** tab.
2. Mark the **Delete previous compounds** check box.
3. Click **Highlight first compound**.
4. Clear all of the check boxes in the Chromatograms and spectra group box.
5. Clear the **Display only the largest** check box if you intend to create a CEF file.

Note: You must create a CEF file in order to import your molecular features into Mass Profiler Professional for the next step of the metabolomics workflow.

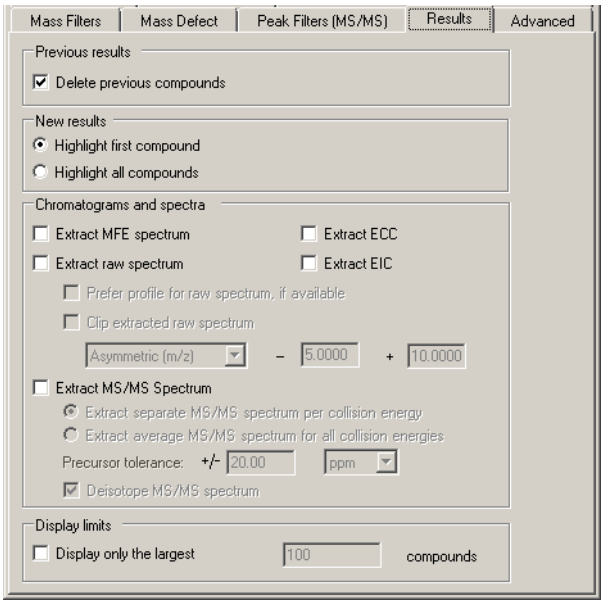


Figure 24 *Parameter values for the MFE Results tab*

Advanced tab

f Edit the parameters on the **Advanced** tab.

The parameters on this tab let you filter features by ion count and indeterminate neutral mass.

1. Click the **Advanced** tab.
2. Click **Include all** under the Compound ion count threshold group box. Filtering by two or more ions is a very useful feature, but it can filter out valid ions with small molecular weights.
3. Click **Exclude** under the Compounds with indeterminate neutral mass group box, especially when you click **Include all** under the Compound ion count threshold group box. Exclude disregards features to which molecular feature extraction has not been able to assign a neutral mass.

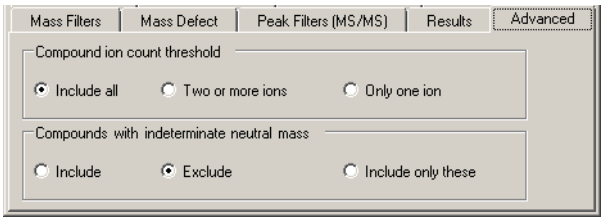


Figure 25 *Parameter values for the MFE Advanced tab*

Save your Find Compounds by Molecular Feature method

Save your method.

After you have edited your method to Find Compounds by Molecular Feature (MFE), it is recommended you save the method using a name that is readily distinguished from the name that is used later in this workflow for the method Find Compounds by Formula. Two distinct methods let you readily reprocess your data, or new data, without having to edit the workflow actions every time you switch between running MFE and FbF in the worklist.

---

a Click **Method > Save As**.

b Select the folder and type a method name in the **Save Method** dialog box. It is recommended to add the text `MFE` at the end of your file name to distinguish it from the file name that is recommended in "Save your Find Compounds by Formula method" on page 109.

c Click **Save**.

Set the Export CEF Options

1. Open the Method Editor for exporting CEF options.

2. Enter the export destination settings for your method.

Export CEF Options specifies where MassHunter DA Reprocessor stores the resulting .CEF feature files and whether the files replace or overwrite any prior files.

---

a Click **Export** from within the **Method Explorer** window.

b Click **CEF Options**.

---

a Click **At the location of the data file**.

b Click **Auto-generate new export file name**.

c Save your method. Click the save method icon or click **Method > Save**.

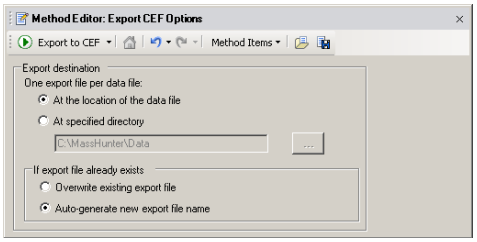


Figure 26 *Export CEF Options for use with DA Reprocessor*

Enable the method to run in MassHunter DA Reprocessor

1. Open the Method Editor to assign actions to run from the worklist.
2. Remove all actions from the **Actions to be run list**.
3. Add new actions to the **Actions to be run** list.

Agilent MassHunter software can most efficiently perform computationally intensive tasks, such as feature finding, on multiple data files using MassHunter DA Reprocessor. The following steps enable your method to run using DA Reprocessor.

---

a Click **Worklist Automation** from within the **Method Explorer** window.

b Click **Worklist Actions**.

---

a Double-click on an action in the **Actions to be run** list. The action is automatically removed from the **Actions to be run** list. As an alternate to the double-click, you can click on an action in the **Actions to be run** list and then click the delete icon  .

b Repeat action removal until the **Actions to be run** list is empty.

c Save your method. Click the save method icon  or click **Method > Save**.

---

a Double-click the **Find Compounds by Molecular Feature** action in the **Available actions** list.

The action is automatically added to the **Actions to be run** list. As an alternate to the double-click, you can click the action and then click the down arrow button  to add the action to the **Actions to be run** list.

b Double-click the **Export to CEF** action in the **Available actions** list. The **Export to CEF** action must be listed after the **Find Compounds by Molecular Feature** action as shown in Figure 27.

c Save your method. Click the save method icon  or click **Method > Save**.

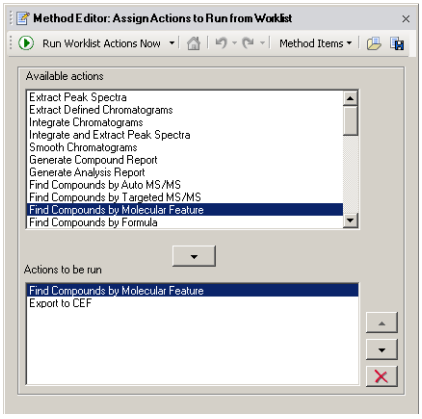


Figure 27 *Assign Actions to Run from Worklist for use with DA Reprocessor*

Confirm the MFE method on a single data file

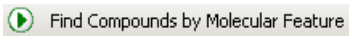
Metabolomics involves the analysis of a large number of sample files with each sample containing a large number of compounds. Find Compounds by Molecular Feature is therefore run on the entire metabolomics sample set using MassHunter DA Reprocessor. However, before the entire sample set is run in MassHunter DA Reprocessor, you can process a single file within MassHunter Qualitative Analysis to verify the new parameters.

1. Find Compounds by Molecular Feature for a single data file.

a Click **File > Open Data File**.

b Click on a single data file in the **Open Data File** dialog box.

c Click **Open**.

d Click **Actions > Find Compounds by Molecular Feature**, or click the Find Compounds by Molecular Feature button  in the **Method Editor: Find Compounds by Molecular Feature** section in the Method Editor window. Molecular feature extraction begins immediately and the progress is shown in an Operation in Progress status box as shown in Figure 28.

If no data file is open, or an inappropriate data file is open, a message box appears as shown in Figure 29. Click **OK** and open a single data file.

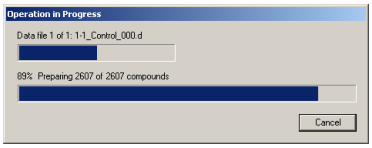


Figure 28 *Find Compounds by Molecular Feature progress box*

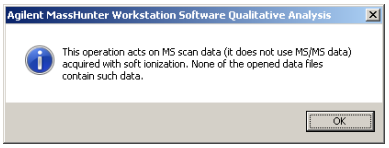


Figure 29 *Message box*

2. Display and review the Compound List.

When molecular feature extraction finishes processing the data file, the results are displayed in several windows within MassHunter Qualitative Analysis. The results may be reviewed and arranged to meet your preferences.

a Set up the recommended columns for viewing your data in the Compound List.

1. Right-click anywhere in the **Compound List** window.
2. Click **Add/Remove Columns** to open the **(Enhanced) Add/Remove Columns** dialog box.
3. Click **Clear All**.
4. Click the **Column Name** column header twice to sort the column names in ascending alphabetical order.
5. Mark the check boxes for at least the following Column Names: **Abund**, **Area**, **Base Peak**, **Cpd**, **File**, **Height**, **Ions**, **Mass**, **RT**, **Saturated**, **Show/Hide**, **Vol** and **Width**.
   - Each of these columns is documented in the MassHunter Qualitative Analysis Help in Reference > Columns > Compound List Table Columns under the Contents tab.
   - It is normal at this time for the Area and Abund columns to be blank.

- All of the columns are exported to the CEF file.

6. Click **OK**.

b Arrange the order of the columns and your compound data in the Compound List.

1. Click and drag the column names left or right so that they are arranged in the order you like. A useful order is shown in Figure 30.

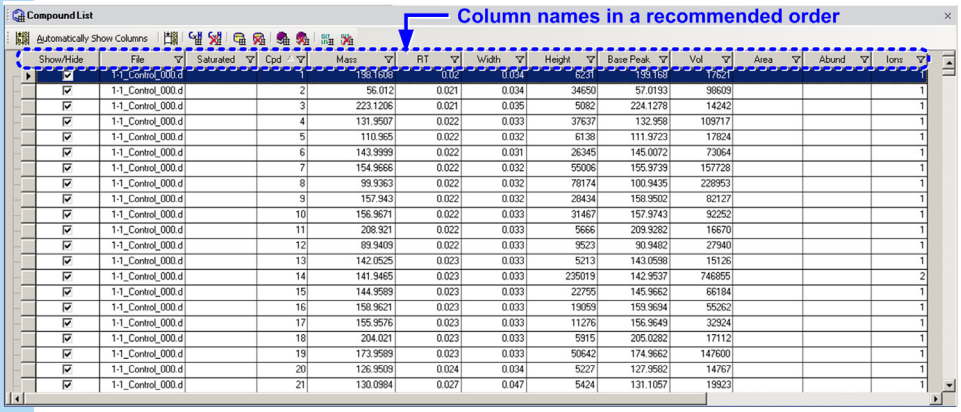


Figure 30 *Columns arranged in the Compound List window*

2. Click the **Height** column heading to sort the compounds by ascending height (the abundance value of the base peak). The compounds with a lower height value are shown at the top of the list.

3. Extract results.

a Select the compounds to view and compare chromatogram and MS results.

1. Click and drag across the first few compound rows to select multiple compounds (e.g., select around ten compounds). The selected compounds are highlighted.
2. Right-click the **Compound List** window and click **Extract Complete Result Set**.

The results are displayed in the **Chromatogram Results** and **MS Spectrum Results** windows. These windows are updated when you use the arrow keys to move up and down the Compound List. See Figure 31 on page 55.

- In the **Chromatogram Results** window, the extracted ion chromatogram (EIC) for each compound is compared to the extracted compound chromatogram (ECC) for the ions contained in the molecular features.
- In the **MS Spectrum Results** window, the compound spectrum for each compound is compared to the scan data spectrum.
- A compound may be deleted and removed from the features available for exporting by highlighting the compound and then pressing the delete key.

b If a significant number of compounds are too weak to provide confidence as a molecular feature, adjust the Find by Molecular Feature parameters. Weak compounds have small values for Height and Vol in the **Compound List** (see Figure 30) and have low count values in the **Chromatogram Results** and **MS Spectrum Results** (see Figure 31 on page 55).

1. Adjust the parameters entered in the "Compound Filters tab" on page 46.
2. Re-run **Run Find Compounds by Molecular Feature**. Molecular feature extraction runs very quickly if no changes are made to the tabs that control compound finding.

3. Review the new results by repeating step 2 - "Display and review the Compound List." on page 53.
4. Repeat these steps until you are satisfied with the molecular feature results.

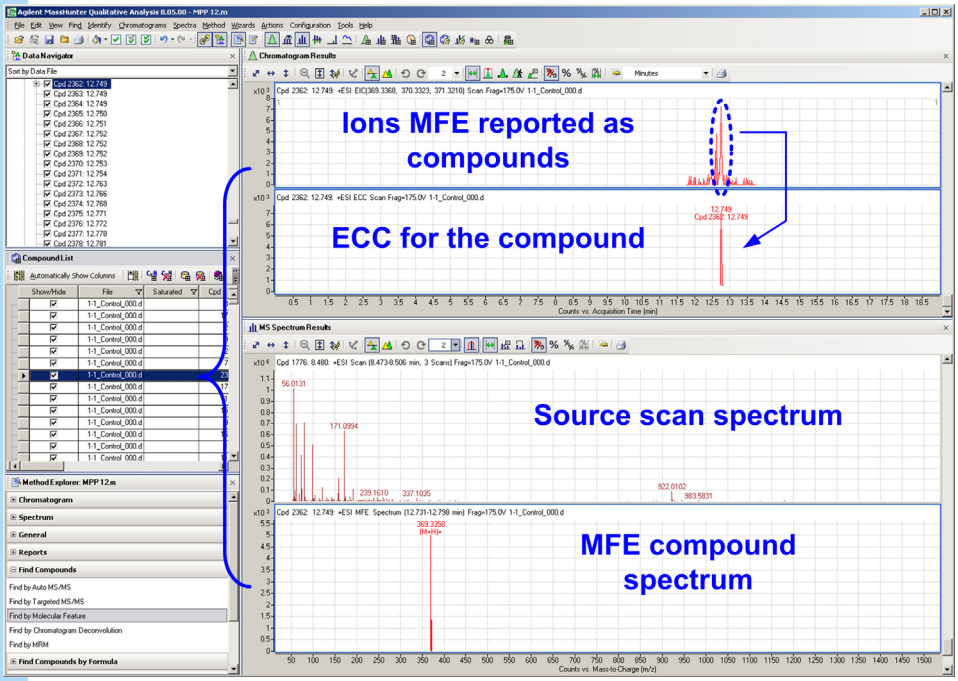


Figure 31 *Display of the Extract Complete Result Set from the Compound List*

4. *Optional* - Export the results for the single sample to a CEF file.

This step is optional. The CEF files for all of the samples are generated in "Find compounds using DA Reprocessor"

a Click **File > Export > as CEF**. The Export CEF Options dialog box is opened.

b Select the data files to be exported from the **List of opened data files**. It is recommended to create a new folder for the exported CEF files to aid documentation of the metabolomics workflow and to make it easier to distinguish any new CEF files from previous CEF files.

c Update the other parameters in the **Export CEF Options** dialog box.

d Click **OK**.

You can review the results from this step by importing the CEF back into MassHunter Qualitative Analysis by following step 3 - "Display and review the Compound List after running MassHunter DA Reprocessor." on page 57.

Find compounds using DA Reprocessor

Metabolomics involves applying your method to a large number of sample files, each of which may contain a large number of compounds. MassHunter Qualitative Analysis can be used to process all of your data sets. However, MassHunter DA Reprocessor provides a more efficient and automated means to run your MassHunter Qualitative Analysis method on multiple sample files. Therefore your method is run on the entire metabolomics sample set using DA Reprocessor.

1. Close your data file.

---

a Click **File > Close Data File**.

b Click **No**. Do not save the results.

---

2. Find Compounds by Molecular Feature using MassHunter DA Reprocessor.

a Click the DA Reprocessor icon located on the desktop, or click **Start > All Programs > Agilent > MassHunter Workstation > Acq Tools > DA Reprocessor**.

Press F1 from within MassHunter DA Reprocessor to start on-line help. For example, click **DA Reprocessor > Shortcut Menu for Worklist (from the top left cell)** for instructions on creating a worklist containing multiple samples.

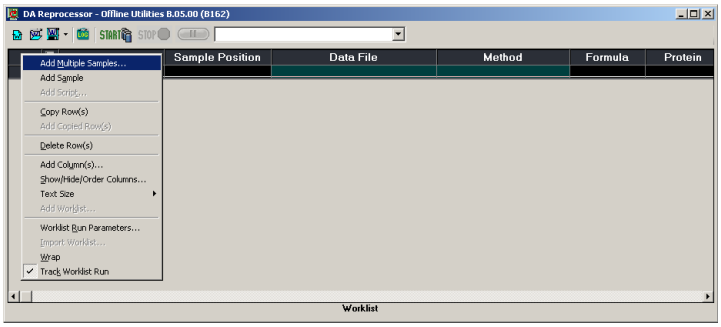


Figure 32 *Adding samples to the MassHunter DA Reprocessor worklist*

b Right-click the top left cell of the worklist and click **Add Multiple Samples** as shown in Figure 32.

c Select the folder and file names that refer to your samples.

d Click **Open**.

e Click the Method name for the first sample in row 1 and select the name of the method saved from MassHunter Qualitative Analysis. If the method you saved is not in the immediate list, select "Other" and then you can select the folder and method using the **Open File** dialog box as shown in Figure 33 on page 57.

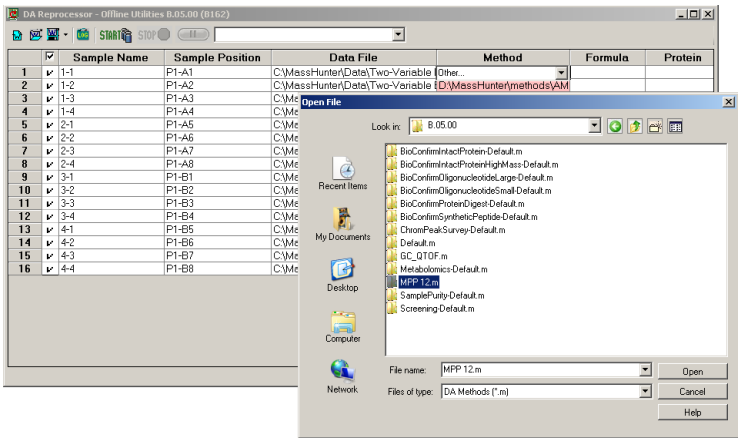


Figure 33 *Selecting the MassHunter Qualitative Analysis method*

f Copy the method from the first sample to each of the samples in the worklist; right-click on the method in row 1 and click **Fill > Column** (see Figure 34).

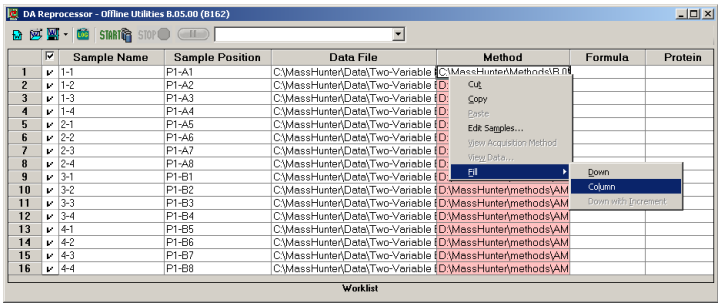


Figure 34 *Copying the data analysis method to each sample in the worklist*

g Click the **Start** icon in the toolbar to run the worklist. The progress is indicated on the worklist sheet as each sample is completed.

The CEF files containing the molecular features from the samples are automatically placed in the folder containing the .d data files. Each CEF file has the same root name as the sample data file. You import the CEF files into Mass Profiler Professional for feature selection in the next step of the metabolomics workflow.

---

3. Display and review the Compound List after running MassHunter DA Reprocessor.

a Return to MassHunter Qualitative Analysis. If you closed the MassHunter Qualitative Analysis program, do the following:

- Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.05.00**.
- Click **Cancel** when the **Open Data File** dialog box opens.

b Click **File > Close All** to close any open data files. Do not save any results.

c Click **File > Open Data File** to open one of the original sample data files including the chromatographic data and the results of Find Compounds by Molecular Feature. Mark the **Load result data** check box.

or

Click **File > Import Compound** to open one of the CEF files that contains the molecular feature results of Find Compounds by Molecular Feature.

Note: Because of the large number of features in a typical metabolomics sample file, it is recommended to open only one file at a time to review the results. Close the open file and then open the next file.

d Display and review the Compound List as described previously in step 2 - "Display and review the Compound List." on page 53. The chromatographic results are only visible if the original data file is opened.

4. Extract results.

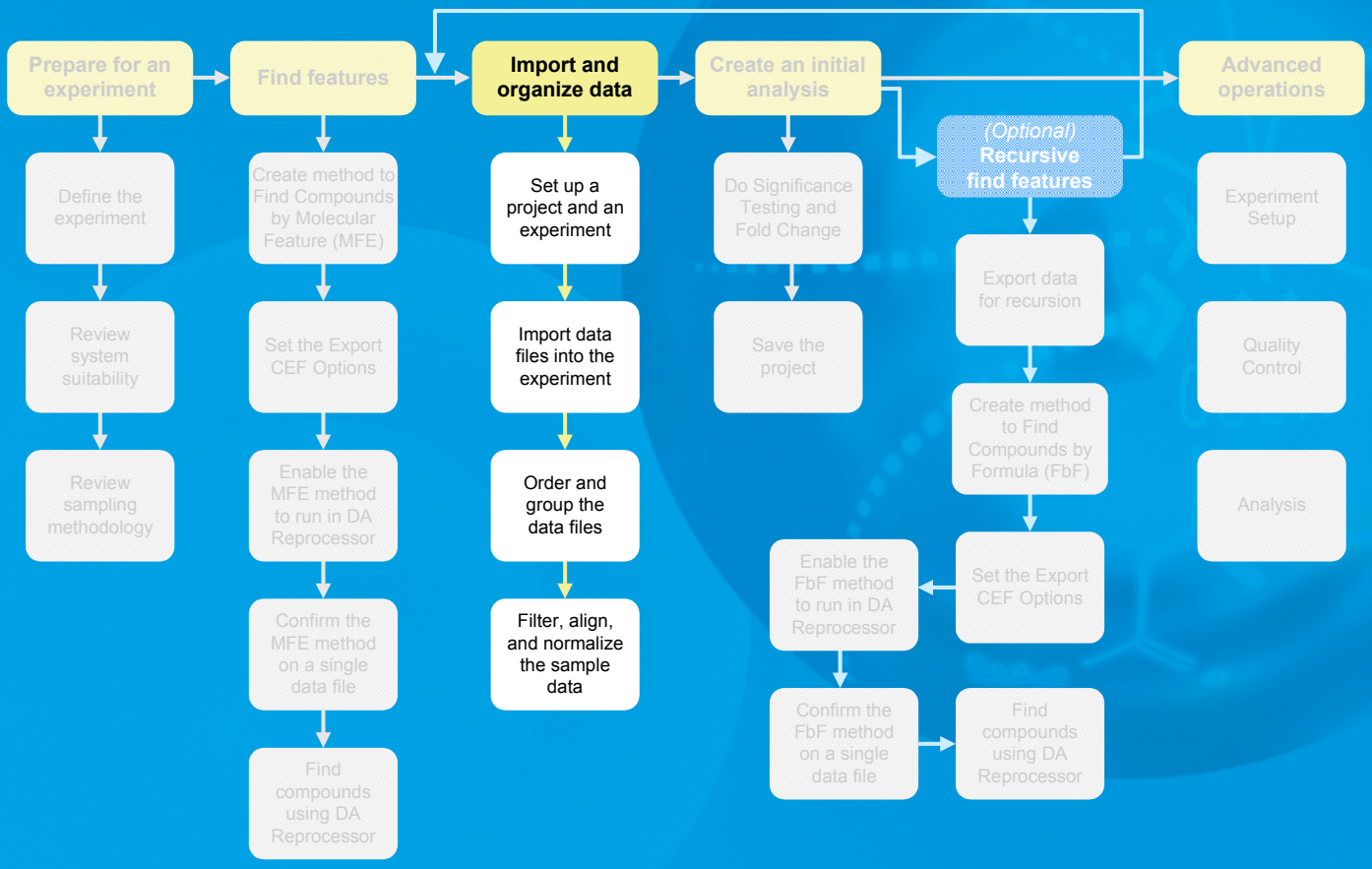
Extract your MS results as described previously in step 3 - "Extract results." on page 54.

Next step...

You have now completed the second step of the metabolomics workflow. In the next workflow step you import your Find Compounds by Molecular Feature (MFE) results into Mass Profiler Professional.

Import and organize data

After you create a project and an experiment, the "MS Experiment Creation Wizard" guides you through the necessary steps to organize your experiment, import your data, define your experimental variables, and prepare your data for analysis. The data preparation includes filtering, alignment, normalization, and baselining.



Agilent Technologies

Start Agilent Mass Profiler Professional

During the data import step of the metabolomics workflow, Mass Profiler Professional imports CEF files created from MassHunter DA Reprocessor, based on the method created in MassHunter Qualitative Analysis, and performs statistical and graphical analyses. Mass Profiler Professional exploits the high information content of chromatography/mass spectrometry data and can easily import, analyze, and visualize GC/MS, LC/MS, CE/MS, and ICP-MS data from large sample sets and complex MS data sets.

Because the advanced operations available in the Workflow Browser do not guide you through the initial steps of data import and differential analysis, it is not recommended to skip the "Import and organize data" or "Create an initial analysis" steps of the metabolomics workflow. All parameters, including the default parameters used during the MS Experiment Creation Wizard, can be edited at the conclusion of the Metabolomics Workflow by using the operations available in the Workflow Browser (see Figure 67 on page 92).

Note: To obtain help and detailed information regarding the various fields and statistical treatments press the F1 key on the keyboard or refer to the *Mass Profiler Professional User Manual*.

---

1. Start Mass Profiler Professional software.

The following examples use Agilent Mass Profiler Professional B.12.01 running on 64-bit Windows 7 Professional.

a Double-click the Mass Profiler Professional icon  located on the desktop, or click **Start > All Programs > Agilent > MassHunter Workstation > Mass Profiler Professional > Mass Profiler Professional**

Set up a project and an experiment

A project is a container for a collection of experiments. A project can have multiple experiments on different sample types and organisms. You are guided through four steps to create a new project and experiment to receive your imported data:

- **Startup:** Select creation of a new project.
- **Create New Project:** Type descriptive information about the project.
- **Experiment Selection Dialog:** Select create a new experiment as part of the project.
- **New Experiment:** Type and select custom information to store with the experiment.

1. Create a new project in the **Startup** dialog box.

a Click **Create new project**.

b Click **OK**.

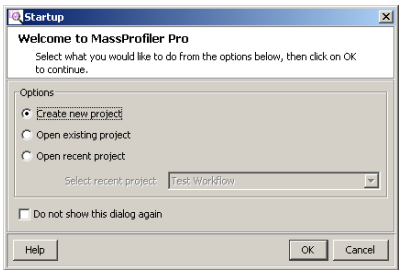


Figure 35 *Welcome to Mass Profiler Professional startup dialog box*

2. Enter descriptive information in the **Create New Project** dialog box.

a Type a descriptive **Name** for the project.

b Type descriptive **Notes** for the project.

c Click **OK**.

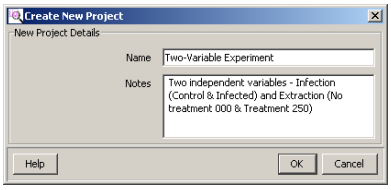


Figure 36 *Create New Project dialog box*

3. Select your experiment origin in the **Experiment Selection Dialog** dialog box.

Specify whether the wizard guides you through creating a new experiment or whether the wizard opens an existing experiment.

a Click **Create new experiment**.

b Click **OK**. If you clicked the **Open existing experiment** button, you are prompted for the experiment to add to the analysis.

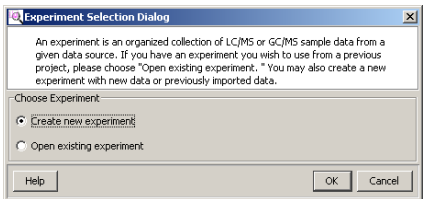


Figure 37 *Experiment Selection Dialog dialog box*

4. Type and select information that guides the experiment creation in the **New Experiment** dialog box.

Available entry options for the New Experiment dialog box depend on your experiment type and data sources as outlined in Table 1and Table 2 on page 63.

a Type a descriptive name for the experiment in **Experiment name**. This entry may be different from the project name previously entered.

b Select **Mass Profiler Professional** for the **Analysis type** to enable metabolomics and proteomics analyses. Only your licensed analysis types are available.

c Select **Unidentified** for the **Experiment type**. Unidentified is the proper selection when the compound features have only been identified by their neutral mass and retention time using molecular feature extraction. The experiment type selection determines how Mass Profiler Professional manages the data. Use Combined (Identified + Unidentified) when you are unsure if the data is identified in full or in part or when MassHunter Qualitative Analysis has been used previously to identify some of the compound features.

d Select **Analysis: Significance Testing and Fold Change** for the **Workflow type**.

When you select Analysis: Significance Testing and Fold Change the workflow still takes you through the MS Experiment Creation Wizard first.

Regardless of your personal expertise, the Analysis: Significance Testing and Fold Change workflow provides you with quality control to your analysis that improves your results. You may customize the entire analysis at the conclusion of the Analysis: Significance Testing and Fold Change workflow.

e Type descriptive notes for the experiment in the **Experiment notes**.

f Click **OK**.

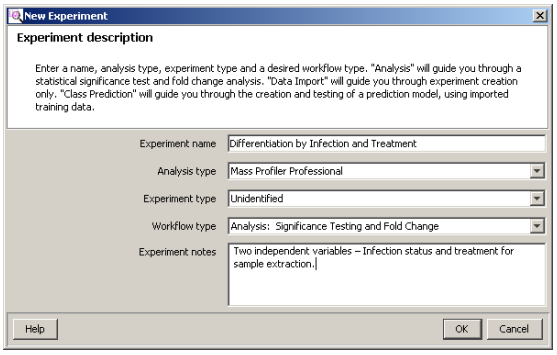


Figure 38 *Experiment description in the New Experiment dialog box*

Table 1 *Table of selections and entries for the New Experiment dialog box*

Dialog Box Field	Your Choices	Comments
Experiment name	<none>	Edit field to describe this experiment
Analysis type	Mass Profiler Professional <other choices depending on Order IDs>	"Mass Profiler Professional" must be selected
Experiment type	Combined (Identified and Unidentified) Identified Unidentified	<see next table>
Workflow type	Analysis: Significance Testing and Fold Change Class Prediction: Build and Test Model Data Import Wizard	
Experiment notes		Edit field to enter other experimental notes

Table 2 *Table of data sources and file extensions based on Experiment Type*

Experiment Type	Data Source	File Types	Comments
Identified	MH Quant		Compounds identified by MassHunter Quantitative Analysis
	Chemstation	*.FIN	Compounds identified by Chemstation Quantification or Screener processes
	MH Qual	*.CEF	Find by Formula
	MH Qual (GC Scan)	*.CEF	Identify by Unit Mass Library
	ICP-MS	*.CSV	Identified by ICP-MS software
	AMDIS	*.FIN	Compound identified by an AMDIS target library
	Generic	*.XLS *.XLSX *.CSV *.TXT	Entries identified by Compound (column C), Formula (column D), CASID (column E)
Unidentified	MH Qual	*.CEF	Find By Molecular Feature Extractor (MFE)
	MH Qual (GC Scan)	*.CEF	Find by Chromatographic Deconvolution
	ICP-MS	*.CSV	Identified by ICP-MS software
	AMDIS	*.ELU	Components identified by AMDIS that are not identified by an AMDIS target library
	Generic	*.XLS *.XLSX *.CSV *.TXT	Entries NOT identified by Compound (column C), Formula (column D), CASID (column E)
Combined	MH Qual	*.CEF	Find By Molecular Feature Extractor (MFE) and Find By Formula
	MH Qual (GC Scan)	*.CEF	Find by Chromatographic Deconvolution and Library Search
	ICP-MS	*.CSV	Identified by ICP-MS software
	AMDIS	*.FIN *.ELU	Targets and components discovered by AMDIS
	Generic	*.XLS *.XLSX *.CSV *.TXT	Combination of entries identified by and not identified by Compound (column C), Formula (column D), CASID (column E)

Overview of the MS Experiment Creation wizard

Importing and organizing your data consists of sequential steps that defines the experiment containing your samples (data files), interpretations, and associated entity lists. More than one experiment may be created within a project. Up to eleven steps are involved in the MS Experiment Creation Wizard. The steps you use with your experiment depend on your description and data source.

Importing your data and creating your experiment from the features found using MassHunter Qualitative Analysis involves only the steps presented below:

Step 1. Select Data Source: Select the data source that generated the molecular features you are using for your experiment.

Step 2. Select Data to Import: Select the molecular feature sample files.

Step 5. Sample Reordering: Organize your samples by selecting and deselecting individual samples and reordering the selection to group the samples based on the independent variables.

Step 6. Experiment Grouping: Define the sample grouping with respect to your independent variables, including the replicate structure of your experiment.

Step 7. Filtering: Filter the molecular features by abundance, mass range, number of ions per feature, and charge state.

Step 8. Alignment: Align the features across the samples based on tolerances established by retention time and mass. This step is omitted when the experiment type is "identified" because identified compounds are treated as aligned by identification.

Step 9. Sample Summary: Display a mass versus retention time plot, spreadsheet, and compound frequency for the distribution of aligned and unaligned entities in the samples. Compound Frequency charts provide a quick view into the effectiveness of the alignment of unidentified experiment types. The back and next buttons in the wizard let you easily review the effects of different alignment and filter options.

Step 10. Normalization Criteria: Scale the signal intensity of sample features to a value calculated by the specified algorithm or an external scalar.

Step 11. Baselining Options: Compare the signal intensity of each sample to a representative value calculated across all of the samples or the control samples.

Import data files into the experiment

Your data files are imported in to Mass Profiler Professional during Step 1 and Step 2 of the MS Experiment Creation Wizard.

1. Select the data source in the **MS Experiment Creation Wizard (Step 1 of 11)**.

a Click **MassHunter Qual**.

b Select the **Organism** represented by your samples. Selection of an organism is important if you plan to use pathways.

c Click **Next**.

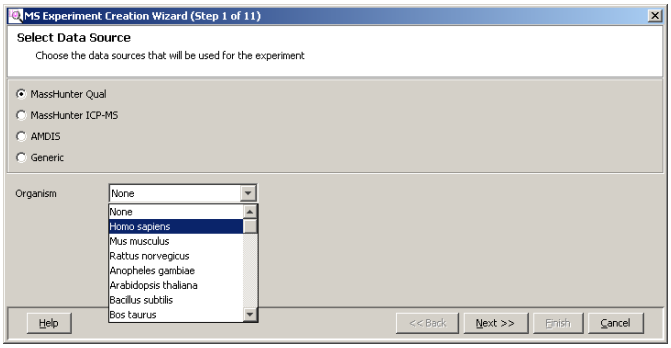


Figure 39 *Select Data Source page in the MS Experiment Creation Wizard*

2. Select the sample data to import in the **MS Experiment Creation Wizard (Step 2 of 11)**.

a Click **Select Data Files**.

b Select the data file or data files to open.

Note: Orderly naming of the data files with respect to the parameters related to the independent variables helps you make sure that all of the data is selected.

c Click **Open**.

d Click **Next**.

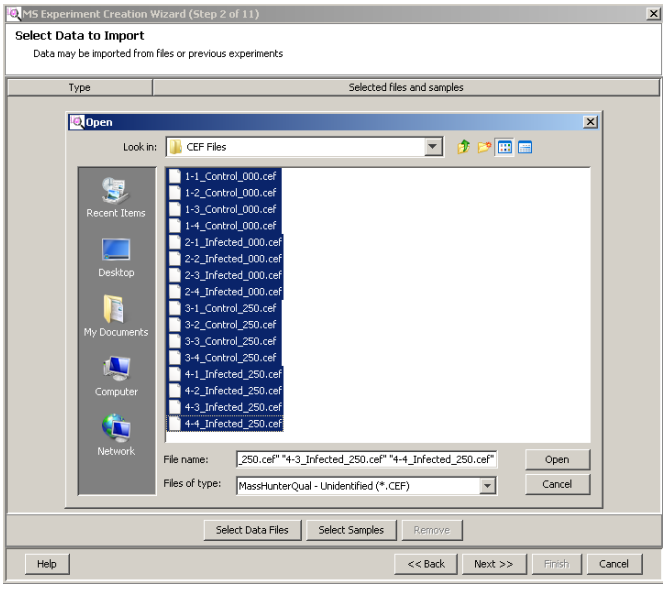


Figure 40 *Selection of several CEF files from MassHunter Qualitative Analysis*

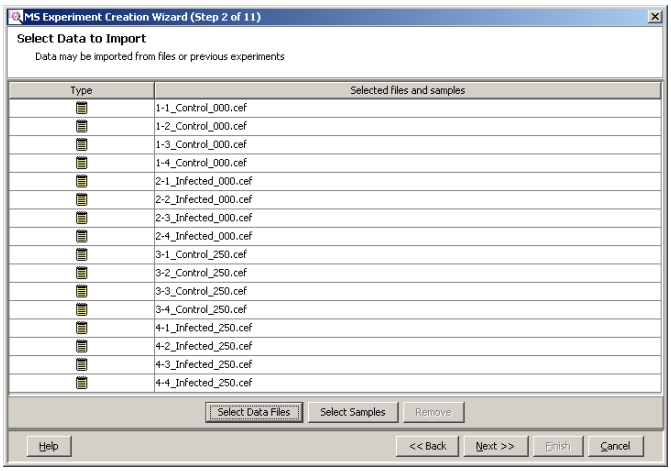


Figure 41 *Selected sample files ready for the next step*

Order and group the data files

1. Review and order the selected files that are imported in the **MS Experiment Creation Wizard (Step 5 of 11)**.

2. Group samples based on the independent variables and replicate structure of your experiment in the **MS Experiment Creation Wizard (Step 6 of 11)**.

Your data files are ordered in Step 5 and your experimental grouping is entered in Step 6 of the MS Experiment Creation Wizard.

This step presents the only opportunity you have to reorder your samples. After completing the MS Experiment Creation Wizard you create a new project and repeat this process to reorder your samples.

a Click one or more samples that you want to reorder. Selected sample rows are highlighted.

b Click the **Up** or **Down** buttons to reorder the selected sample or samples.

c Click the **Restore** button at any time to return the sample order to your starting point when this step was begun.

d Repeat the reordering steps as often as necessary to obtain your order.

e Mark the **Select** check box in the same row as the **Sample Name** for the samples to import for your analysis, or click **Select All**.

f Click **Next**.

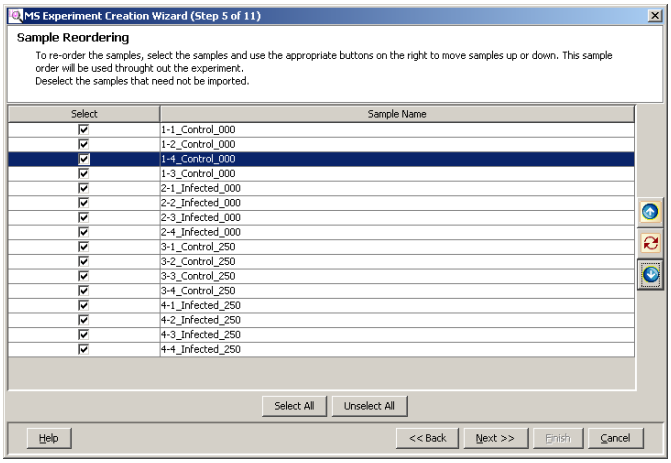


Figure 42 *Selection and reordering of the sample files*

Your sample grouping is determined by your experiment definition. Review the section "Define the experiment" on page 22 for an overview of a metabolomics experiment to help you enter your sample grouping. An independent variable is referred to as a parameter name. The attribute values within an independent variable are referred to as parameter values. Samples with the same parameter values within a parameter name are treated as replicates. In order to proceed, at least one parameter with two values must be assigned.

Only the first two parameter names (independent variables) are presented in the summary at the conclusion of the MS Experiment Creation Wizard. All parameters and values entered at this time can be edited during the Analysis: Significance Testing and Fold Change workflow and at the conclusion of the workflow by using operations available in the Workflow Browser.

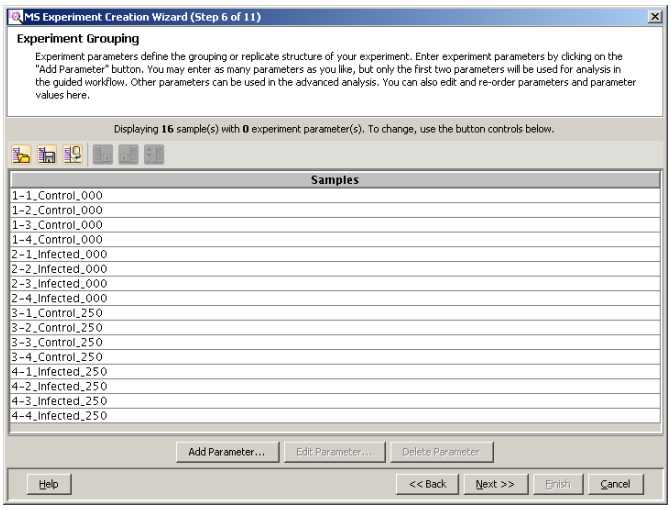


Figure 43 *Experimental Grouping dialog box*

Assign parameter values for the first, or only, independent variable

Note: When entering Parameter Names and parameter Assign Values, it is very important that the entries use identical letters, numbers, punctuation, and case in order for the Experiment Grouping to function properly. Click **Back** or **Experiment Setup > Experiment Grouping** to return to Experiment Grouping if an error is identified later in the Significance Testing and Fold Change workflow or when performing operations available in the Workflow Browser, respectively.

To apply previously saved experiment parameters and parameter values saved in a tab separated value (.tsv) file, click the **Load experiment parameters** button, or the **Import parameters from samples** button, and skip most of the following steps.

a Click **Add Parameter**. The Grouping of Samples dialog box is opened.

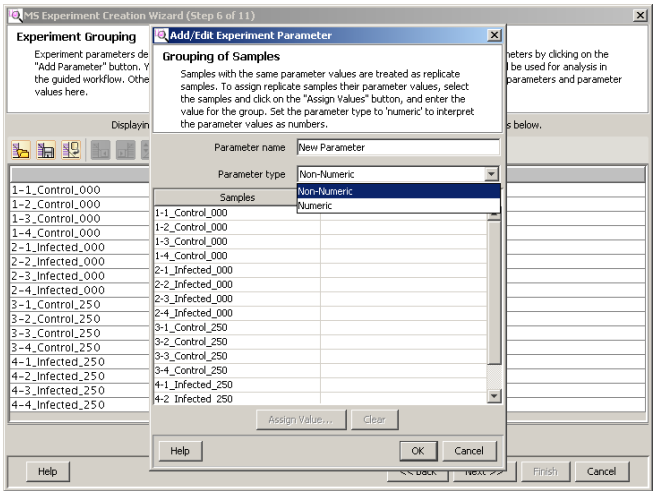


Figure 44 *Grouping of Samples dialog box*

b Type a brief, descriptive name for the first independent variable into the **Parameter name**. Type `Infection` for the two-variable experiment example.

c Select **Non-Numeric** for the **Parameter type** for your grouping when the grouping is not a quantitative value.

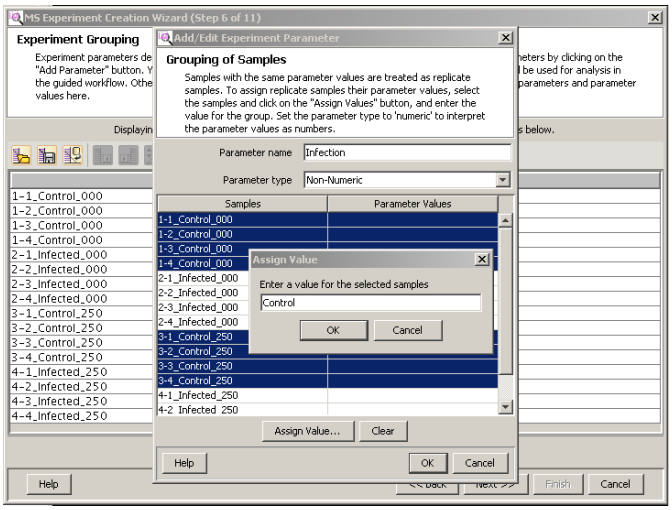


Figure 45 *Non-Numeric Parameter Type during Experiment Grouping*

d Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the samples that are part of the first attribute of the typed parameter name.

e Click **Assign Value** after the rows have been selected.

f Type `Control` in the Assign Value dialog box.

g Click **OK**.

Assign additional parameter values for the first, or only, independent variable

a Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the samples that are part of the next attribute value within the parameter name.

b Click **Assign Value** after the rows have been selected.

c Type `Infected` in the Assign Value dialog box.

d Click **OK**.

e Repeat the row selection and Assign Value process as necessary to complete the assignment of the samples to each of the attribute values for the independent variable.

f Click **OK** when all of the attribute values (Assign Value) are assigned to the current independent variable (Parameter Name).

Assign parameter values for the second independent variable

a Click **Add Parameter** to begin parameter assignment for the next independent variable. Otherwise skip to "Assign parameter values for the remaining independent variables" on page 70.

b Type a brief, descriptive name for the second independent variable into the **Parameter Name**. Type `Treatment` for the two-variable experiment example.

c Select **Numeric** for the **Parameter type** for your grouping when the values are quantitative or reflect a degree of proportionality among the samples with respect to the independent variable.

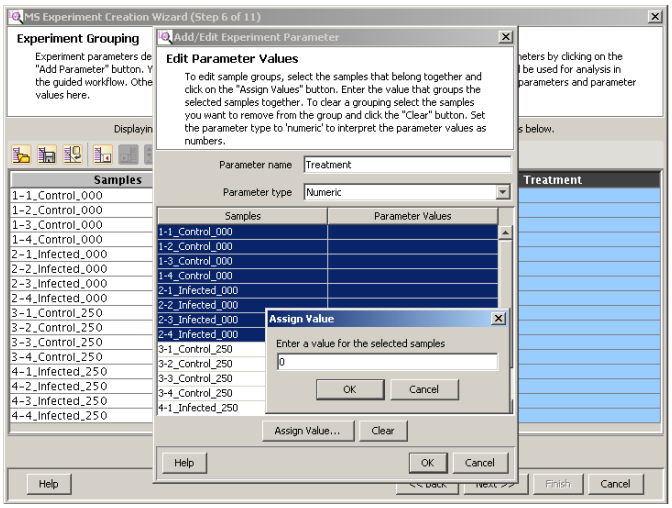


Figure 46 *Numeric Parameter Type during Experiment Grouping*

d Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the samples that are part of the first attribute of the typed parameter name.

e Click **Assign Value** after the rows have been selected.

f Type `0` in the Assign Value dialog box.

g Click **OK**.

Assign additional parameter values for the second independent variable

a Click the sample rows, while pressing the **Shift** or **Ctrl** key as necessary, to select the samples that are part of the next attribute value within the parameter name.

b Click **Assign Value** after the rows have been selected.

c Type `250` in the Assign Value dialog box.

d Click **OK**.

e Repeat the row selection and Assign Value process as necessary to complete the assignment of the samples to each of the attribute values for the independent variable.

f Click **OK** when all of the attribute values (Assign Value) are assigned to the current independent variable (Parameter Name).

Assign parameter values for the remaining independent variables

a Repeat "Assign parameter values for the second independent variable" through "Assign additional parameter values for the second independent variable" as often as necessary to assign all of the independent variable parameter names and to assign their concomitant attribute values.

b To save your experiment parameters and parameter values to a .tsv file, click the **Save experiment parameters** button.

c Click **Next**.

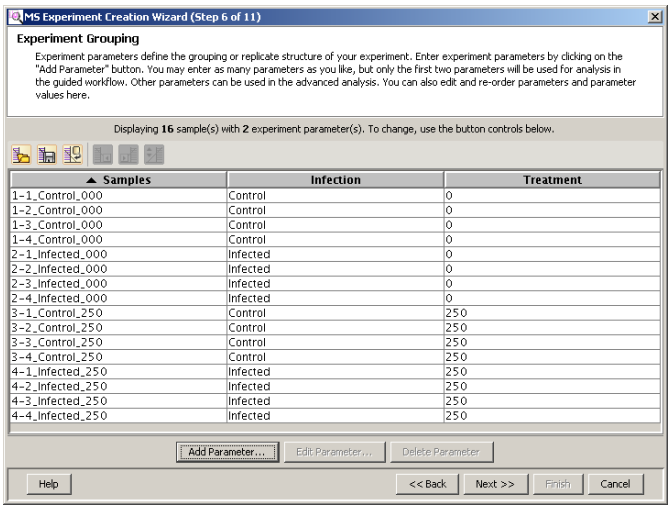


Figure 47 *Assigned parameter values in the Experiment Grouping*

Filter, align, and normalize the sample data

1. Select and enter the data filter parameters in the **MS Experiment Creation Wizard (Step 7 of 11)**.

You filter, align, and normalize your sample data in Step 7 through 11 of the MS Experiment Creation Wizard. At each step of the process you can view your progress and return to prior steps to adjust your results.

Filtering during the data import process may be used to reject low-intensity data or restrict the range of data. After data is imported, several filtering options may be applied: Abundance, Retention Time, Mass, Flags, Number of ions, Mass and Minimum Quality Score.

Note: Filtering works with both GC/MS and LC/MS data. The term abundance actually refers to volume as stored in an MFE generated CEF file. The term abundance actually refers to area as stored in a FbF generated CEF file. The parameters may be cleared to preserve prior filtering that was used to generate the CEF file.

a Mark the **Minimum absolute abundance** check box and type a value of `5000` counts.

b Clear the **Limit to the largest** check box. It is not recommended to set an arbitrary limit with metabolomics. Limiting the number of compounds in metabolomics is based on statistics and occurs later in the workflow during your analysis.

c Clear the **Minimum relative abundance** check box under the Abundance filtering group box.

d Mark the **Use all available data** check box.

e Clear the **Use all available data** check box and type `50.01` for the **Min Mass** and `1000` for the **Max Mass**. Filtering by maximum mass improves the statistical analysis by rejecting masses that are not significant to the experiment.

f Click **Minimum number of ions** and type `2`. The mass filter does not need to include reference ions.

g Click **Multiple charge states forbidden**. Metabolomics involves singly charged ions.

h Click **Next**.

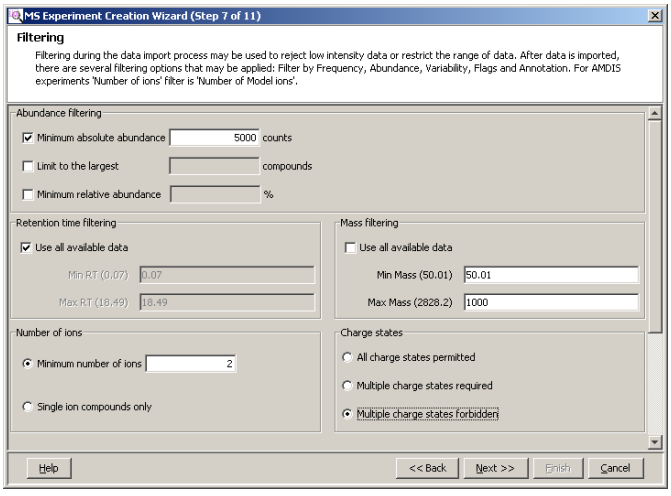


Figure 48 *Recommended filtering parameters*

2. Select and enter the retention time and mass alignment parameters in the **MS Experiment Creation Wizard (Step 8 of 11)**.

Unidentified compounds from different samples are aligned or grouped together if their retention times are within the specified tolerance window and the mass spectral similarity as determined by a simple dot product calculation is above the specified level.

Note: The alignment methodologies for the data types are explained in Section 3.1.7 in the *Mass Profiler Professional User Manual*. Retention alignment rewrites the retention times in the data file so that your input or algorithmically selected features are used to correct the retention times.

a Clear the **Perform RT correction** check box. A larger retention time shift may be used to compensate for less than ideal chromatography.

If retention time correction is used, it is recommended to perform retention time correction with standards, such as 9 anthracine carboxolic acid, provided that at least two widely spaced standards exist, and those standards must be present in every sample. With standards the correction is based on a piecewise linear fit.

b Type `0.1` % and `0.15` min for RT Window under Compound alignment. Smaller values result in reduced compound grouping among the samples leading to a larger list of unique compounds in the experiment.

c Type `5.0` ppm and `2.0` mDa for **Mass Window**. It is not recommended to set the mass window less than 2.0 mDa for higher masses.

d Click **Next**.

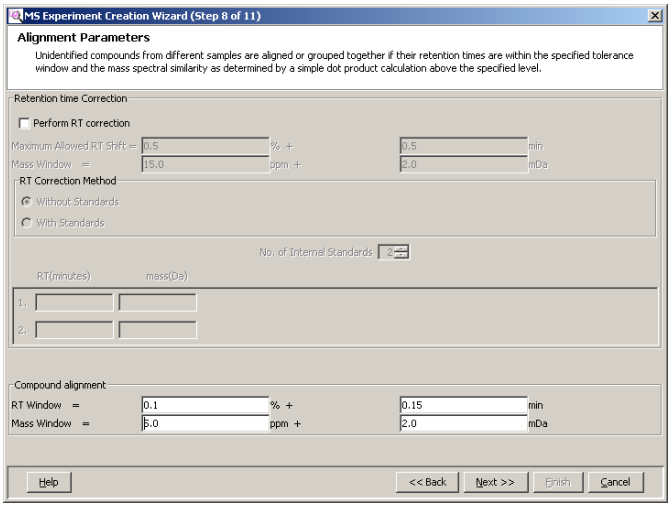


Figure 49 *Recommended alignment parameters*

3. View and review the compounds present and absent in each sample in the **MS Experiment Creation Wizard (Step 9 of 11)**.

This step lets you review a summary of the compounds present and absent in each of the samples based on the experiment parameters including the application of the filter and alignment parameters.

Note: It is useful to click **Back** to make changes in the **Filtering (Step 7 of 11)** page and the **Alignment Parameters (Step 8 of 11)** page parameters and then return to this **Sample Summary (Step 9 of 11)** page several times to develop a feel for how each of the parameters affects the compound summary.

You can independently assess the effects of retention time alignment versus compound alignment. For example, with the **RT Window** to `0.15` minutes in the **Alignment Parameters (Step 8 of 11)** page, the Total number of Aligned Compounds is 4414. If you decrease the **RT Window** to `0.05` minutes, the Total number of Aligned Compounds increases to 4524. If you increase the **RT Window** to `0.30` minutes the Total number of Aligned Compounds decreases to 4361.

a Clear the **Export for Recursion** check box. It is not recommended to export the compounds for recursion at this step in the metabolomics workflow. Better results are obtained after the data has been filtered for significance in the following steps.

b Click **Next**.

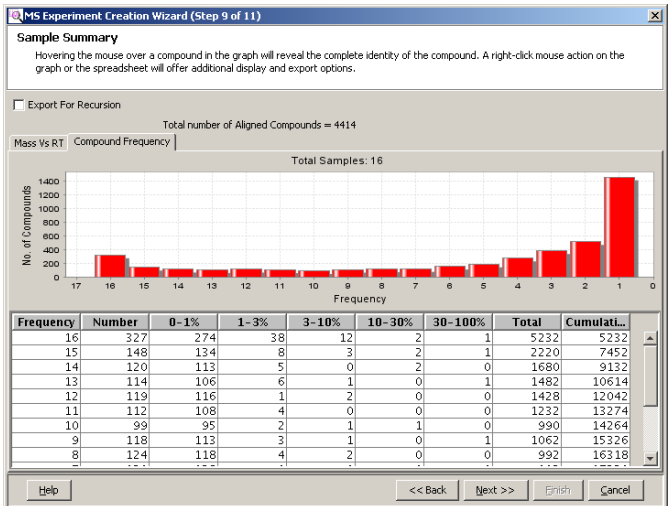


Figure 50 *Compound frequency view in the sample summary page*

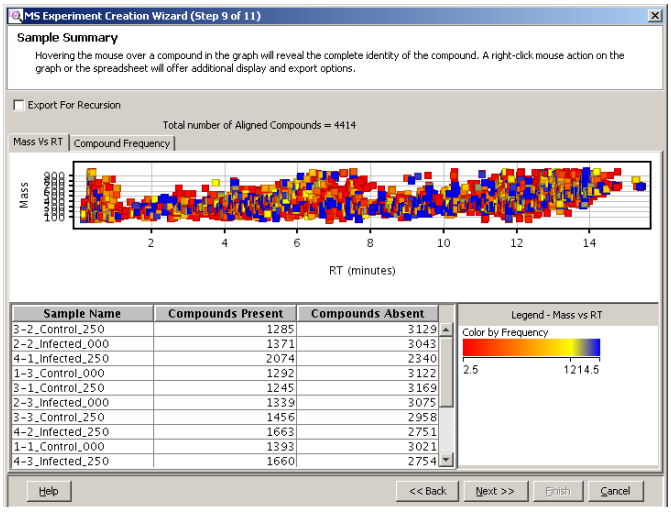


Figure 51 *Mass versus retention time view in the sample summary page*

Replicates should have similar numbers of compounds present and absent. You can see this easily if the files have a systematic naming system that lets replicates be sorted together.

Note: To export any part of the sample summary (Mass vs. RT, Spreadsheet, and Legend), right-click on that part of the summary, and click Export As and click the image type (see Figure 52). The option Image lets you enter the image file location, file name, image size and resolution to meet your organization and publication requirements.

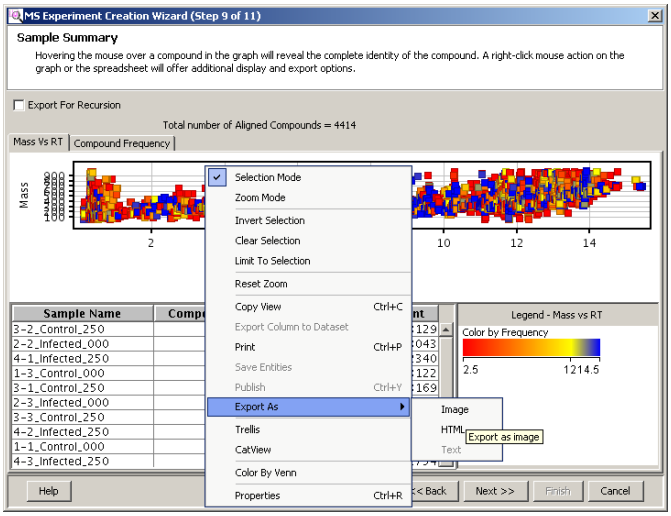


Figure 52 *Export an image of the sample summary for publication*

4. Select whether to normalize the data in the **MS Experiment Creation Wizard (Step 10 of 11)**.

Normalizing the data reduces the variability caused by sample preparation and instrument response. From the list of compounds present in all of the samples you may pick one as an internal standard. No internal or external standard is selected at this time.

Note: Creatinine is a good choice for an internal standard for urine samples.

a Select **None** for the **Normalization Algorithm** in the Normalization tab.

b Clear the **Use External Scalar** check box on the External Scalar tab.

c Click **Next**.

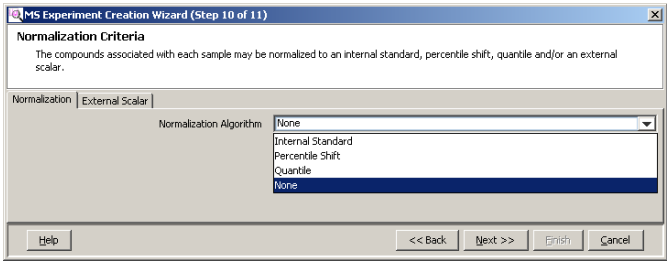


Figure 53 *Normalization tab*

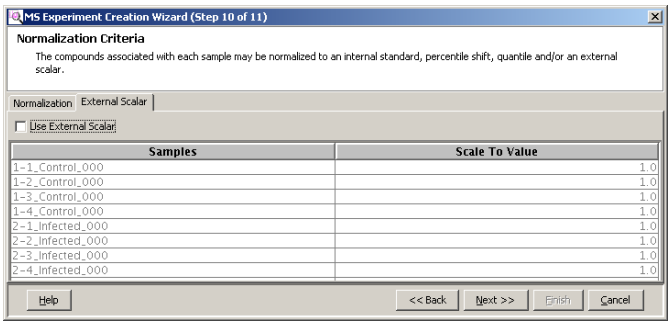


Figure 54 *External Scalar tab*

5. Select whether to compare features in each sample to the response of the features across multiple samples in the **MS Experiment Creation Wizard (Step 11 of 11)**.

Baselining is a technique used to view and compare data. It involves converting the original data values to values that are expressed as changes in the data values relative to a calculated statistical value derived from the data. The calculated statistical value is referred to as the baseline.

There are four baselining options:

1. None: Recommended if only a few features in the samples exist.

2. Z-Transform: Recommended if the data sets are very dense, data where very few instances of compounds are absent from any sample, such as a quantitation data set from recursion.

3. Baseline to ____ of all samples: The abundance for each compound is normalized to its selected statistical abundance (median or mean) across all of the samples. This has the effect of reducing the weight of very large and very small compound features on later statistical analyses.

4. Baseline to ____ of control samples: The abundance for each compound is normalized to its selected statistical abundance (median or mean) across just the samples selected as the control samples. This has the effect of weighting the compound features to a known value that is considered to be normal in the population while reducing the effect of large and small compound features.

a Click **Baseline to ____ of all samples**.

b Select **median** for the **Baseline to ____of all samples**.

c Click **Finish**.

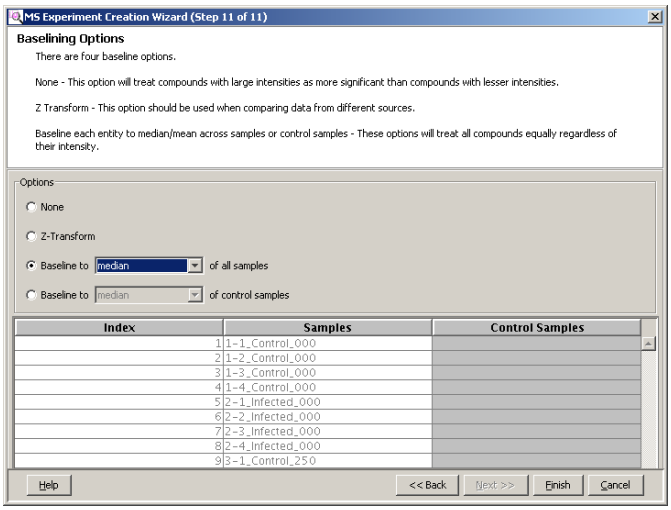


Figure 55 *Selecting baselining options*

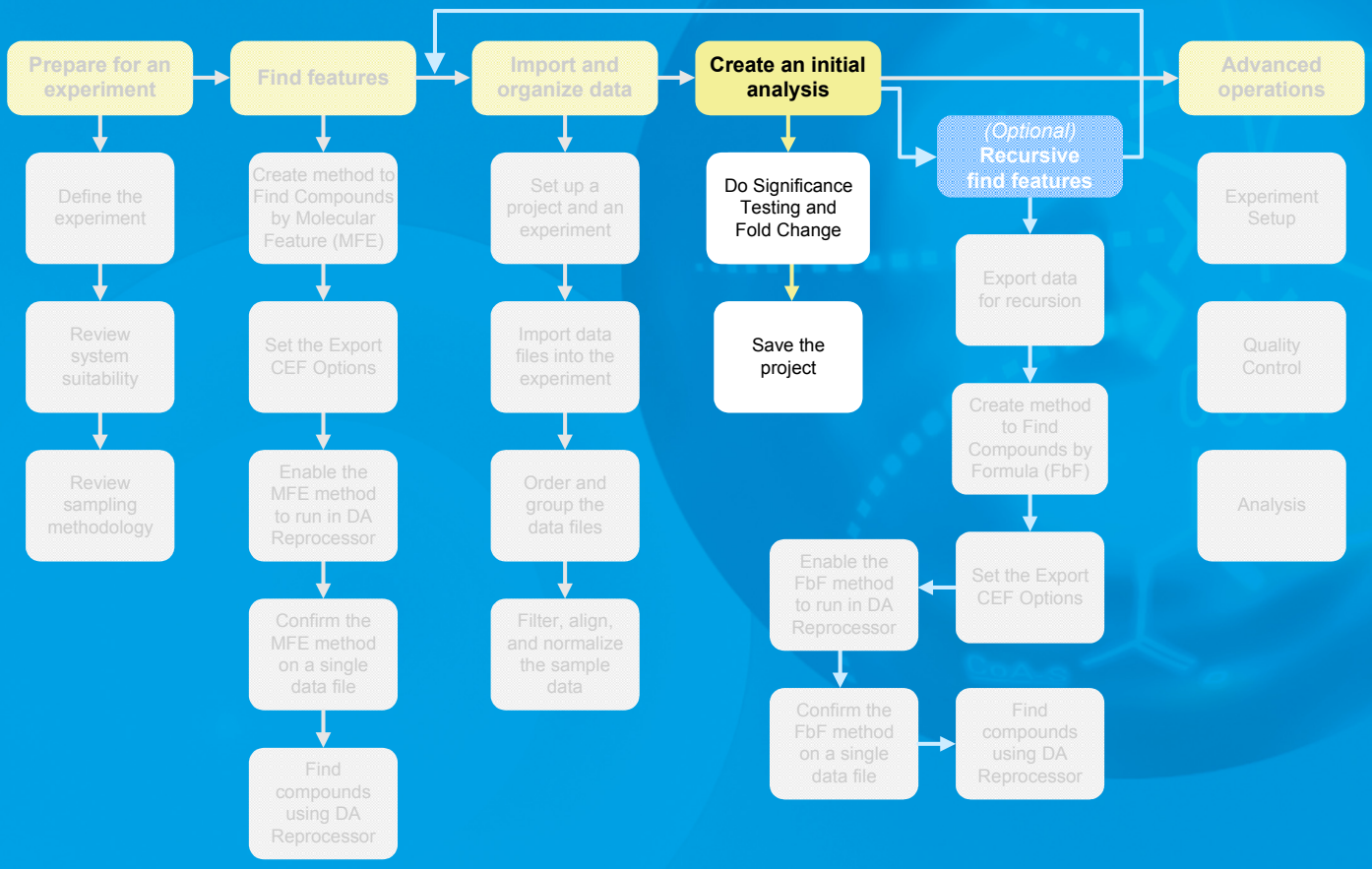
Note: The Significance Testing and Fold Change Wizard immediately starts after the Import Data Wizard if you selected **Analysis: Significance Testing and Fold Change** for the **Workflow type** in the **New Experiment dialog box**.

Next step...

You have now completed the third step of the metabolomics workflow. In the next workflow step you create an initial differential expression from your data using Mass Profiler Professional.

Create an initial analysis

The "Significance Testing and Fold Change Wizard" guides you through the steps necessary to enter parameters and values that improve the quality of your results and produce an initial differential expression for your analysis.



Overview of Significance Testing and Fold Change

The Significance Testing and Fold Change workflow helps you create an initial differential expression from your data and identify the most significant features from among all of the features previously found using molecular feature extraction. The steps necessary to create your initial differential expression are predetermined and based on the experiment type, experiment grouping, and conditions you entered when creating your project and setting up your experiment.

The workflow displays the sequence of steps on the left-hand side navigator with the current step highlighted (see Figure 56 on page 82). Some steps may be automatically skipped for your experiment. All of the parameters can be edited at the conclusion of the Significance Testing and Fold Change workflow by using the operations available in the Workflow Browser (see Figure 67 on page 92).

Feature selection for recursion

The main objective of this initial differential analysis is to export the significant features identified in your data so that they can be used by MassHunter Qualitative Analysis as targeted features to improve your feature finding. It is recommended to process the molecular features in Mass Profiler Professional through at least "Enter the parameters for Filter Flags in the Analysis: Significance Testing and Fold Change (Step 3 of 8) workflow." on page 84. The Filter Flags step is used to require that a feature must be present in at least two samples, which removes "one-hit wonder" features and lets recursive finding in MassHunter Qualitative Analysis run efficiently. A "one-hit wonder" is an entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis.

Note: Importing several thousand features back into MassHunter Qualitative Analysis for targeted finding may cause MassHunter Qualitative Analysis to run out of memory. You use DA Reprocessor to process your complete data set.

Step 1. Summary Report: Displays a summary view of your experiment based on the parameters you provided in the Import Data wizard. A profile plot with the samples on the x-axis and the log normalized abundance values on the y-axis is displayed. If the number of samples is more than 30, the data is represented by a spreadsheet view instead of a profile plot.

Step 2. Experiment Grouping: Independent variables and the attribute values of the independent variables must be specified to define grouping of the samples. An independent variable is referred to as a parameter name. The attribute values within an independent variable are referred to as parameter values. Samples with the same parameter values within a parameter name are treated as replicates.

Step 3. Filter Flags: The compounds created during the experiment creation are now referred to as entities. The entities are filtered (removed) from further analysis based on their presence across samples and parameter values (now referred to as a condition).

Step 4. Filter by Frequency: Entities are further filtered based on their frequency of presence in specified samples and conditions. This filter removes irreproducible entities.

Step 5. Quality Control on Samples: The samples are presented by grouping and the current Principal Component Analysis (PCA). PCA calculates all the possible principal components and visually represents them in a 3D scatter plot. The scores shown

by the axes scales are used to check data quality. The scatter plot shows one point per sample colored-coded by the experiment grouping. Replicates within a group should cluster together and be separated from samples in other groups

Step 6. Significance Analysis: The entities are filtered based on their p-values calculated from a statistical analysis. The statistical analysis performed depends on the samples and experiment grouping.

Step 7. Fold Change: Compounds are further filtered based on their abundance ratios or differences between a treatment and a control that are greater than a specified cut-off or threshold value.

Step 8. ID Browser Identification: The final entity list is directly imported into ID Browser for identification and returned to Mass Profiler Professional.

The Significance Testing and Fold Change workflow lets you to proceed through each step using the **Next** button. A summary of your analysis is presented in each subsequent step. After review of your analysis progress you may return to any previous step and make changes by using the **Back** button. To become more familiar with the analysis parameters and how the parameters affect your data it is recommended that you frequently use the **Back** and **Next** buttons.

To exit the wizard and skip the later steps in the wizard, click **Finish** at any step. When you click **Finish**, the entity list is saved and you may commence analysis using the advanced operations available in the **Workflow Browser**.

Do Significance Testing and Fold Change

Note: The Significance Testing and Fold Change workflow does not start if **Data Import Wizard** was selected as the **Workflow type** in the New Experiment dialog box (Figure 38 on page 62).

Since this step in the metabolomics workflow is to identify the most significant features, Steps 7 and 8 are skipped during the creation of an entity list for recursion.

1. Review the summary report in the **Analysis: Significance Testing and Fold Change (Step 1 of 8)** workflow.

   a To review your data, change the plot view, export selected data, or export the plot to a file, click and right-click features available on the plot.

   It is recommended that you try the review options presented to become familiar with the tools available to you.

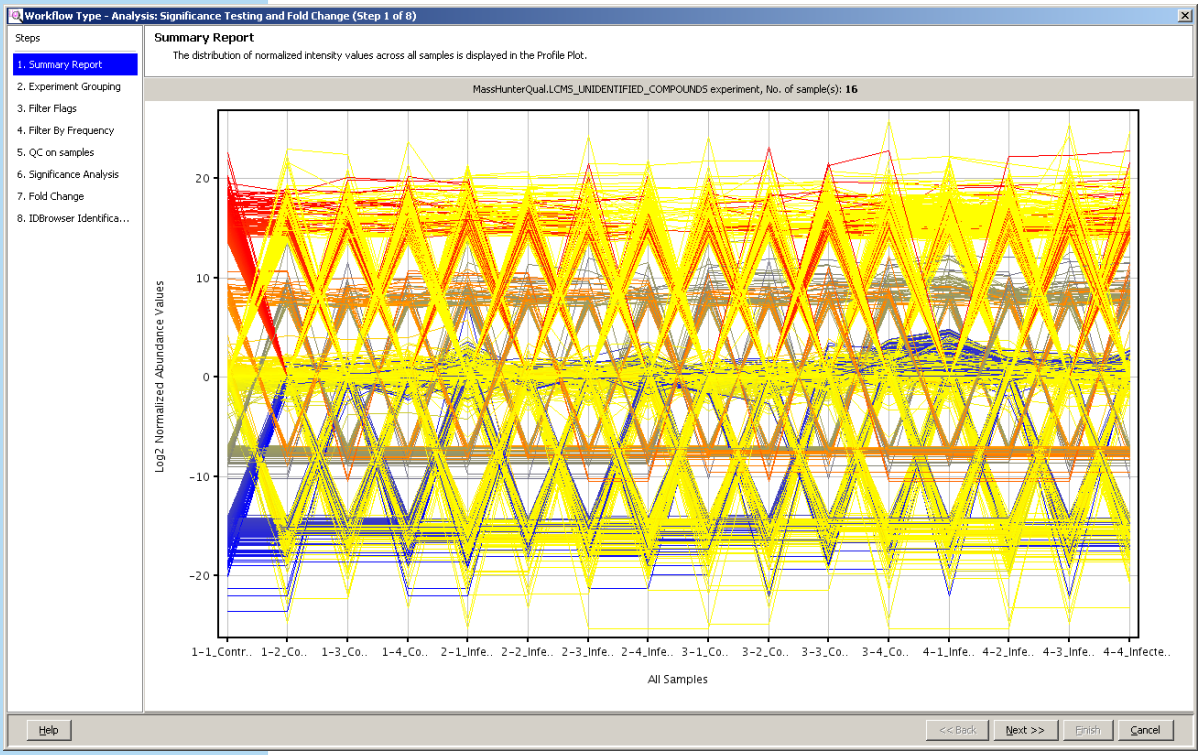


Figure 56 *Summary Report profile plot of the two-variable experiment example showing sixteen samples*

The graphical plot operations available are described in Section 7 Data Visualization in the *Mass Profiler Professional User Manual*. Specifically, the operations available are described in Section 7.5 The Profile Plot View in the *Mass Profiler Professional User Manual*.

Some of the tools available for you to review the data are:

- **Click**: Click any line in the Profile Plot to select an entity line. Multiple entities may be selected by pressing the **Ctrl** key during successive clicks.
- **Double-click**: Double-click any one line within the Profile Plot to inspect a single entity. You may also double-click the last entity selected while pressing the **Ctrl** key to inspect multiple entities. The Entity Inspector lets you view the specific entity Annotation, Data, Profile Plot and Spectra.

- **Right-click**: Right-click anywhere on the Profile Plot to change the selection mode, entity selection, and plot view or to export the plot to a file.

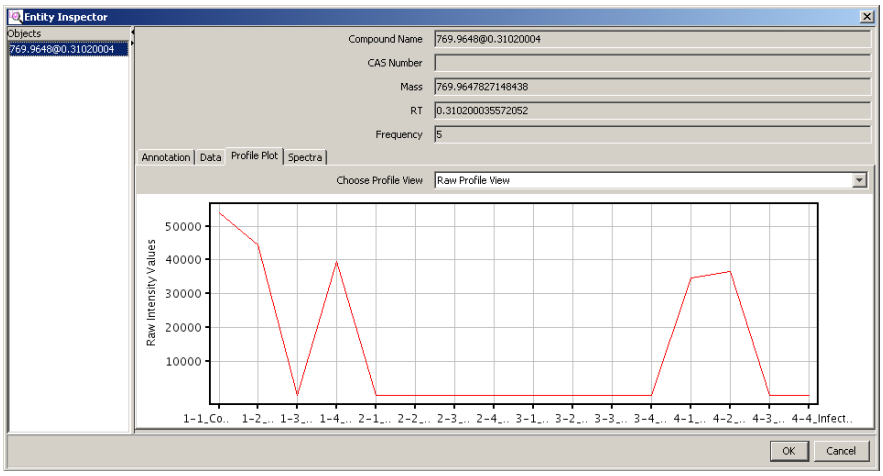


Figure 57 *Entity Inspector from the Summary Report (double-click)*

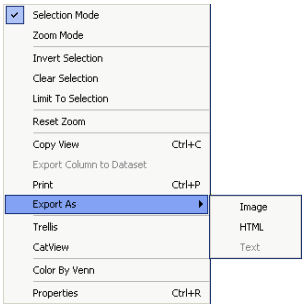


Figure 58 *Shortcut menu for the Summary Report profile plot (right-click)*

b Click **Next**.

---

2. Enter the experiment grouping parameters associated with the independent variables and their attribute values in the **Analysis: Significance Testing and Fold Change (Step 2 of 8)** workflow.

In this step you have an opportunity to edit or change your experiment grouping. An independent variable is referred to as a parameter name. The attribute values within an independent variable are referred to as parameter values. Samples with the same parameter values within a parameter name are treated as replicates.

Only the first two parameter names (independent variables) are used for analysis in the Significance Testing and Fold Change workflow. All of the parameters are available in the Workflow Browser at the completion of the Significance Testing and Fold Change workflow.

Note: In order to proceed, at least one parameter with two values must be assigned.

Note: When entering Parameter Names and parameter Assign Values, it is very important that the entries use identical letters, numbers, punctuation, and case in order for the Experiment Grouping to function properly. Click **Back** or **Experiment Setup > Experiment Grouping** to return to Experiment Grouping if an error is identified later in the Significance Testing and Fold Change workflow or while performing operations in the Workflow Browser, respectively.

a Click **Add Parameter**. The Grouping of Samples dialog box is opened.

b Edit or change your experiment grouping by following the procedure presented in chapter "Import and organize data" step "Group samples based on the independent variables and replicate structure of your experiment in the MS Experiment Creation Wizard (Step 6 of 11)." on page 67.

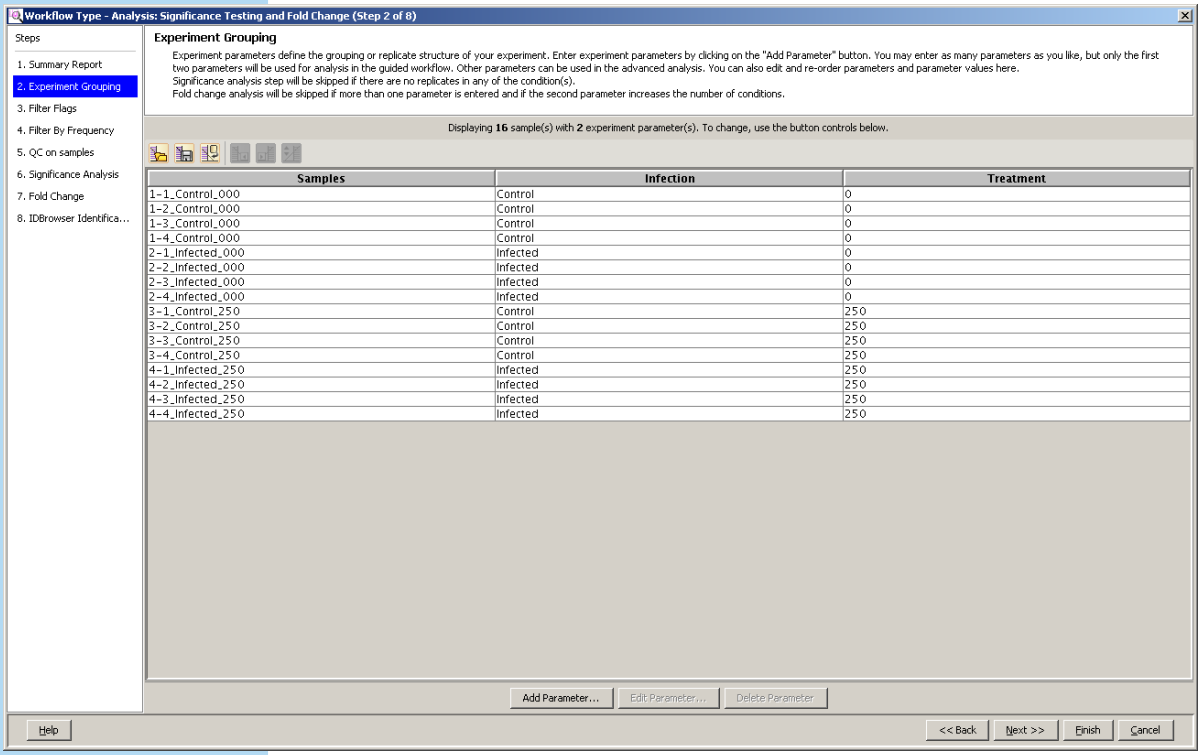


Figure 59 *Experiment grouping for the two-variable experiment example*

c Click **Next**.

3. Enter the parameters for Filter Flags in the **Analysis: Significance Testing and Fold Change (Step 3 of 8)** workflow.

The entities may now be filtered (removed) from further analysis based on their presence or absence across the samples and parameter values (now referred to as a condition). A flag is a term used to denote the quality of an entity within a sample. A flag indicates if the entity was detected in each sample as follows: Present means the entity was detected, Absent means the entity was not detected, and Marginal means the signal for the entity was saturated. See "Definitions" on page 180 for more definitions and relationships of the terms used by the metabolomics workflow.

Note: Before using the **Re-run Filter** button, you can review the data, change the plot view, export selected data, or export the plot to a file using the click and right-click features available on the plot in the same manner as presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. The graphical plot operations available are described in Section 7 Data Visualization in the *Mass Profiler Professional User Manual*.

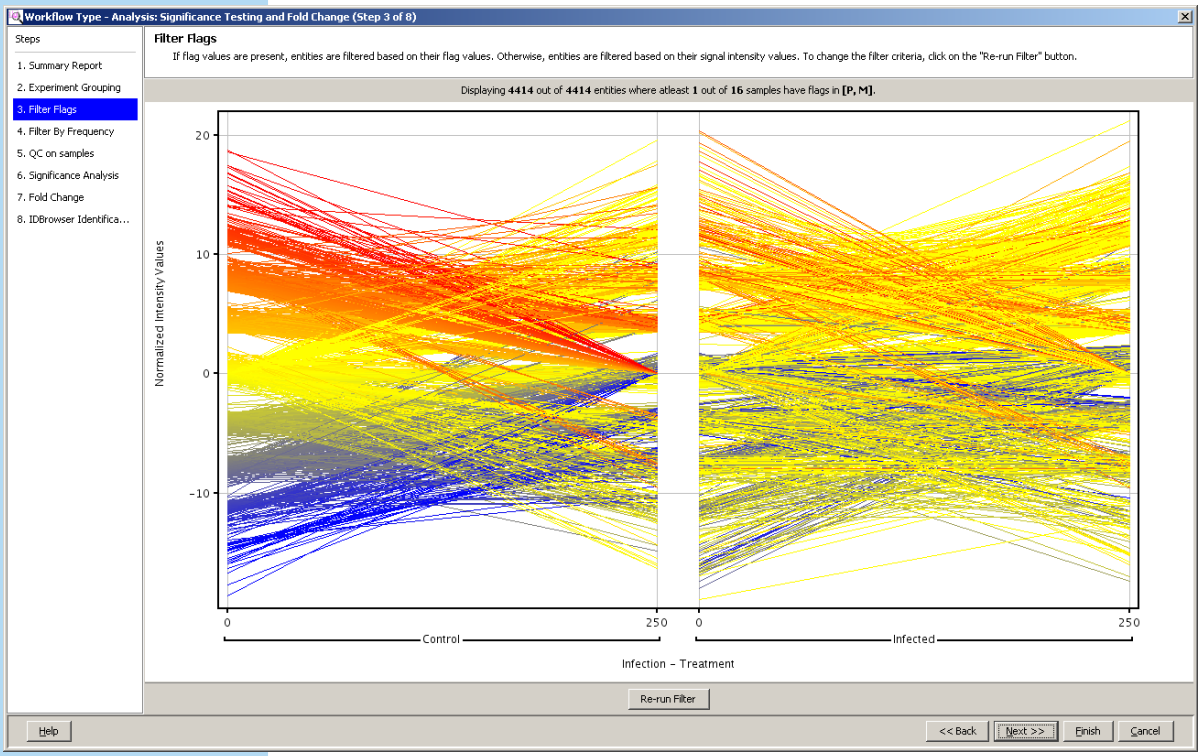


Figure 60 *Profile Plot after Experiment Grouping showing the final four permutations (based on two independent variables, each with two attribute values) from the sixteen samples*

A major objective of Filter Flags is to remove "one-hit wonders" from further consideration. A "one-hit wonder" is an entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis.

a Click **Re-run Filter**.

b Mark the **Present** check box.

c Mark the **Marginal** check box.

d Clear the **Absent** check box. This flag is useful when you want to identify entities that are missing in the samples. You can use this flag in conjunction with the **Next** and **Back** buttons to review the entities that are missing in some samples.

e Click **at least ___ out of X samples have acceptable values**. The value "X" is replaced in your display with the total number of samples in your data set.

f Type 2 in the entry box. By setting this parameter to a value of two or more, one-hit wonders are filtered.

g Click **OK**.

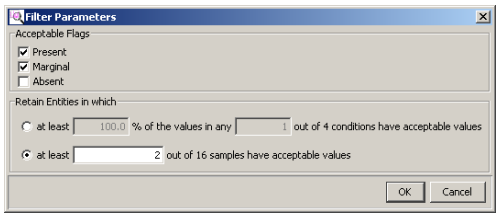


Figure 61 *Recommended filter parameters for Filter Flags*

Note: With the two-variable experiment example, the number of displayed entities changes to 2930 out of 4165 entities if you started from the original data files. If you started with the sample CEF files the display changes to 2957 out of 4414 entities. The reduction in entities displayed reflects successful filtering of one-hit wonders.

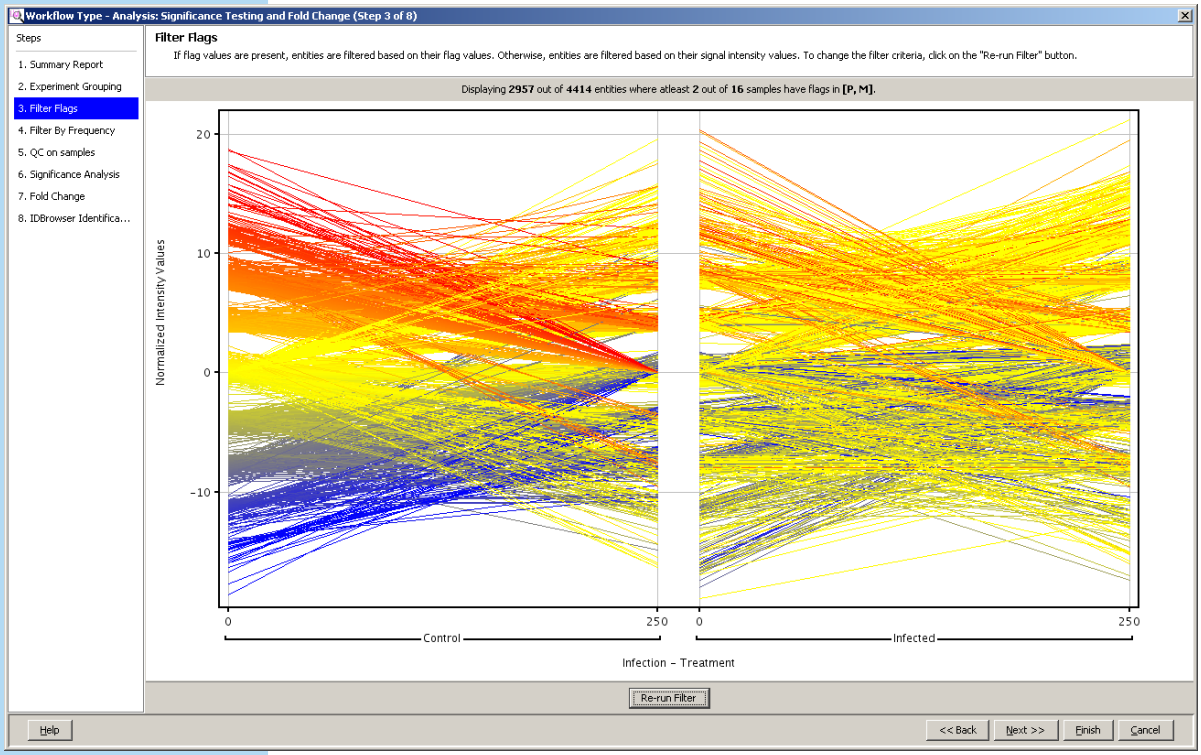


Figure 62 *Profile plot after removing one-hit wonders*

h (optional) To re-adjust the filter parameters again, click **Re-run Filter** until the results displayed in the Profile Plot are satisfactory. It is recommended that the filter be re-run several times with differing parameters to develop an understanding of how each parameter affects the results.

i Click **Next**.

---

4. Enter the parameters for Filter By Frequency in the **Analysis: Significance Testing and Fold Change (Step 4 of 8)** workflow.

The entities may now be filtered from further analysis based on their frequency of occurrence among the samples and conditions. See "Definitions" on page 180 for definitions and relationships of the terms used by the metabolomics workflow.

Filter by Frequency defines the filter by the minimum percentage of samples an entity must be present in to pass the filter. The filter is specified by typing the mini-

mum percentage and selecting the applicable condition of the samples for which each entity must be present, i.e., Retain entities that appear in at least _____ %:

- of all the samples (conditions are not evaluated)
- of samples in only one condition (one and only one condition)
- of samples in at least one condition (one or more conditions)
- of samples within each condition (all conditions)

Filter by Frequency is set by default to retain the entities that appear in at least 100% of all the samples in at least one condition. This is the recommended percentage for experiments that contain five or fewer replicates. A larger percentage removes more entities from further statistical consideration. For experiments with a larger number of replicates the filter frequency percentage may be lowered to reflect the required occurrence.

a Click **Re-run Filter**.

b Type `100` in the **Retain entities that appear in at least** box.

c Click **of samples in at least one condition**.

d Click **OK**.

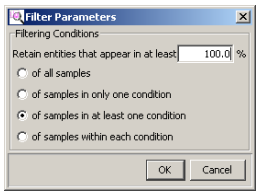


Figure 63 *Recommended Filter by Frequency parameters for experiments with few replicates*

Note: With the two-variable experiment example, the Profile Plot changes to reflect displaying 1220 out of 2957 entities, reflecting the successful filtering by frequency (see Figure 64 on page 88).

e Click **Re-run Filter** to re-adjust the filter parameters until the results displayed in the Profile Plot are satisfactory. It is recommended that the filter be re-run several times with different parameters to develop an understanding of how each parameter affects the results.

You can review the data, change the plot view, export selected data, or export the plot to a file by using left-click and right-click features available on the plot in the same manner as presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. The graphical plot operations available are described in Section 7 Data Visualization in the *Mass Profiler Professional User Manual*.

f Click **Next**.

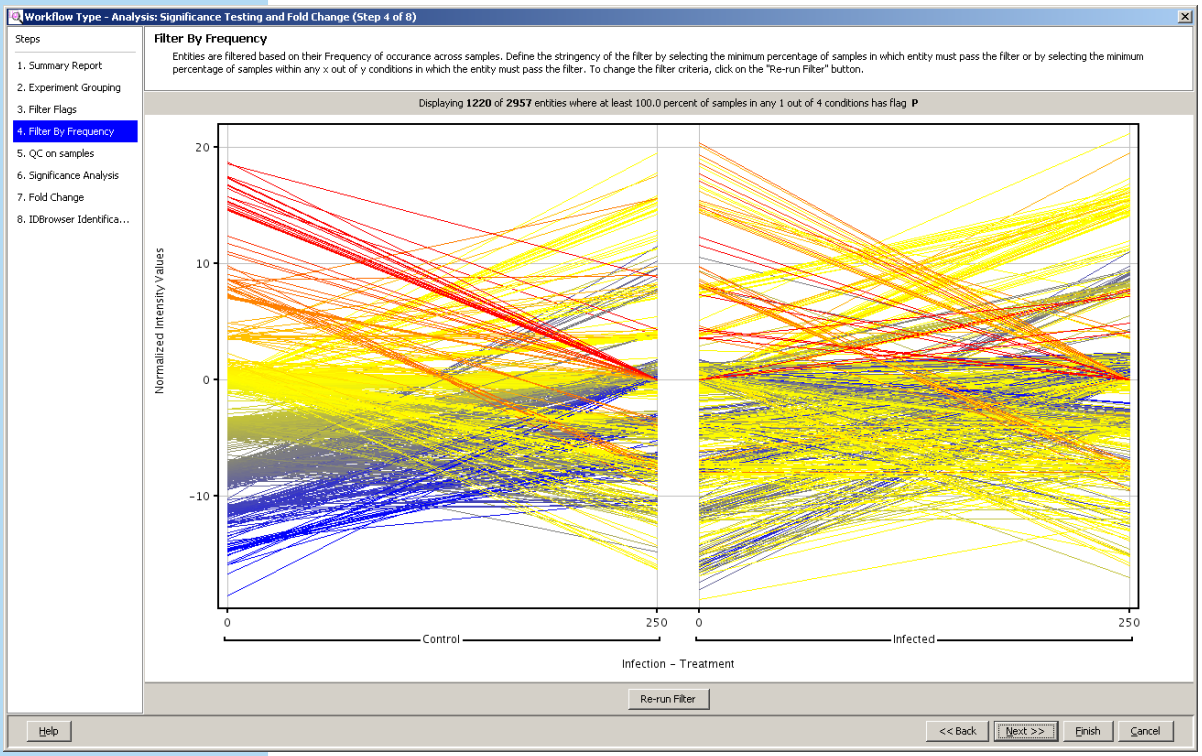


Figure 64 *Profile plot after removing entities that do not appear in at least 100% of all the samples in at least one condition*

5. Review the sample quality in QC on samples in the **Analysis: Significance Testing and Fold Change (Step 5 of 8)** workflow.

This step provides the first view of the data using a Principal Component Analysis (PCA). PCA lets you assess the data by viewing a 3D scatter plot of the calculated principal components. The PCA scores are shown in each of the selection boxes located along the bottom of the 3D PCA Scores window. A higher score indicates that the principal component contains more of the variability of the data. The components generated in the 3D PCA Scores graph are represented in the X, Y, and Z axes and are numbered 1, 2, 3 ... in order of their decreasing significance.

Principal component analysis: The mathematical process by which data containing a number of potentially correlated variables is transformed into a data set in relation to a smaller number of variables called principal components that account for the most variability in the data. The result of the data transformation leads to the identification of the best explanation of the variance in the data, e.g. identification of the components in the data that contain the meaningful information providing differentiation.

Principal component: Transformed data into axes, principal components, so that the patterns between the axes most closely describe the relationships between the data. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components are viewed and interpreted in 3D graphical axes with additional dimensions represented by different colors and/or shapes representing the parameter names.

QC on samples display is divided into three viewing areas

a Review the **Experiment Grouping view**. This table lets you view each of the samples within a parameter (now referred to as a group). Ideally, replicates within a group should cluster together and be separated from samples in other groups. The spreadsheet operations available are described in section 7.3 The Spreadsheet View in the *Mass Profiler Professional User Manual*.

b Review the **3D PCA Scores scatter plot view**. You may change the plot view or export the plot to a file by using the left-click and right-click features available on the plot in the same manner presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. Additional controls available are:

- To customize the 3D PCA scores plot, right-click and then click **Properties**.
- To zoom into the 3D scatter plot, press the **Shift** key and simultaneously click the mouse button and drag the mouse upwards.
- To zoom out, press the **Shift** key and simultaneously click the mouse button and drag the mouse downwards.
- To rotate, press the **Ctrl** key and simultaneously click the mouse button and drag the mouse around the plot.

c Review the **Legend - 3D PCA Scores view**. This window shows the legend of the scatter plot.

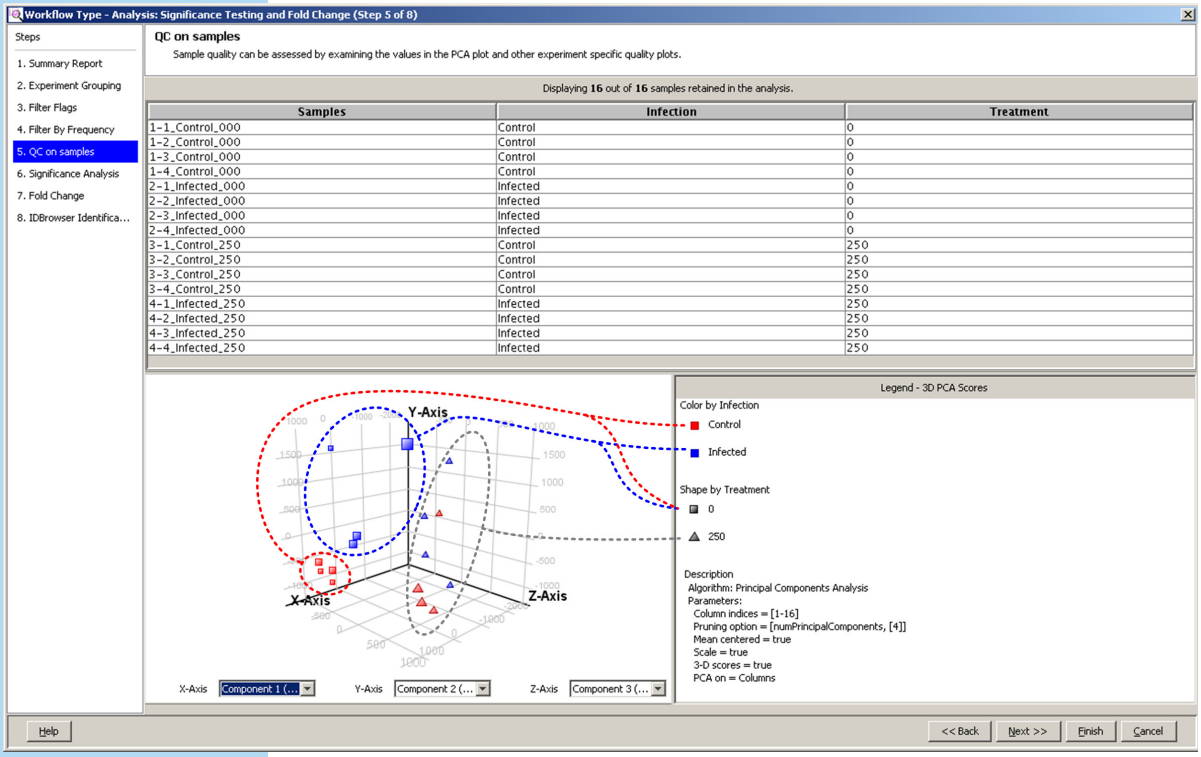


Figure 65 *QC on samples PCA Score showing the initial separation of the infection parameters and separation of the treatment values*

Note: It is recommended to click **Back** to make changes in the parameters for **Filter Flags** on page 84 and **Filter By Frequency** on page 86 then return to **QC on samples** step several times to understand how each of the parameters affects your compound summary.

d Click **Next**.

6. Assess the differential Significance Analysis in the **Analysis: Significance Testing and Fold Change (Step 6 of 8)** workflow.

The entities are filtered based on their p-values calculated from a statistical analysis that is selected based on the samples and experiment grouping.

The statistical analysis is either a T-test or an Analysis of Variance (ANOVA) based on the samples and experiment grouping. The statistical analysis applied is described in section 4.7 Significance Analysis in the *Mass Profiler Professional User Manual*.

a Review the Significance Analysis display. The display is divided into four viewing areas:

Test Description view: The statistical test applied to the samples is described.

Result Summary view: A summary table that organizes the results by p-value. A p-value of 0.05 is similar to stating that if the mean values for each parameter value (a condition of an independent variable) are identical, then a 5% chance or less exists of observing a difference in the mean of the parameter values as large as you observed. In other words, statistical treatment of random sampling from identical populations with a p-value set at 0.05 leads to a difference smaller than you observed in 95% of the experiments and larger than you observed in 5% of the experiments.

The last row of data in the Result Summary (see Figure 66 on page 91) shows the number of entities that would be expected to meet the significance analysis by random chance based on the p-value specified in each column heading. If the number of entities expected by chance is much smaller than the number expected based on the corrected p-value you have realized a selection of entities that show significance in the difference of the mean values of the parameter values.

The spreadsheet operations available are described in section 7.3 The Spreadsheet View in the *Mass Profiler Professional User Manual*.

Compounds p-Values Table view: Each entity that survived the filters is now presented by compound along with the p-values expected and corrected for each of the interpretation sets. Each entity is uniquely identified by its average neutral mass and retention time from across the data sets.

Venn Diagram view: Display of the Venn Diagram, or other plot, depends on the samples and experiment grouping for the analysis (see Figure 66 on page 91). The entities that make up each selected section of the Venn diagram are highlighted in the p-values spreadsheet. The Venn diagram is a graphical view of the most significant entities in each of the samples. Where entities in common to the analyses exist, they are depicted as overlapping sections of the circles. Fewer entities in the regions of overlap are an indication that the samples support the hypothesis that a difference exists in the samples based on the experimental parameters.

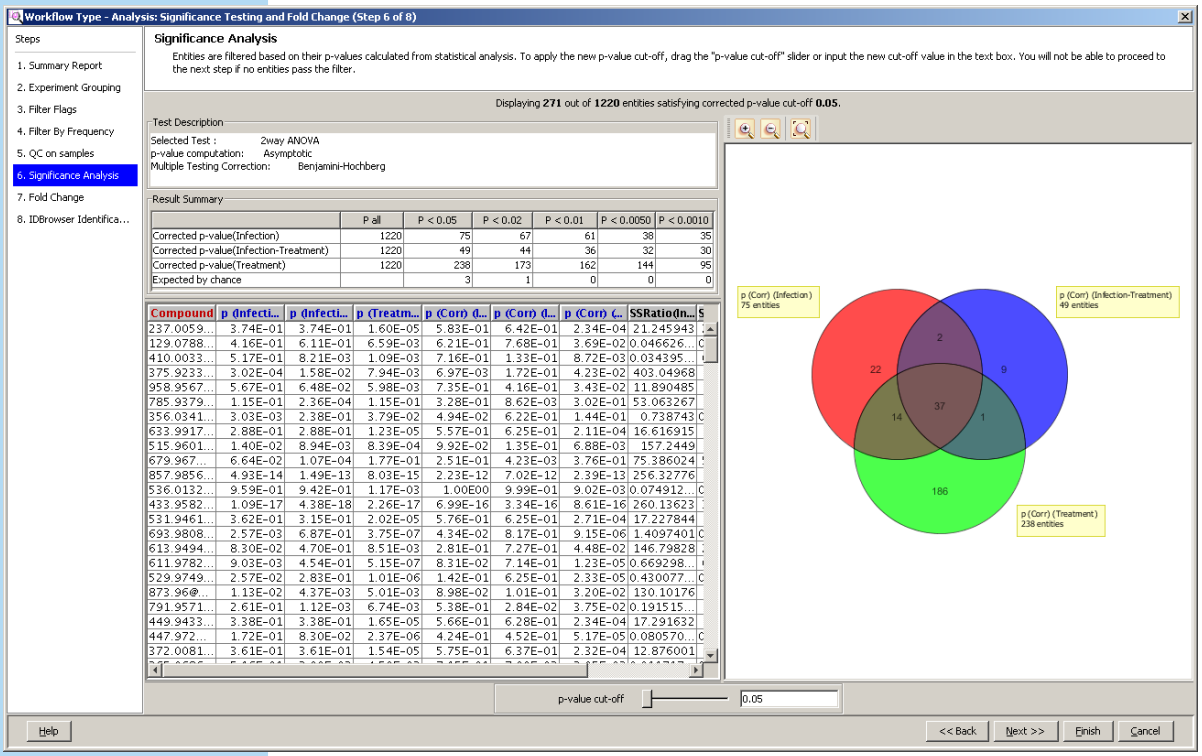


Figure 66 *Significance analysis based on a 2-way ANOVA using the two-variable experiment example. The results show a Venn diagram. A 1-way ANOVA significance analysis does not present a graphical representation of the entities relationships.*

Note: To change the plot view or export the plot, click and right-click features available on the plot in a same manner similar to that presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. The graphical plot operations available are described in Section 7 Data Visualization in the *Mass Profiler Professional User Manual*.

b Click and move the **p-value cut-off** slider or type in the p-value cut-off value and press the **Enter** key. The default value is 0.05. The results in the display window are automatically updated.

c Re-adjust the p-value cut-off until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the p-value cut-off affects the results. A larger p-value passes a larger number of entities.

d Click **Next**.

7. Skip Fold Change in the **Analysis: Significance Testing and Fold Change (Step 7 of 8)** workflow.

Fold Change is skipped by the Significance Analysis and Fold Change workflow for the two-variable example.

8. Skip ID Browser Identification in the **Analysis: Significance Testing and Fold Change (Step 8 of 8)** workflow.

You do not select ID Browser at this time because the object of this step in the workflow with Mass Profiler Professional is to generate a CEF file containing the most significant features for recursion in MassHunter Qualitative Analysis.

a Click **Finish**.

Layout of the Mass Profiler Professional screen

You are now in the advanced workflow mode and have access to all features available in Mass Profiler Professional through the Workflow Browser. Figure 67 shows the layout of Mass Profiler Professional.

Further information about may be obtained by pressing the **F1** key or reviewing the *Mass Profiler Professional User Manual*.

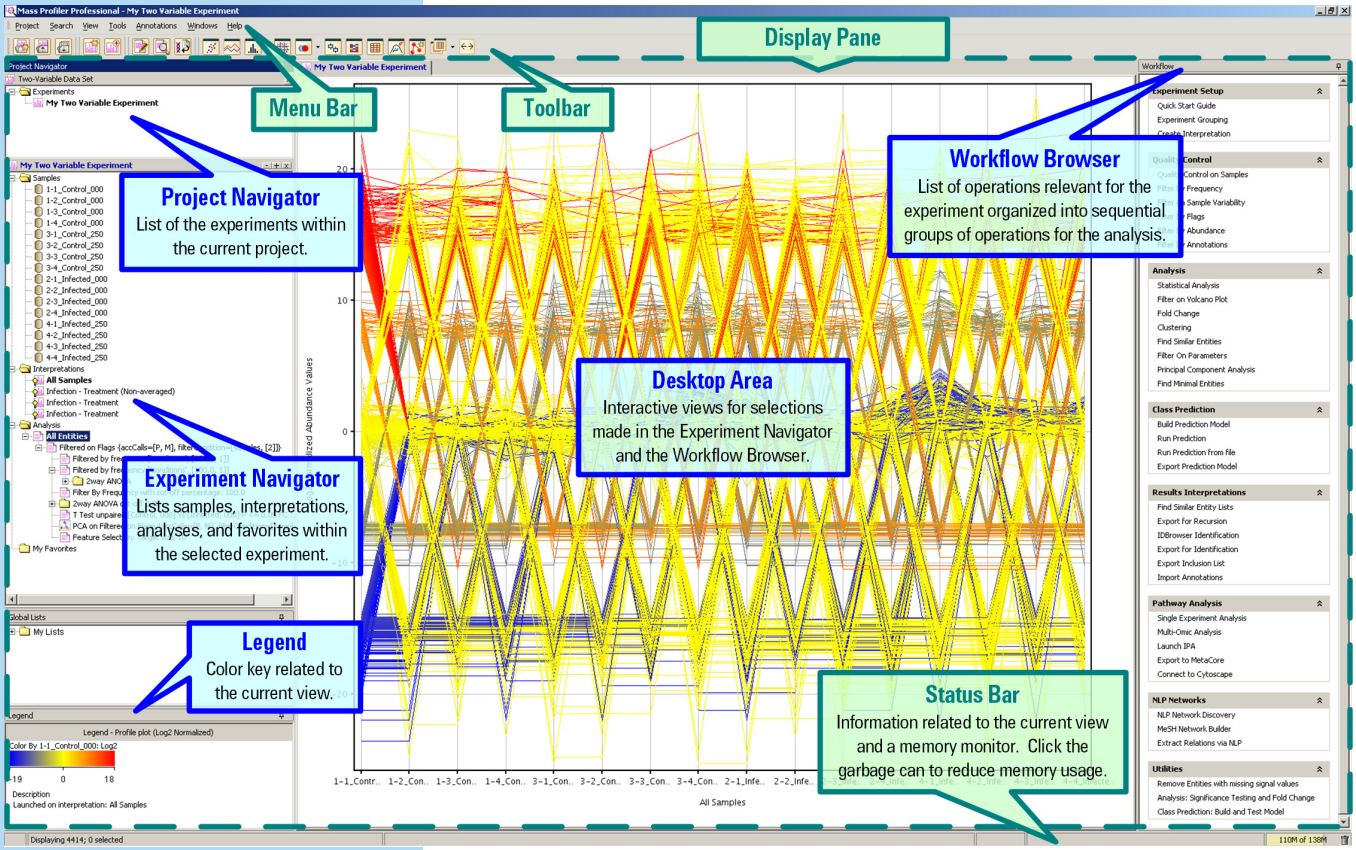


Figure 67 *The main functional areas of Mass Profiler Professional*

Save the project

Save your current analysis as a TAR file for archiving, restoration of any future analysis to the current results, sharing the data with a collaborator, or sharing the data with Agilent customer support.

a Click **Project > Export Project**.

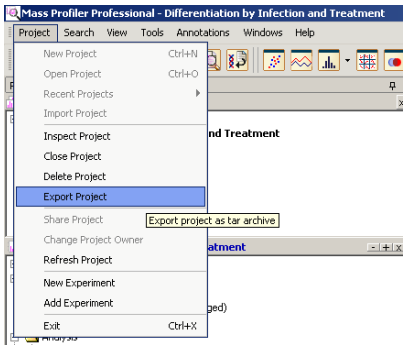


Figure 68 *Menu selection to export your current analysis*

b Mark the check box next to the experiment you wish to save.

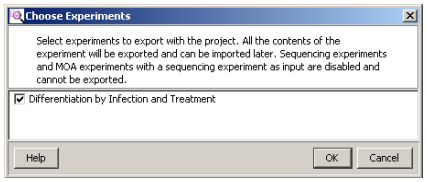


Figure 69 *Choose Experiments dialog box for saving your experiment*

c Click **OK**.

d Select or create the file folder.

e Type the **File name**.

f Click **Save**.

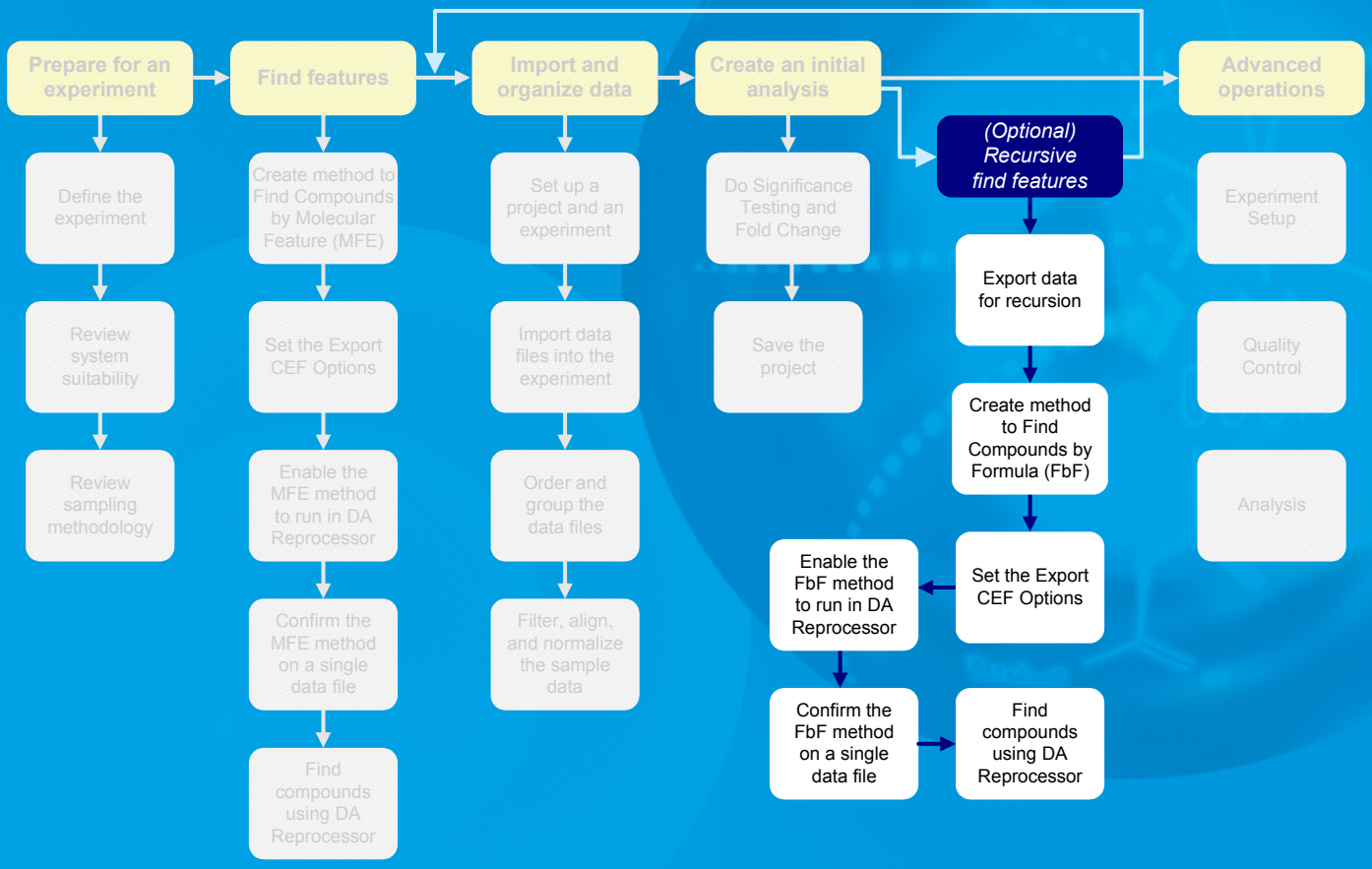
Next step...

You have now completed the fourth step of the metabolomics workflow. In the next workflow step you export your most significant features back into MassHunter Qualitative Analysis as targeted features to improve finding of the features from your samples.

If your analysis is ready to perform operations available under the Workflow Browser you may skip "Recursive find features" and proceed to "Advanced operations" on page 117.

Recursive find features

Importing the most significant features back into MassHunter Qualitative Analysis as targeted features improves finding the features in your samples. This repeated feature finding is referred to as recursion. Improved reliability in finding your features leads to improvement in the accuracy of your analysis.



Agilent Technologies

Overview of recursive find features

If your analysis does not involve recursive feature finding you may skip this step in the workflow and continue with your analysis at "Advanced operations" on page 117.

Recursive feature finding combined with replicate samples improves the statistical accuracy of your analysis and reduces the potential for obtaining a false positive or false negative answer to your hypothesis.

Recursive finding consists of three steps

1. Untargeted **Find Compounds by Molecular Feature** in MassHunter Qualitative Analysis (see "Find features" on page 39) to find your initial entities.
2. Filtering by **Significance Testing and Fold Change** using abundance, retention time, sample variability, flags, frequency, and statistical significance in Mass Profiler Professional (see "Create an initial analysis" on page 79) to find your most significant entities.
3. Targeted **Find Compounds by Formula** in MassHunter Qualitative Analysis (this section "Recursive find features") to improve the reliability of finding your features and subsequently improve your statistical analysis accuracy.

As you follow the steps involved in recursive find features, keep in mind that the term "feature" is used synonymously with metabolite, compound, element, or entity.

Find Compounds by Formula (FbF) typically uses molecular formula information to calculate the ions and isotope patterns derived from the formula as the basis to find features in the sample data file. When the input molecular features consist of mass and retention time, instead of molecular formula, FbF calculates reasonable isotope patterns and uses these patterns with retention time tolerances to find the target features in the sample data files. When the input molecular features are filtered from a find process that was previously untargeted, the molecular features found using this repeated process of finding molecular features is referred to as recursive finding.

Export data for recursion

You export the entities identified in the initial differential analysis performed in step three of the metabolomics workflow and use these entities to perform a targeted feature find from your original data sets.

Export the entities that have been identified as the most relevant to the differential analysis for recursive finding in MassHunter Qualitative Analysis

---

a Click **Export for Recursion** in the Workflow Browser under the Results Interpretations group heading. This displays the Export dialog box.

b Click **Choose** to select the Entity List for exporting.

c Click **Filtered by frequency** from the entity lists in the Choose EntityList dialog box. For more significance in your analysis, select an entity list that has at least been filtered by flags to remove one-hit wonders.

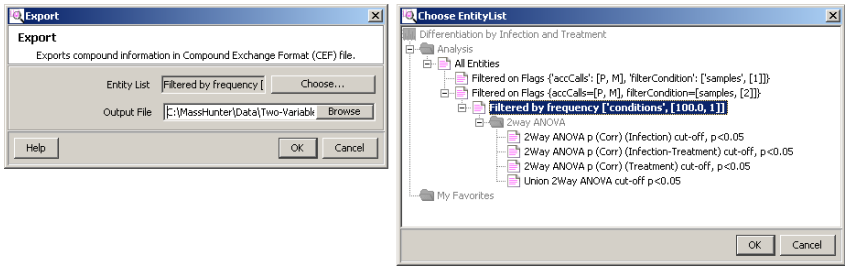


Figure 70 *Export for recursion and Choose EntityList dialog boxes*

d Click **OK**.

e Click **Browse** in the Export dialog box.

Do not type a file name at this location.

f Select the folder to which to save the file.

g Type the **File name**. For example, you can type `Two-Variable Filtered by Frequency.cef` for your current example data set.

h Click **Save**.

i Click **OK**.

Start Agilent MassHunter Qualitative Analysis

The following examples use Agilent MassHunter Qualitative Analysis B.05.00 running on 64-bit Windows 7 Professional.

1. Start MassHunter Qualitative Analysis Software.

a Double-click the Qualitative Analysis icon  located on the desktop,

or (for Qualitative Analysis version B.05.00 or later on Windows 7)

Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.05.00**,

or (for Qualitative Analysis version B.03.01 on Windows XP)

Click **Start > Programs > Agilent > MassHunter Workstation > Qualitative Analysis.**

b Click **Cancel** in the **Open Data File** dialog box to start MassHunter Qualitative Analysis without opening any data files. To open data files later click **File > Open Data File.**

You do not need to open a data file at this time. You are prompted to open a data file in "Confirm the FbF method on a single data file" on page 112.

2. Enable advanced parameters in the user interface.

Advanced parameters must be enabled in MassHunter Qualitative Analysis in order to show tabs labeled Advanced in the Method Editor and to enable compound importing for recursive finding of molecular features.

a Check to make sure that **File > Import Compound** is an available command. See Figure 14 on page 41.

b If **File > Import Compound** is not available follow the instructions in "Enable advanced parameters in the user interface." on page 40.

c Continue with the next step.

Create a method to Find Compounds by Formula

FbF involves chromatographic deconvolution as described in "Capabilities of the metabolomics workflow" on page 18 (see Figure 6 on page 19) based on the target entity list exported at the beginning of this step in the workflow. Find Compounds by Formula automatically finds related co-eluting ions, sums the related ion signals into single values, creates compound spectra, and reports results for each molecular feature.

All of the parameters involved in Find Compounds by Formula are accessed in the tabs presented in four Method Editor sections that are selected from the Method Explorer window:

Find by Formula - Options: Specify the rules that are applied to match the data based on isotope patterns (*m/z* and abundance) and retention time

Find by Formula - Chromatograms: Enter parameters that are applied to the chromatographic component of the data to extract features.

Find by Formula - Mass Spectra: Enter parameters that are applied to the mass spectral component of the data to extract features.

Find by Formula - Sample Purity: Not used in this workflow.

After the parameters are entered in all four Method Editor sections, to Find Compounds by Formula on a single sample data file click the **Find Compounds by Formula** button from within any one of these Method Editor sections.

---

1. Enter the parameters in the Find by Formula - Options section.

The parameters in this section specify the rules that are applied to the formula database to match against the data based on isotope patterns (*m/z* and abundance) and retention time. This is the first part of the recursive refinement of finding features.

a Click **Find Compounds by Formula > Find by Formula - Options** in the Method Explorer window. The input options specify the rules that are applied to the input molecular formula database.

Formula Source tab

b Enter the parameters on the **Formula Source** tab.

The parameters on this tab let you use a molecular formula or previously created databases as the source of targeted features to find. In the metabolomics workflow the source of targeted features is the .CEF file you exported for recursion.

1. Click the **Formula Source** tab.
2. Click **Compound exchange file (.CEF)**.
3. Type the folder and file name of the CEF file or click **Browse** and select a CEF file from the **Open CEF file** dialog box.
4. Open the CEF file that contains the most significant features created using Mass Profiler Professional from the "Export data for recursion" on page 97.
5. Click **Open**.
6. Click **Mass and retention time (retention time required)**.

Note: The parameters under the **Values to match** group heading are only active if the **Compound exchange file (.CEF)** button or the **Database** button is clicked.

Note: **Mass and retention time (retention time required)** is the proper selection for a CEF file. Mass and retention time (retention time optional) is the proper selection for a database source.

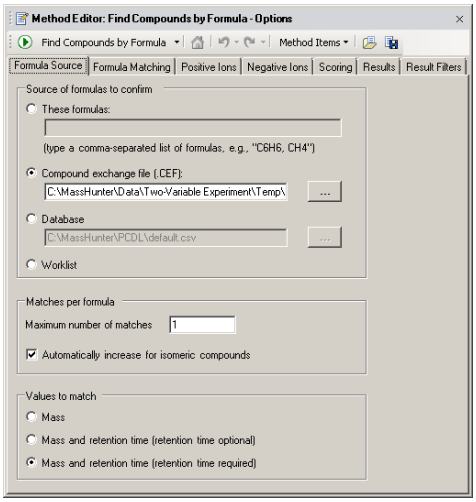


Figure 71 *Formula Source tab in the Find by Formula - Options section*

Formula Matching tab

c Enter the parameters on the **Formula Matching** tab.

The parameters in this tab specify the tolerances that are used to match the input values for mass and retention time against those found in the data.

1. Click the **Formula Matching** tab.
2. Type `20` for **Masses** tolerance and select **ppm** as the match tolerance units. It is important to set this value wider than the measured instrumental acquisition mass tolerance to avoid losing valid feature matches.
3. Type `0.15` for **Retention times**. This parameter should be no less than two times the measured retention time tolerance.
4. Select **Symmetric (ppm)** and ± **20** ppm for **Possible m/z**. The parameters under the Expansion of values for chromatographic extraction group box are used to direct the algorithm on how to handle saturated chromatographic data.
5. Mark the **Limit EIC extraction range** check box.
6. Type a value of `1.0` minutes for **Expected retention time**. The value may be between `1.0` and `1.5` minutes.

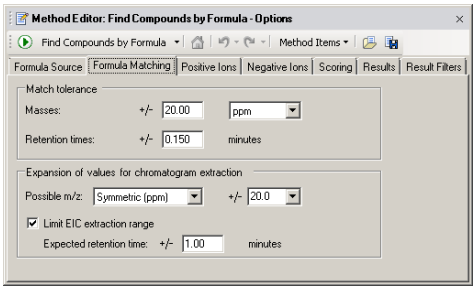


Figure 72 *Formula Matching tab in the Find by Formula - Options section*

Positive Ions tab

d Enter the parameters on the **Positive Ions** tab.

The parameters in this tab specify the positive ion adducts that the algorithm uses with the molecular formula to confirm that the feature was found in the data. Better results are derived from acquisition methodologies that minimize adducts, especially sodium and potassium.

1. Click the **Positive Ions** tab.
2. Mark the charge carriers **+H**, **+Na**, and **+K** that are known to be present in the data. Typically positive protonated is the ideal selection. Non-adducted molecular ions, loss of an electron, are an option in Find by Formula.
3. Enter the molecular formulas for specific charge carriers in the input box below the charge carriers selection.
4. Clear Neutral losses. Neutral losses are not typically used. An exception is when a facile loss is expected.
5. Enter the molecular formulas for specific neutral losses in the input box below the neutral losses selection.
6. Type `1` for **Charge state range**.
7. Clear the **Dimers** check box.
8. Clear the **Trimers** check box.

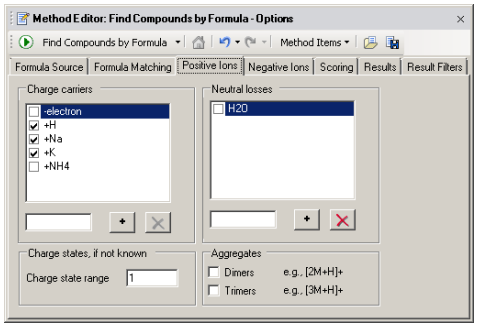


Figure 73 *Positive Ions tab in the Find by Formula - Options section*

Negative Ions tab

e Enter the parameters on the **Negative Ions** tab.

The parameters in this tab specify the negative ion adducts that the algorithm uses with the molecular formula to confirm that the feature was found in the data. Better results are derived from acquisition methodologies that minimize adducts.

1. Click the **Negative Ions** tab.
2. Mark the charge carrier **-H** that is known to be present in the data. Typically negative deprotonated is the ideal selection. Non-adducted molecular ions, attachment of an electron, are an option in Find by Formula.
3. Enter the molecular formulas for specific charge carries in the input box below the charge carriers selection.
4. Clear Neutral losses. Neutral losses are not typically used. An exception is when a facile loss is expected.
5. Enter the molecular formulas for specific neutral losses in the input box below the neutral losses selection.
6. Type `1` for **Charge state range**.
7. Clear the **Dimers** check box.

8. Clear the **Trimers** check box.

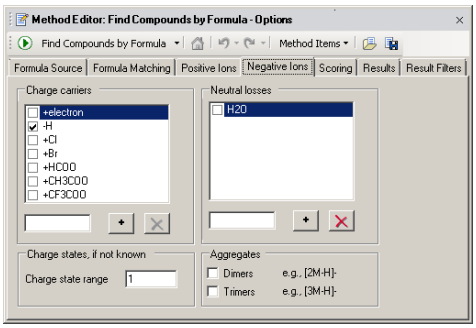


Figure 74 *Negative Ions tab in the Find by Formula - Options section*

Scoring tab

f Enter the parameters on the **Scoring** tab.

The parameters in this tab specify how to rate whether the spectral pattern is correct for the molecular formula. The scoring determines a goodness of fit between observed ions compared to the expected ions in the database. As signal levels decrease the scoring parameters entered may not match as well. The defaults provided are adequate.

1. Click the **Scoring** tab.
2. Type `100` for the **Mass score.**
3. Type `60` for the **Isotope abundance score**.
4. Type `50` for the **Isotope spacing score.**
5. Type `100` for the **Retention time score**.

Note: If you set values of 100 for the Mass score and 0 for Isotope abundance score, the Isotope spacing score, and the Retention time score, then the latter three scores are not included when calculating the Score.

6. Type the default values of `2.0` for mDa and `5.6` ppm for **MS mass**.
7. Type the default value of `7.5` % for **MS isotope abundance**.
8. Type the default values of `5.0` for mDa and `7.5` ppm for **MS/MS mass**.
9. Type the default value of `0.15` min for **Retention time**.

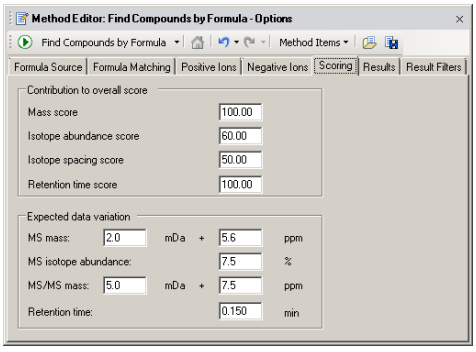


Figure 75 *Scoring tab in the Find by Formula - Options section*

Results tab

g Enter the parameters on the **Results** tab.

The parameters in this tab specify how the results are saved. This affects the ease reviewing results.

1. Click the **Results** tab.
2. Mark the **Delete previous compounds** check box to delete prior compound results. Clear the **Delete previous compounds** check box when you want to concatenate the Find Compounds by Formula results to the Find by Molecular Feature results and thereby manually review whether the feature was found in both instances.
3. Click **Highlight first compounds** under the New results group box.
4. Mark the **Extract EIC** check box.
5. Mark the **Extract cleaned spectrum** check box. Extracting chromatograms or spectra slows the processing. Once you are comfortable with the results, processing time is reduced by clearing these check boxes.
6. Clear the **Include structure** check box.
7. Clear the **Extract raw spectrum** check box.
8. Clear the **Extract MS/MS spectrum** check box.

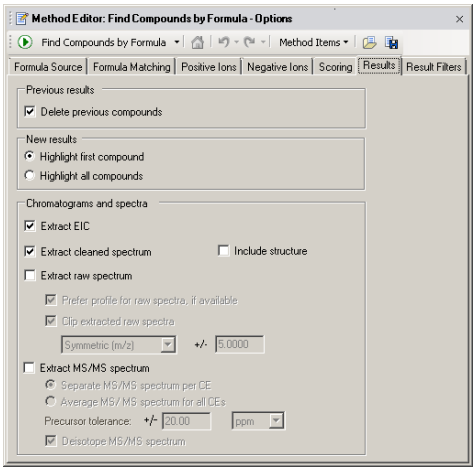


Figure 76 *Results tab in the Find by Formula - Options section*

Result Filters tab

h Enter the parameters on the **Result Filters** tab.

The parameters in this tab specify whether compounds are generated and/or whether you receive warning notations based on the matching score. Mass Profiler Professional does not need the matching.

1. Click the **Result Filters** tab.
2. Clear the **Only generate compounds for matched formulas** check box.
3. Mark the **Warn if score is** check box.
4. Type 75 for Warn if score **<**.
5. Clear the **Do not match if score is** check box.
6. Mark the **Warn if the second ion's expected abundance is** check box.
7. Type 50 for Warn if the second ion's expected abundance is **>**.
8. Clear the **Do not match if the second ion's expected abundance is** check box.

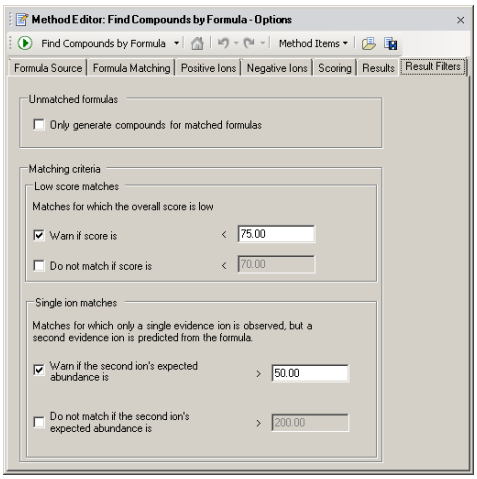


Figure 77 *Result Filters tab in the Find by Formula - Options section*

2. Enter the parameters in the Find by Formula - Chromatograms section.

In this section, you enter integrator parameters that are applied to the data to extract features for matching to the input formula. This is the second and most critical part of the recursive refinement of the feature finding.

a Click **Find Compounds by Formula > Find by Formula - Chromatograms** in the Method Explorer window. The input options specify the integrator parameters.

EIC Smoothing tab

b Enter the parameters on the **EIC Smoothing** tab.

The parameters in this tab specify which algorithm to use to smooth the extracted ion chromatogram results.

1. Click the **EIC Smoothing** tab.
2. Select **Gaussian** from the Smoothing function selection.
3. Type `15` for **Function width** points.
4. Type `5` for **Gaussian width** points.

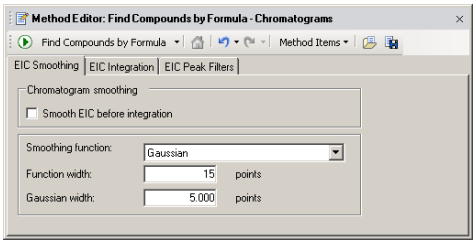


Figure 78 *EIC Smoothing tab in the Find by Formula - Chromatograms section*

EIC Integration tab

c Enter the parameters on the **EIC Integration** tab.

The parameters in this tab specify which integrator to use for the data extraction.

1. Click the **EIC Integration** tab.
2. Select **Agile** under the Integrator selection. Agile is the preferred metabolomics integrator. No additional user parameters are associated with this integration.

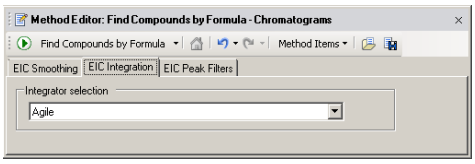


Figure 79 *EIC Integration tab in the Find by Formula - Chromatograms section*

The General integrator may be selected as an alternate. When selected, an **Options** tab is presented below the Integrator selection heading.

Under the Detector heading:

1. Type `2` for **Point sampling**, `0.02` for **Start threshold**, and `0.1` for **Stop threshold**.
2. Select **7 point** for **Filtering** and **Top** for **Peak location**.

Under Baseline allocation:

3. Type `5` for **Baseline reset >** and `100` % for **If either edge <**.
4. Click **Drop else tangent skim**.

The other integrator integration selections are not recommended for metabolomics.

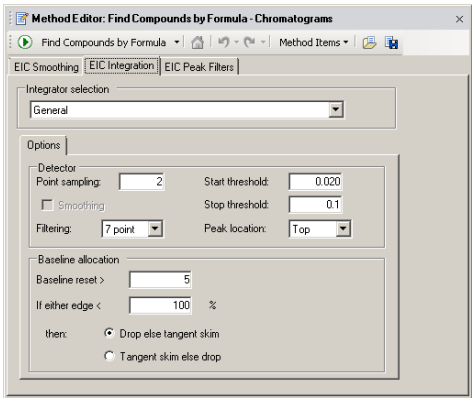


Figure 80 *General integrator selection in the EIC Integration tab*

EIC Peak Filters tab

d Enter the parameters on the **EIC Peak Filters** tab.

The parameters in this tab specify which ions to filter out of the chromatogram integrator results.

1. Click the **EIC Peak Filters** tab.
2. Click **Peak height**.
3. Mark the **Absolute height** check box.
4. Type in a value of `1000` counts for the **Absolute Height**. With targeted feature finding, the minimum **Absolute height** of the feature may be smaller than the absolute height used in the compound filters for untargeted molecular feature extraction.
5. Mark the **Limit (by height) to the largest** check box.
6. Type in a value of `5` for the **Limit (by height) to the largest**. If more than five peaks are found and pass the isotope test and retention times are not used, then the most abundant peaks are reported.

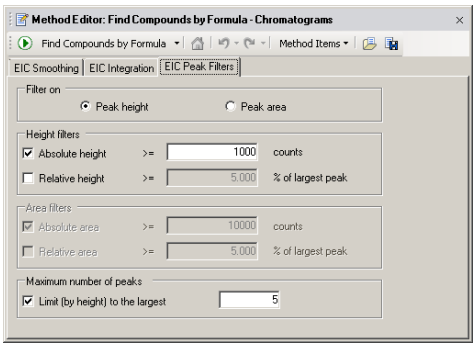


Figure 81 *EIC Peak Filter tab*

3. Enter the parameters in the Find by Formula - Mass Spectra section.

The parameters in the Find by Formula - Mass Spectra section are applied to the data to extract features for matching to the input formula. This is the third and final part of the recursive refinement of the feature finding.

a Click **Find Compounds by Formula > Find by Formula - Mass Spectra** in the Method Explorer window. The input options specify criteria for mass spectra to include in the feature processing.

Peak Spectrum tab

b Enter the parameters in the **Peak Spectrum** tab.

The parameters in this tab specify which spectra from the extracted ion chromatograms to include in the feature processing and whether to perform background subtraction on the spectra.

1. Click the **Peak Spectrum** tab
2. Click **Average scans >**.
3. Type `10` for the **% of peak height**. Averaging scans provides mass accuracy.
4. Clear the **Exclude if above X% of saturation** under the TOF spectra group box. If this check box is marked, any spectrum containing a peak within the given percentage of being saturated is excluded from processing for any compound feature.
5. Select **None** for **MS** under the Peak spectrum background group box.

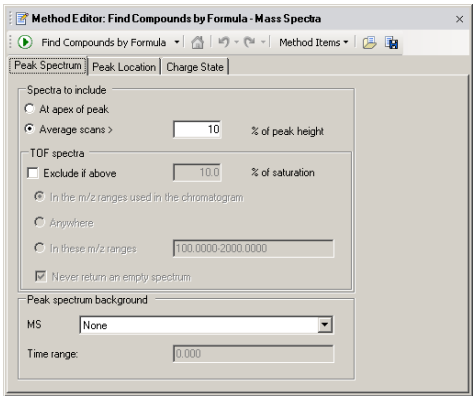


Figure 82 *Peak Spectrum tab in the Find by Formula - Mass Spectra section*

Peak Location tab

c Enter the parameters on the **Peak Location** tab.

The parameters in this tab specify the *m/z* values in a spectrum that are considered peaks. These parameters are only applicable to profile data files. They are not applicable to centroid collected data files and may be left at the defaults.

1. Click the **Peak Location** tab.
2. Type the default of `2` for **Maximum spike width**.
3. Type the default of `0.70` for **Required valley**.

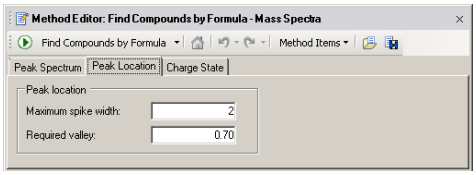


Figure 83 *Peak Location tab in the Find by Formula - Mass Spectra section*

Charge State tab

d Enter the parameters on the **Charge State** tab.

The parameters in this tab specify isotope grouping tolerances and charge state limits. It is important to set the maximum charge state to one (1). Adjustments to the grouping model can change the compound results.

1. Click the **Charge State** tab.
2. Type `0.0025` for *m/z* and `7.0` ppm for the **Peak spacing tolerance**.
3. Select **Common organic molecules** for the **Isotope model**. You select **Unbiased** if the compounds are known to contain metals.
4. Mark the **Limit assigned charge state to a maximum of** check box.
5. Type `1` for the **Limit assigned charge state to a maximum of**. This parameter should match the value typed into the Charge state range in the "Positive Ions tab" on page 101 and "Negative Ions tab" on page 101 in the "Enter the parameters in the Find by Formula - Options section.".
6. Clear the **Treat ions with unassigned charge as singly-charged** check box.

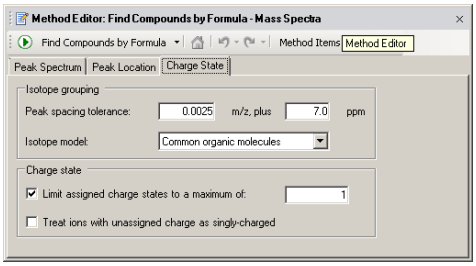


Figure 84 *Charge State tab in the Find by Formula - Mass Spectra section*

4. Turn off the sample purity calculations in the Find by Formula - Sample Purity section.

The Find by Formula - Sample Purity is not used in metabolomics analyses.

a Click **Find Compounds by Formula > Find by Formula - Sample Purity** in the Method Explorer window. The input options specify criteria for mass spectra to include in the feature processing.

b Turn off sample purity calculations on the Options tab.

1. Click the **Options** tab.
2. Clear the **Compute sample purity** check box. If this check box is cleared, then sample purity is not calculated. All other options on this tab are unavailable and the entries in the remaining tabs are not considered.

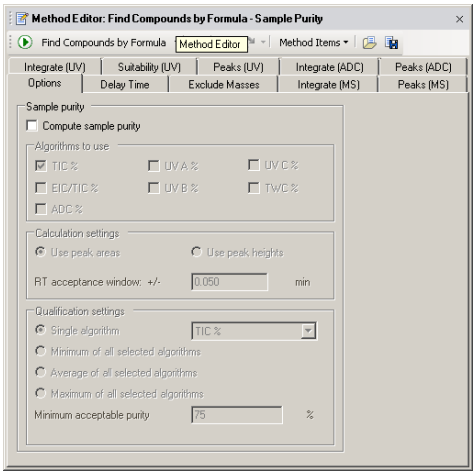


Figure 85 *Options tab in the Find by Formula - Sample Purity section*

Save your Find Compounds by Formula method

After you have edited your FbF method it is recommended you save the method using a name different from the name you previously used for your MFE method so that you can readily reprocess your data or new data without having to edit the Worklist Automation actions.

a Click **Method > Save As**.

b Select the folder and type a method name in the **Save Method** dialog box. It is recommended to change the `MFE` at the end of the method file name used in "Save your Find Compounds by Molecular Feature method" on page 50 to `FbF`.

c Click **Save**.

Set the Export CEF Options

1. Open the Method Editor for exporting CEF options.

2. Enter the export destination settings for your method.

Export CEF Options specifies where MassHunter DA Reprocessor stores the resulting .CEF feature files and whether the files replace or overwrite any prior files.

---

a Click **Export** from within the **Method Explorer** window.

b Click **CEF Options**.

---

a Click **At the location of the data file**.

b Click **Auto-generate new export file name**.

c Save your method. Click the save method icon or click **Method > Save**.

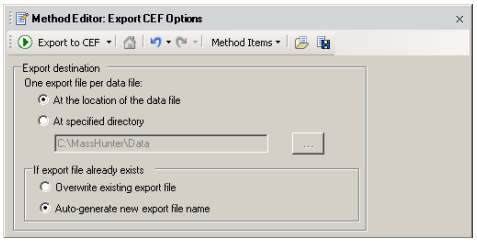


Figure 86 *Export CEF Options for use with DA Reprocessor*

Enable the method to run in MassHunter DA Reprocessor

1. Open the Method Editor to assign actions to run from the worklist.
2. Replace the **MFE** action with **FbF** action.

Agilent MassHunter software can most efficiently perform computationally intensive tasks, such as feature finding, on multiple data files using MassHunter DA Reprocessor. The following steps enable your method to run using DA Reprocessor.

Click **Worklist Automation** from within the **Method Explorer** window.

a Click **Worklist Actions**.

---

a Double-click on the **Find Compounds by Molecular Feature** action in the **Actions to be run** list. The action is automatically removed from the **Actions to be run** list. As an alternate to the double-click, you can click on the action and then click the delete icon  .

b Double-click on the **Find Compounds by Formula** action in the **Available actions** list. The action is automatically added to the **Actions to be run** list. As an alternate to the double-click, you can click the action and then click the down arrow button  to add the action to the **Actions to be run** list.

c Move the actions in the **Available actions** list so that the **Export to CEF** action is listed after the **Find Compounds by Formula** action as shown in Figure 87.

d Save your method. Click the save method icon  or click **Method > Save**.

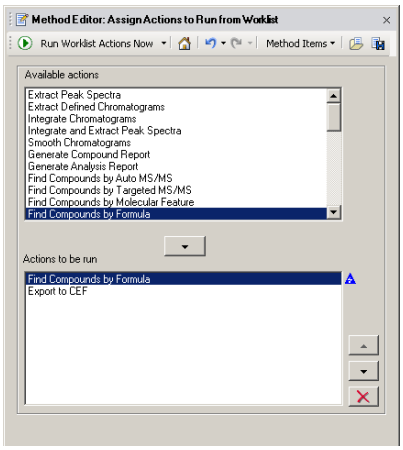


Figure 87 *Assign Actions to Run from Worklist for use with DA Reprocessor*

Confirm the FbF method on a single data file

Metabolomics involves the analysis of a large number of sample files with each sample containing a large number of compounds. Find Compounds by Formula is therefore run on the entire metabolomics sample set using MassHunter DA Reprocessor. However, before the entire sample set is run in MassHunter DA Reprocessor, a single file is processed within MassHunter Qualitative Analysis to verify the new parameters.

1. Find Compounds by Formula on a single sample.

a Click **File > Open Data File**.

b Click on a single data file in the **Open Data File** dialog box.

c Click **Open**.

d Click **Actions > Find Compounds by Formula**, or click the Find Compounds by Formula button  in the **Method Editor: Find Compounds by Formula** section in the Method Editor window. Feature extraction begins immediately and the progress is shown in an Operation in Progress status box (see Figure 28 on page 53).

If no data file is open, or an inappropriate data file is open, a message box appears as shown in Figure 29 on page 53. Click **OK** and open a single data file.

2. Display and review the .

When the FbF method finishes processing the data file, MassHunter Qualitative Analysis displays the results in several windows. You may review and arrange the results to meet your preferences.

If the window is not displayed, click **View >** .

The options for reviewing the are identical to those described in the first metabolomics workflow step "Find features", step 2 - "Display and review the Compound List." on page 53.

3. *Optional* - Export the results for the single sample to a CEF file.

This step is optional. The CEF files for all of the samples are generated in step 1 - "Find Compounds by Formula using MassHunter DA Reprocessor."

a Click **File > Export > as CEF**. The Export CEF Options dialog box is opened.

b Select the data files to be exported from the **List of opened data files**. It is recommended to create a new folder for the exported CEF files to aid documentation of the metabolomics workflow and to make it easier to distinguish any new CEF files from previous CEF files.

c Update the other parameters in the **Export CEF Options** dialog box.

d Click **OK**.

You can review the results from this step by importing the CEF back into MassHunter Qualitative Analysis by following the procedure presented in step 2 - "Display and review the after running MassHunter DA Reprocessor." below.

Find compounds using DA Reprocessor

1. Find Compounds by Formula using MassHunter DA Reprocessor.
2. Display and review the after running MassHunter DA Reprocessor.

Metabolomics involves applying your method processing to a large number of sample files whereby each sample file may contain a large number of compounds. MassHunter Qualitative Analysis can be used to process all of your data sets. However, MassHunter DA Reprocessor provides a more efficient and automated means to run your MassHunter Qualitative Analysis method on multiple sample files. Therefore your method is run on the entire metabolomics sample set using DA Reprocessor.

Follow the same procedure presented in "Find features", "Find compounds using DA Reprocessor" on page 56.

---

a Return to MassHunter Qualitative Analysis. If you closed the MassHunter Qualitative Analysis program, do the following:
   - Click **Start > All Programs > Agilent > MassHunter Workstation > Qualitative Analysis B.05.00**.
   - Click **Cancel** when the **Open Data File** dialog box opens.

b Click **File > Close All** to close the open data files. Do not save any results.

c Click **File > Import Compound** to open one of the CEF files that contains the molecular feature results of Find Compounds by Formula.

   Note: Because of the large number of features in a typical metabolomics sample file, it is recommended to open only one file at a time to review the results. Close the open file and then open the next file.

d Follow the same procedure presented in "Find features", step 2 - "Display and review the Compound List." and step 3 - "Extract results." on page 54 to display and review the and the mass spectral results.

Import and organize your data

After you set up a new project and experiment, Mass Profiler Professional provides a "MS Experiment Creation Wizard" that guides you through the operations necessary to organize and prepare your FbF data for analysis. The preparation includes project set up, experiment creation, filtering, alignment, normalization, and baselining. Since the advanced operations available in the Workflow Browser do not guide you through the initial steps of data import and differential analysis, it is not recommended to skip the "Data Import" or "Analysis: Significance Testing and Fold Change" steps of the metabolomics workflow.

Bookmark this location in the workflow and follow the procedures presented in "Import and organize data" on page 59 then return to the next step of this workflow.

Create an initial analysis

Create a differential expression from your recursively found features. The Significance Testing and Fold Change workflow helps you create an initial differential expression from your data and identify the most significant features from among all of the features found using FbF. The steps necessary to create your initial differential expression are predetermined and based on the experiment type, experiment grouping, and conditions you entered when creating your project and setting up your experiment.

Bookmark this location in the workflow and follow the procedures presented in "Create an initial analysis" on page 79 then return to the next step of this workflow.

Save your project

If you did not save your current analysis at the end of "Create an initial analysis", save your current analysis at this time.

a Click **Project > Export Project**.

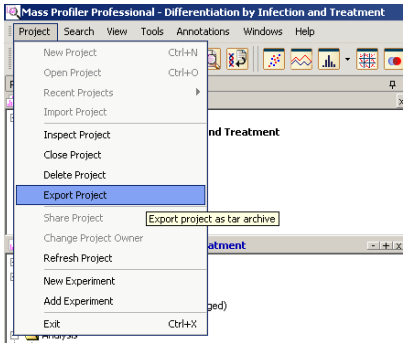


Figure 88 *Menu selection to export your current analysis*

b Mark the check box next to the experiment you wish to save.

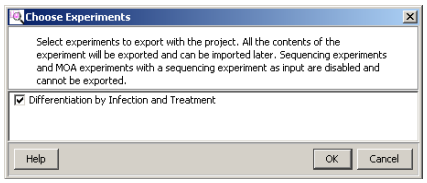


Figure 89 *Choose Experiments dialog box for saving your experiment*

c Click **OK**.

d Select or create the file folder.

e Type the **File name**.

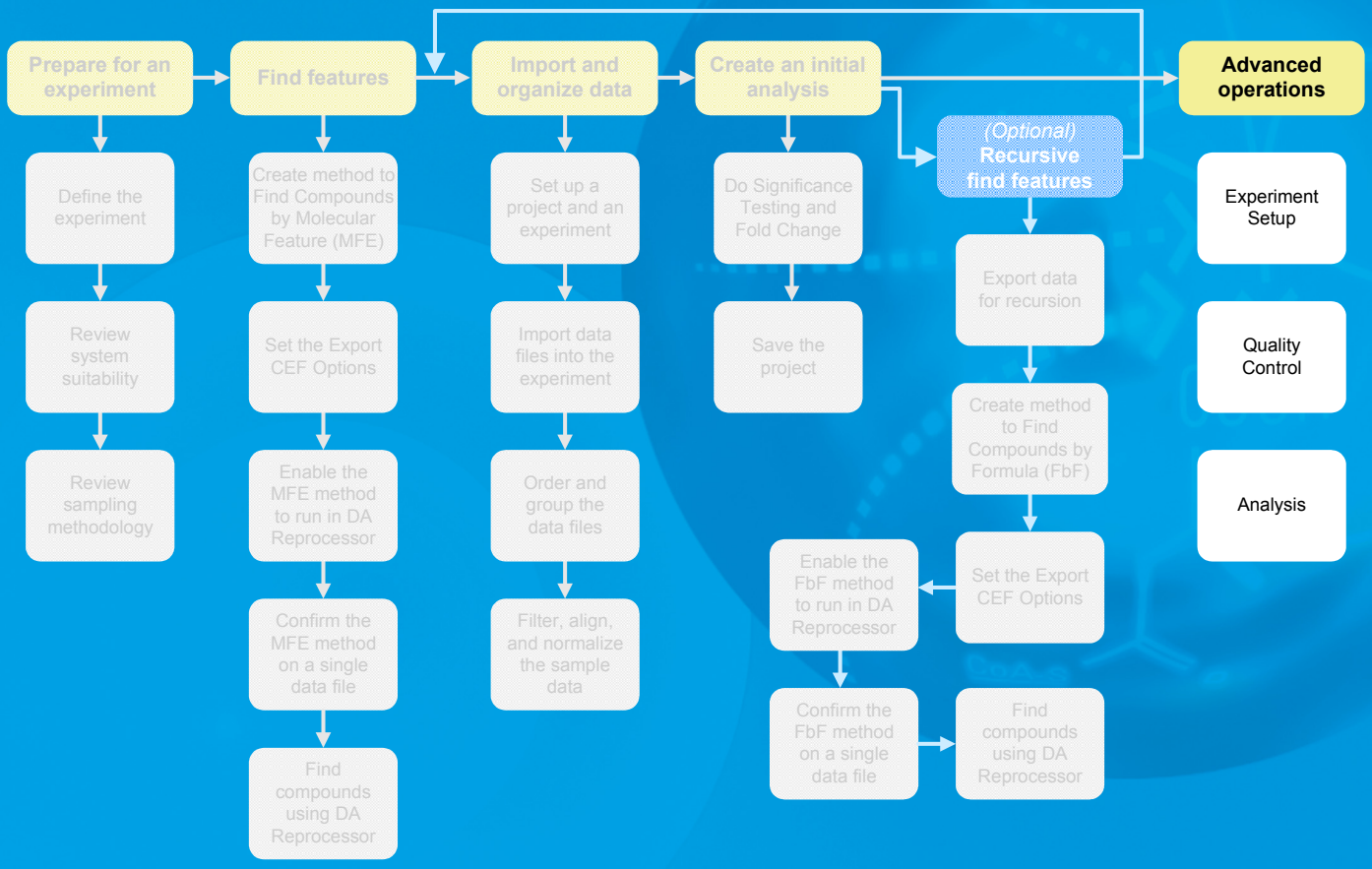
f Click **Save**.

Next step...

You have now completed the fifth step of the metabolomics workflow. In the next workflow step you review the options available in Mass Profiler Professional for advanced analysis.

Advanced operations

The most significant features in your data are processed by Mass Profiler Professional into a final statistical analysis and interpretation. The results from the final interpretation may be used to prove or disprove your hypothesis and may be used to create a sample class prediction model or evaluate pathways.



Agilent Technologies

Overview of advanced operations

The operations available in the Workflow Browser of Mass Profiler Professional provides the tools necessary for analyzing features from your mass spectrometry data depending upon the need and aim of the analysis, the experimental design and the focus of the study. This helps you create different interpretations to carry out the analysis based on the different filtering, normalization, and standard statistical methods.

It is recommended that you follow the procedures in the prior steps of this workflow before proceeding with the operations available in advanced operations. When you click **Finish** during "Create an initial analysis" Mass Profiler Professional automatically makes the operations available under the Workflow Browser, letting you have access to all available operations. Click **Cancel** to stop the workflow at any step in "Create an initial analysis" and immediately enter the Workflow Browser.

Only some of the operations available in the Workflow Browser mode are documented in this section. More information regarding the operations available in the Workflow Browser may be found in the *Mass Profiler Professional User Manual*.

Layout of the Mass Profiler Professional screen

You are now in the advanced workflow mode and have access to all features available in Mass Profiler Professional through the Workflow Browser (see Figure 90).

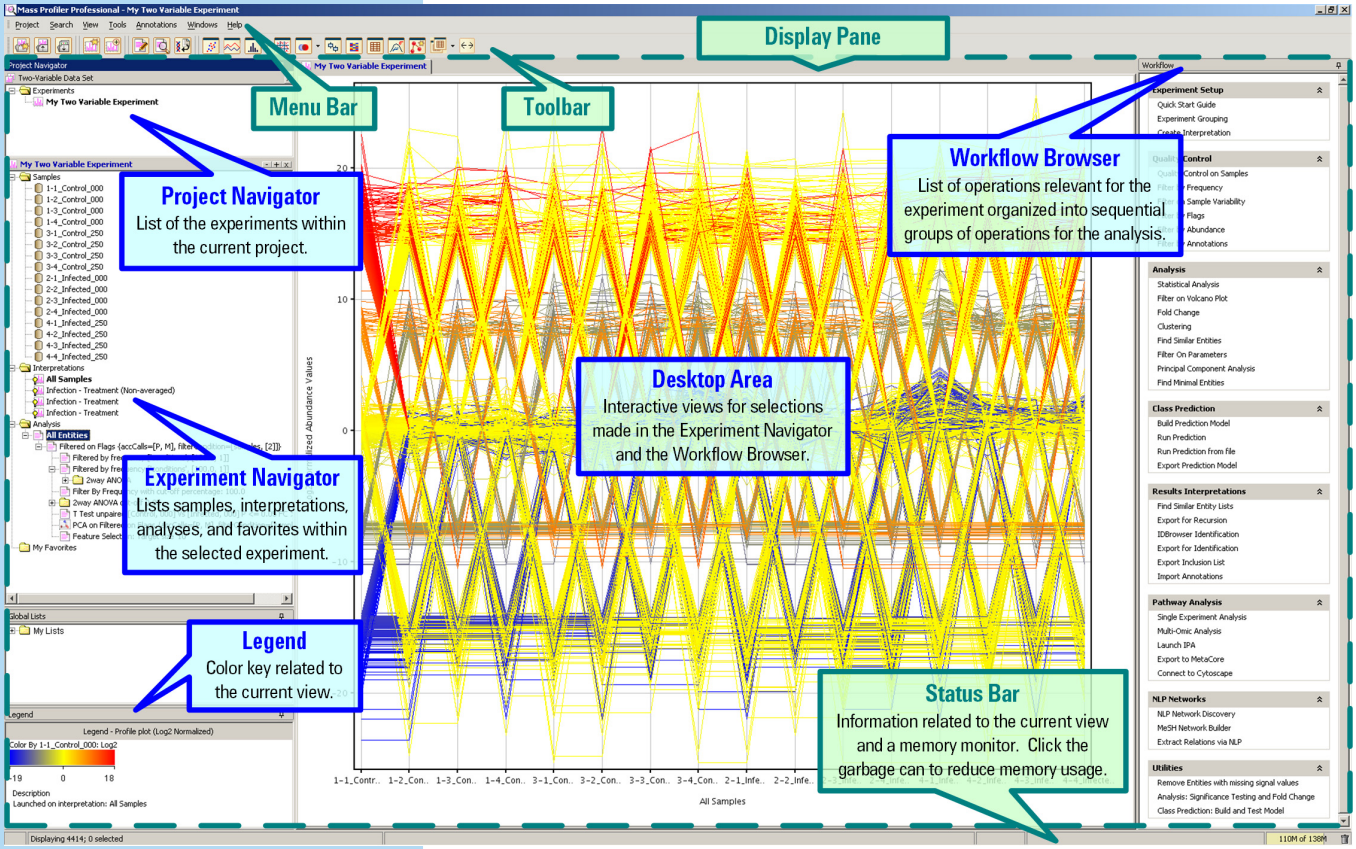
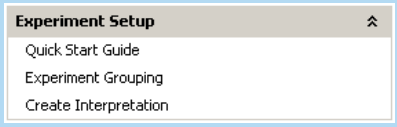


Figure 90 *The main functional areas of Mass Profiler Professional illustrated using the "Two-variable experiment" data set*

Experiment Setup



Experiment Setup lets you review a quick start guide to Mass Profiler Professional, define the sample grouping and replicate structure of your experiment, and specify how the samples are grouped into experimental conditions for visualization and analysis. Experiment setup consists of three operations:

Quick Start Guide

a Click **Quick Start Guide** in the Workflow Browser. The *Mass Profiler Professional Quick Start Guide* opens in your Internet browser.

b Review the *Mass Profiler Professional Quick Start Guide*.

You can also download current Quick Start Guides from the Agilent Literature Library:

Agilent G3835AA MassHunter Mass Profiler Software - Overview and Data Import Quick Start Guide (Agilent publication G3835-90004, Revision A, January 2012)

Agilent G3835AA MassHunter Mass Profiler Software - Significance Testing and Fold Change Quick Start Guide (Agilent publication G3835-90003, Revision A, January 2012)

Experiment Grouping

a Click **Experiment Grouping** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Enter your sample grouping by following the same process presented in "Group samples based on the independent variables and replicate structure of your experiment in the MS Experiment Creation Wizard (Step 6 of 11)." on page 67.

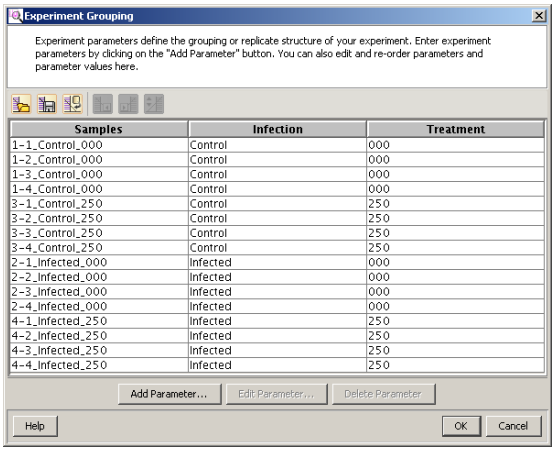


Figure 91 *Experiment Grouping dialog box*

c Click **OK**.

Create Interpretation

An Interpretation specifies how your samples are grouped into experimental conditions for display and tratement by your analysis. All samples with the same parameter value are grouped into an experimental condition.

a Click **Create Interpretation** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Enter parameters on the **Select Parameters** page (Create Interpretation (Step 1 of 4)).

1. Mark the experiment parameters you would like to use as the basis for display and analysis. The experiment parameters are the independent variables in your experimental design.
2. Click **Next**.

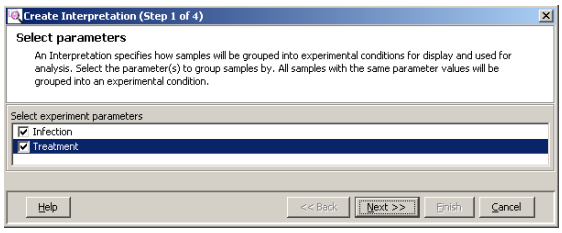


Figure 92 *Select parameters, Create Interpretation (Step 1 of 4)*

c Enter parameters on the **Select Profile Plot Display Modes** page (Create Interpretation (Step 2 of 4)).

1. Click either **Numerical** or **Categorical** for the display mode for each of the parameters you marked in the prior step.
2. Click **Next**.

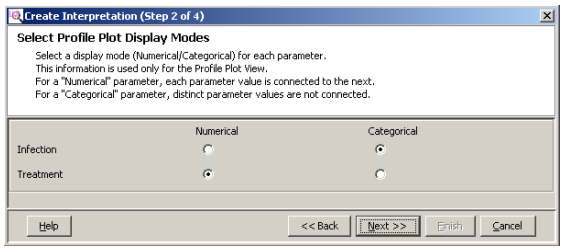


Figure 93 *Select Profile Plot Display Modes, Create Interpretation (Step 2 of 4)*

d Update parameters on the **Select conditions** page (Create Interpretation (Step 3 of 4)).

1. Mark the conditions that you would like to be included in the interpretation. The conditions are the attribute values of the independent variable or permutation of attribute values when more than one independent variable is selected in the Select Parameters step.
2. Click **Averaged** to average the replicates within a condition. A single value based on the average of the entity across the replicates of the condition is presented for each entity with respect to each condition.
3. Mark the measurement permitted flags. A flag is a term used to denote the quality of an entity within a sample; whether the entity is present or absent in the sample. *Present* means the entity was detected, *Absent* means the entity was not detected, and *Marginal* means the signal for the entity was saturated.
4. Click **Next**.

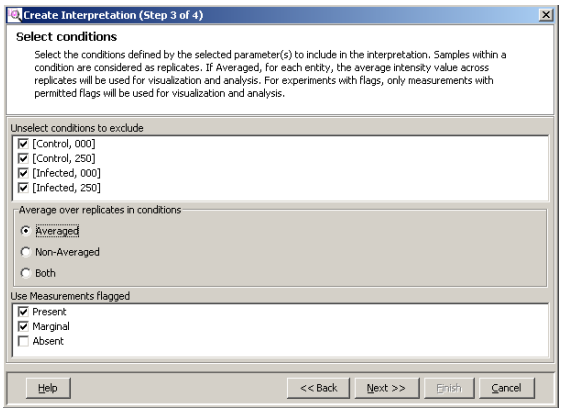


Figure 94 *Select conditions, Create Interpretation (Step 3 of 4)*

e Update parameters on the **Save Interpretation** page (Create Interpretation (Step 4 of 4)).

1. Add or edit details to save as part of the interpretation in the **Notes** box.
2. Review the interpretation summary and tabs.
3. Click **Back**, make changes to the parameters, and click **Next** to return to this view.
4. Click **Finish** when the interpretation parameters are complete.

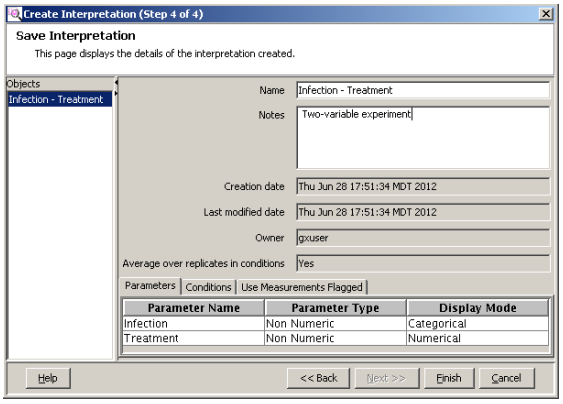
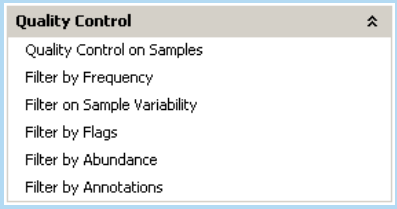


Figure 95 *Save Interpretation, Create Interpretation (Step 4 of 4)*

Quality Control



Quality control lets you decide which samples are ambiguous, or outliers, and which pass the quality criteria. Quality control provides rich, interactive, and dynamic tools to visualize and examine the quality of your data. Quality control consists of six operations:

Quality Control on Samples

a Click **Quality Control on Samples** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

Access to display and printing of the Correlation Coefficients spreadsheet, Correlation Plot, and 3D PCA Scores views. Available options to manipulate the view is described in the "Review the sample quality in QC on samples in the Analysis: Significance Testing and Fold Change (Step 5 of 8) workflow." on page 88.

b Click **Add/Remove Samples** to limit the samples to those that meet the quality criteria based on he PCA scores.

c Click **OK**.

d Click **Close** to complete the quality control.

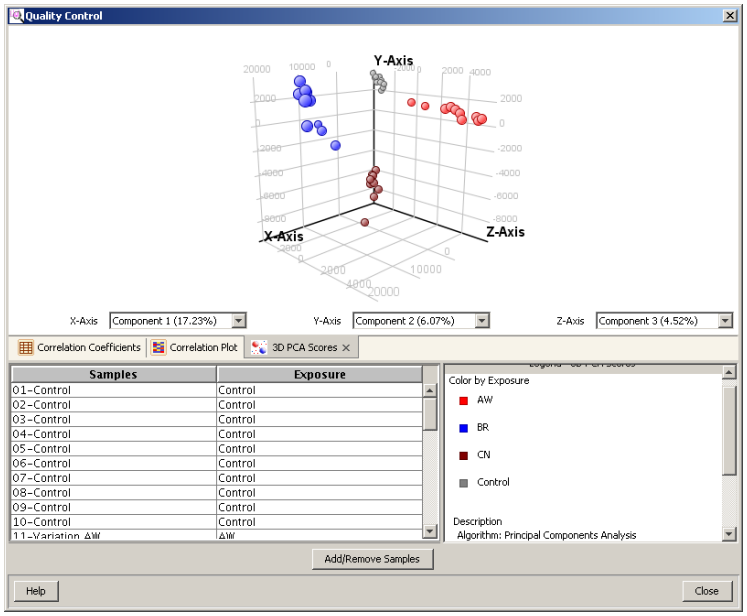


Figure 96 *Quality Control on Samples*

Filter by Frequency

You can filter entities based on the number of times an entity occurs in the samples. The new entity list is placed in the Analysis folder within the Experiment Navigator.

a Click **Filter by Frequency** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Select your entity list and interpretation on the **Entity List and Interpretation** page (Filter by Frequency (Step 1 of 4)):
   1. Click **Choose** to select the **Entity List.**
   2. Click **Choose** to select the **Interpretation**.
   3. Click **Next**.

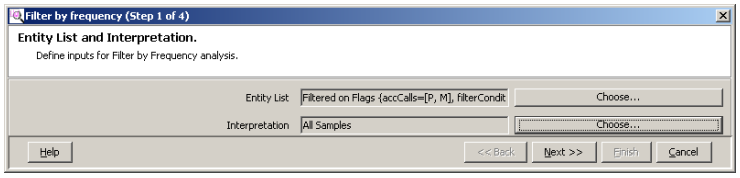


Figure 97 *Entity List and Interpretation, Filter by Frequency (Step 1 of 4)*

c Enter parameters on the **Input Parameters** page (Filter by Frequency (Step 2 of 4)):
   1. Type the minimum percentage of samples in which an entity must be present in order to pass the filter.
   2. Select the applicable condition of the samples for which each entity must be present.
   3. Click **Next**.

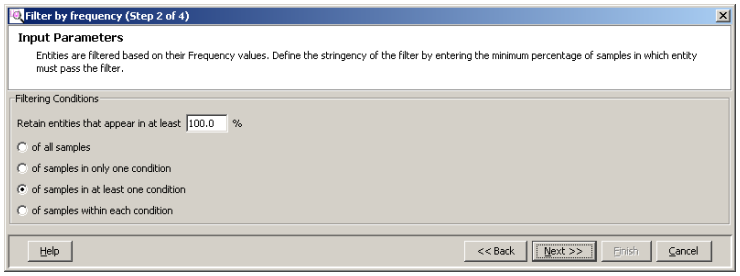


Figure 98 *Input Parameters, Filter by Frequency (Step 2 of 4)*

d Review your filter results on the **Output Views of Filter by Frequency** page (Filter by Frequency (Step 3 of 4)):
   1. Review the entities that passed the filter conditions using the spreadsheet view or the profile plot view. The total number of entities and number of entities passing the filter, along with their filtering condition, are displayed.
   2. Click **Back**, make changes to the parameters, and click **Next** to return to this view.
   3. Click **Next**.

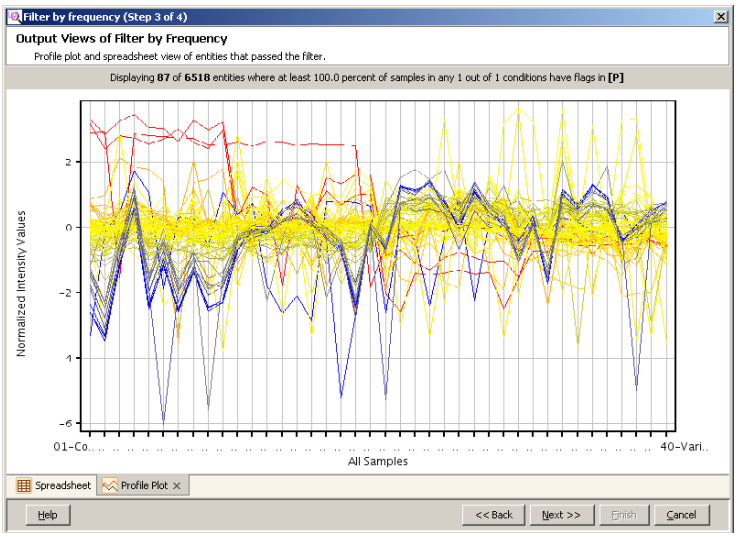


Figure 99 *Output Views of Filter by Frequency, Filter by Frequency (Step 3 of 4)*

The **Save Entity List** page

e Enter parameters on the **Save Entity List** page (Filter by Frequency (Step 4 of 4)):

The review content and parameters you enter in the **Save Entity List** page are the same for many operations within Mass Profiler Professional. The figures and description presented in this step are identical to those in other operations. You are referred back to this section when you are prompted to save your entity list at the completion of other operations available in the Workflow Browser.

1. Add or edit descriptive information that is stored with the saved entity list in the **Name**, **Notes**, and **Experiments** fields.
2. Click **Configure Columns** to add/remove and reorder the columns in the tabular presentation of the entities. This opens the **Select Annotation Columns** dialog box.
3. Select column items to add or to remove from the saved entity list.
4. Reorder the selected columns to your preference.
5. Mark **Save as Default** if you would like this configuration to be saved as the default for future save entity list steps.
6. Select the experiment type for your configuration to be applied.
7. Click **OK**.
8. Click **Finish**.

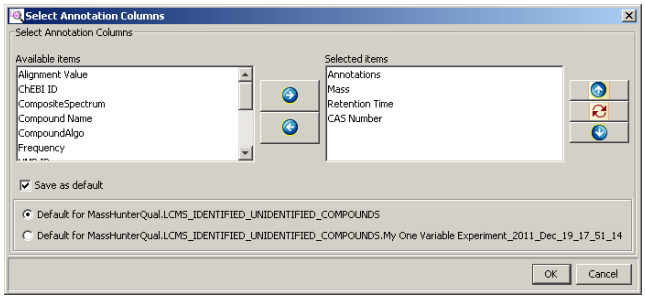


Figure 100 *Select Annotation Columns dialog box*

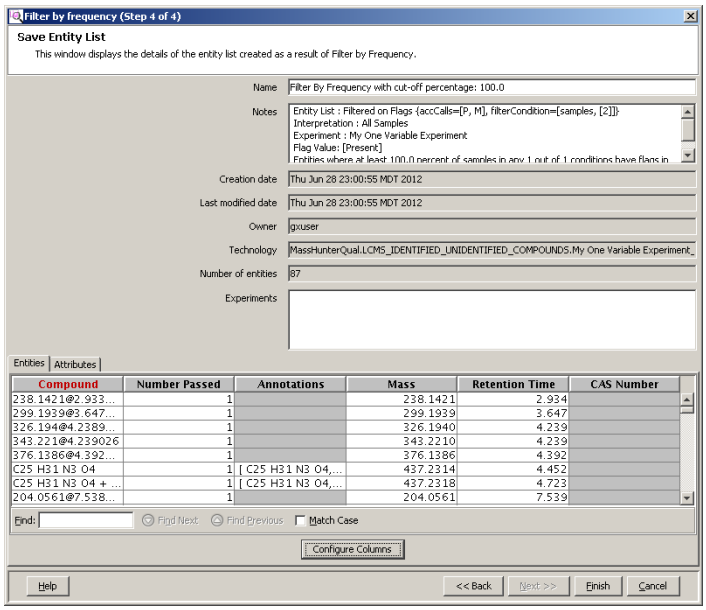


Figure 101 *Save Entity List is similar for many operations*

Filter on Sample Variability

Filter entities based on the standard deviation or coefficient of variation of intensity values within a condition. The new entity list is placed in the Analysis folder within the Experiment Navigator.

a Click **Filter on Sample Variability** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Select your entity list and interpretation on the **Entity List and Interpretation** page (Filter on Sample Variability (Step 1 of 4)):
   1. Click **Choose** to select the **Entity List.**
   2. Click **Choose** to select the **Interpretation**.
   3. Click **Next**.

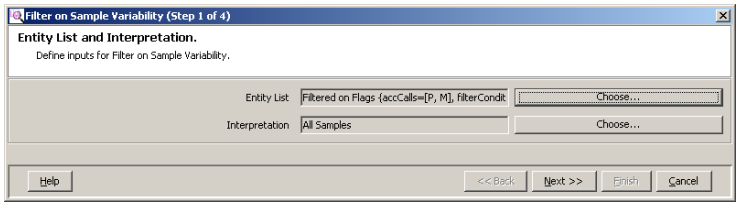


Figure 102 *Entity List and Interpretation, Filter on Sample Variability (Step 1 of 4)*

c Enter parameters on the **Input Parameters** page (Filter on Sample Variability (Step 2 of 4)):
   1. Select **Coefficient of variation <**.
   2. Type `50` **%** for the percent used for the filter.
   3. Type a value for the number of conditions that an entity must meet in **Retain entities in which at least ___ out of # conditions have values within range**. The default value is `1`.
   4. Click **Next**.

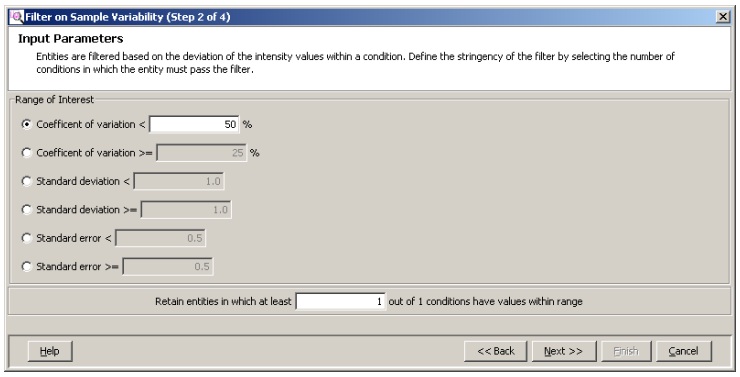


Figure 103 *Input Parameters, Filter on Sample Variability (Step 2 of 4)*

d Review your filter results on the **Output Views of Filter on Sample Variability** page (Filter on Sample Variability (Step 3 of 4)):

1. Review the entities that passed the filter conditions using the spreadsheet view or the profile plot view. The total number of entities and number of entities passing the filter, along with their filtering condition, are displayed.
2. Click **Back**, make changes to the parameters, and click **Next** to return to this view.
3. Click **Next**.

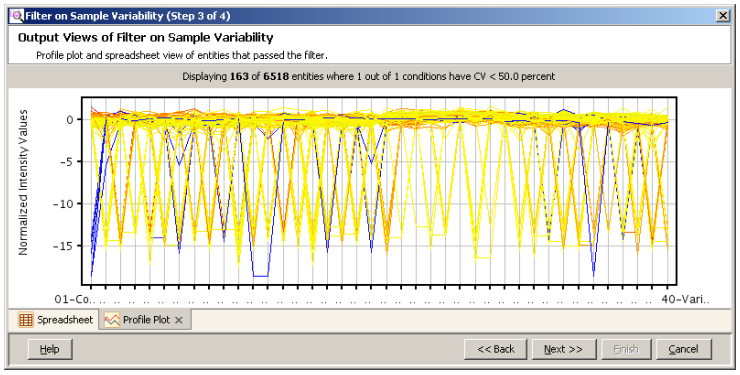


Figure 104 *Output Views of Filter on Sample Variability, Filter on Sample Variability (Step 3 of 4)*

e Enter parameters on the **Save Entity List** page (Filter on Sample Variability (Step 4 of 4)):

1. Follow the steps presented in "The Save Entity List page" on page 124.
2. Click **Finish**.

Filter by Flags

The main goal of Filter by Flags is to remove "one-hit wonders" from further consideration. A one-hit wonder is an entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis. The new entity list is placed in the Analysis folder within the Experiment Navigator.

a Click **Filter by Flags** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Select your entity list and interpretation on the **Entity List and Interpretation** page (Filter by Flags (Step 1 of 4)):

1. Click **Choose** to select the **Entity List.**
2. Click **Choose** to select the **Interpretation**.
3. Click **Next**.

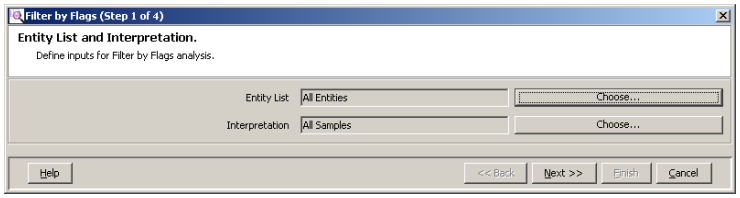


Figure 105 *Entity List and Interpretation, Filter by Flags (Step 1 of 4)*

c Enter parameters on the **Input Parameters** page (Filter by Flags (Step 2 of 4)):

1. Mark the **Present** check box.
2. Mark the **Marginal** check box.
3. Clear the **Absent** check box. This flag is useful when you want to identify entities that are missing in the samples. You can use this flag in conjunction with the **Next** and **Back** buttons to review the entities that are missing in some samples.
4. Click **at least ___ out of X samples have acceptable values**. The value "X" is replaced in your display with the total number of samples in your data set.
5. Type 2 in the entry box. By setting this parameter to a value of two or more, one-hit wonders are filtered.
6. Click **Next**.

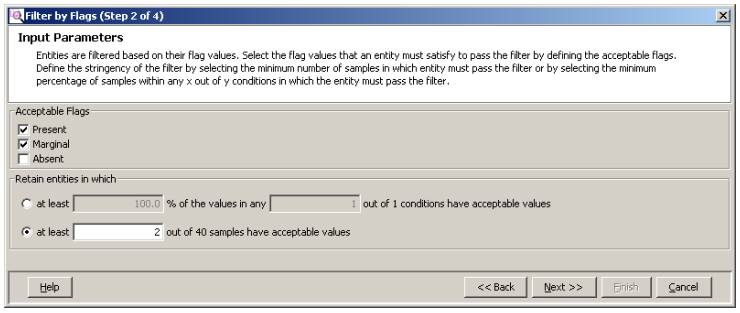


Figure 106 *Input Parameters, Filter by Flags (Step 2 of 4)*

d Review your filter results on the **Output Views of Filter by Flags** page (Filter by Flags (Step 3 of 4)):

1. Review the entities that passed the filter conditions using the spreadsheet view or the profile plot view. The total number of entities and number of entities passing the filter along with their filtering condition are displayed.
2. Click **Back** button, make changes to the parameters, and click **Next** to return to this view.
3. Click **Next**.

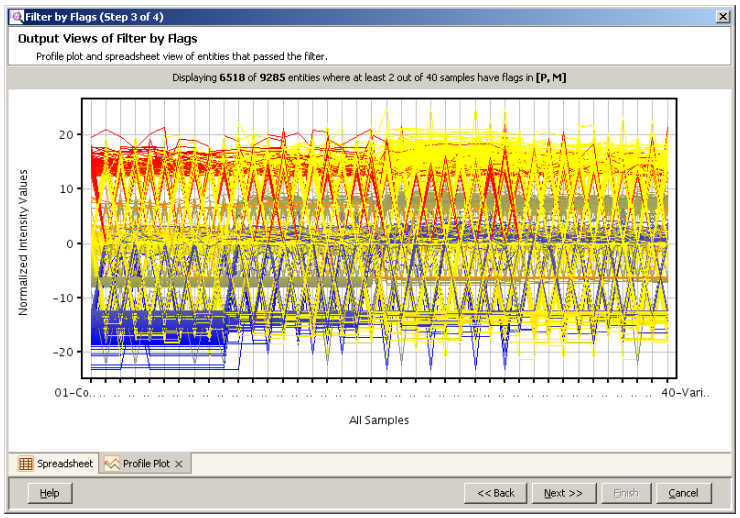


Figure 107 *Output Views of Filter by Flags, Filter by Flags (Step 3 of 4)*

e Enter parameters on the **Save Entity List** page (Filter by Flags (Step 4 of 4)):
   1. Follow the steps presented in "The Save Entity List page" on page 124.
   2. Click **Finish**.

Filter by Abundance

Filter out masses that have abundance values below the reliable detection limit of the mass spectrometer. You can set the proportion of conditions that must meet the criteria you specify. For example, to eliminate masses that do not meet a specified control value at least once in the experiment, you can filter them out by setting a minimum abundance value to be met in at least one condition. You can decide the proportion of conditions that must meet a certain threshold. The new entity list is placed in the Analysis folder within the Experiment Navigator.

a Click **Filter by Abundance** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Select your entity list and interpretation on the **Entity List and Interpretation** page (Filter by Abundance (Step 1 of 4)):
   1. Click **Choose** to select the **Entity List.**
   2. Click **Choose** to select the **Interpretation**.
   3. Click **Next**.

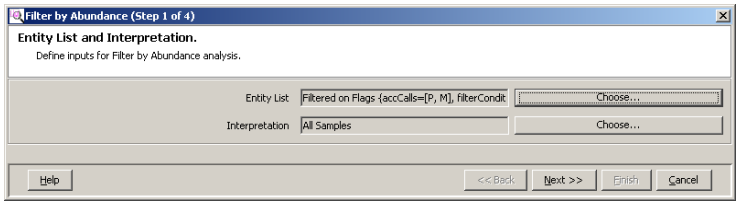


Figure 108 *Entity List and Interpretation, Filter by Abundance (Step 1 of 4)*

c Enter parameters on the **Input Parameters** page (Filter by Abundance (Step 2 of 4)):
   1. Select either **Raw Data** or **Normalized Data** for applying the filter.

2. Select either **Filter by Values** or **Filter by Percentile** with respect to the entity signal intensity (volume).
3. Assign values to the upper and lower cut-off range. For **Filter by Percentile** set the **Upper cut-off** value to `100` percent and the **Lower cut-off** value to `20` percent for the range of interest. Similarly, when using data values, set the upper and lower cut-off values so that the values selected pass approximately 80% of the samples based on their signal intensity.
4. Select retain entities in which **at least ____ of # samples have values within range**. Type `20` in the value.
5. Click **Next**.

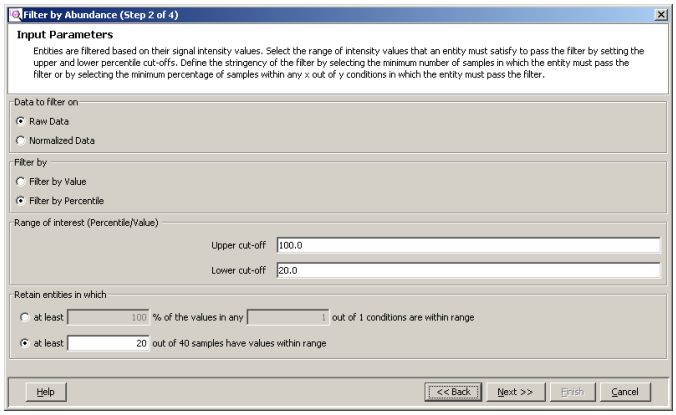


Figure 109 *Input Parameters, Filter by Abundance (Step 2 of 4)*

d Review your filter results on the **Output Views of Filter by Abundance** page (Filter by Abundance (Step 3 of 4)):

1. Review the entities that passed the filter conditions using the spreadsheet view or the profile plot view. The total number of entities and number of entities passing the filter along with their filtering condition are displayed.
2. Click **Back**, make changes to the parameters, and click **Next** to return to this view.
3. Click **Next**.

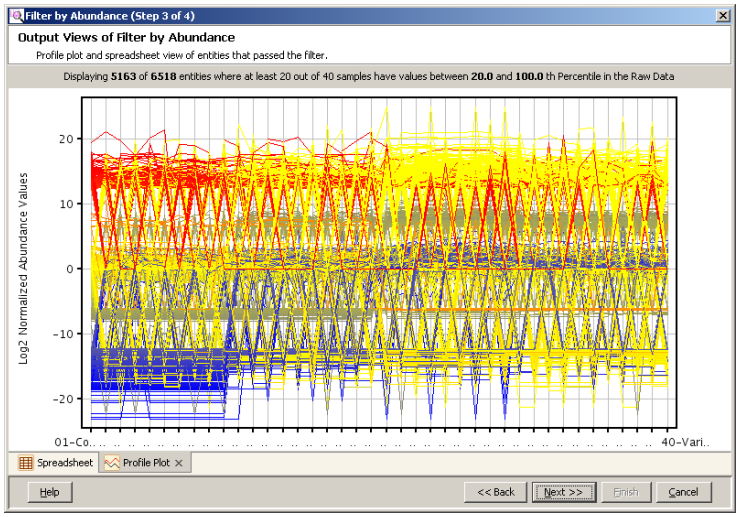


Figure 110 *Output Views of Filter by Abundance, Filter by Abundance (Step 3 of 4)*

e Enter parameters on the **Save Entity List** page (Filter by Abundance (Step 4 of 4)):

1. Follow the steps presented in "The Save Entity List page" on page 124.
2. Click **Finish**.

Filter by Annotations

Filter entities based on search fields, conditions, and search values that are part of the sample descriptions. The new entity list is placed in the Analysis folder within the Experiment Navigator.

a Click **Filter by Annotations** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment".

b Select your entity list on the **Input Parameters** page (Filter by Annotations (Step 1 of 4)):
   1. Click **Choose** to select the **Entity List.**
   2. Click **Next**.

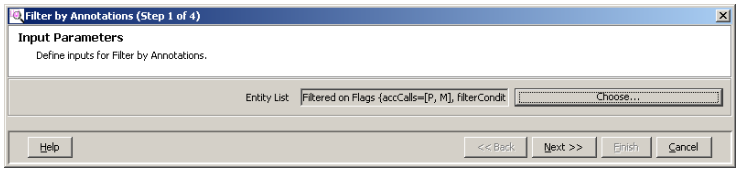


Figure 111 *Input Parameters, Filter by Annotations (Step 1 of 4)*

c Enter parameters on the **Add/Remove filter conditions** page (Filter by Annotations (Step 2 of 4)):
   1. Click the blank cell under the **Search Field** column to add a filter condition.
   2. Select an option from the **Search Field** drop-down list. Available selections for the search field are **Number Passed**, **Compound Name, Annotations**, **Mass**, **Retention Time**, or **CAS Number**.
   3. Select a condition from the **Condition** drop-down list. The available conditions depend on the search field selected in the prior step:
      - **=**, **≠**, **≥**, **≤**, or **in the range** are available when the Search Field value is **Number Passed**, **Mass**, or **Retention Time.**
      - **equals**, **does not equal**, **starts with**, **ends with**, or **includes** are available when the Search Field value is **Compound Name, Annotations**, or **CAS Number.**
   4. Enter the value in the **Search Value** cell. When the **Condition** selected is **in the range** two numerical values are entered separated by a comma.
   5. Click **Add** to add an additional search condition. Click **Remove** to remove search conditions you no longer need.
   6. For multi-line search queries, select **AND** or **OR** from the **Combine search conditions by** drop-down list.
   7. Click **Next**. If you do not add any valid search conditions all of the entities are selected.

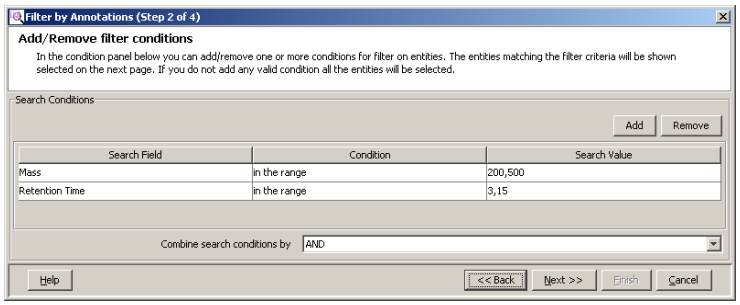


Figure 112 *Add/Remove filter conditions, Filter by Annotations (Step 2 of 4)*

d Review your filter results on the **Filter Results** page (Filter by Annotations (Step 3 of 4)):

1. Review the entities that passed the filter conditions using the spreadsheet view. The entities (rows) that pass the annotation filter are highlighted. The total number of entities (rows) that pass the filter is displayed above the table.
2. Click **Back**, make changes to the parameters, and click **Next** to return to this view.
3. Click **Next**.

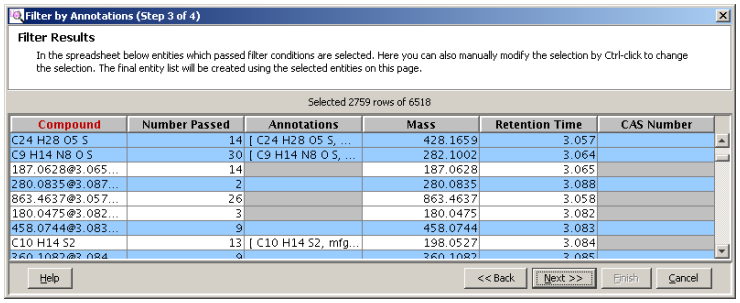
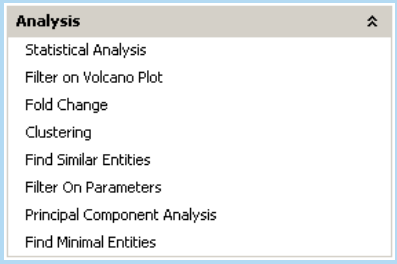


Figure 113 *Filter Results, Filter by Annotations (Step 3 of 4)*

e Enter parameters on the **Save Entity List** page (Filter by Annotations (Step 4 of 4)):

1. Follow the steps presented in "The Save Entity List page" on page 124.
2. Click **Finish**.

Analysis



Mass Profiler Professional supports multiple statistical analytical methods. The statistical methods available to your analysis depend on your experiment design.

Analysis consists of six operations:

- "Statistical Analysis" on page 132
- "Filter on Volcano Plot" on page 142
- "Fold Change" on page 149
- "Clustering" on page 154
- "Find Similar Entities" on page 158
- "Filter on Parameters" on page 162
- "Principal Component Analysis" on page 165
- "Find Minimal Entities" on page 170

Statistical Analysis

Statistical analysis is the pair-wise comparison between two conditions. Because metabolomics involves the analysis of large data sets spanning many conditions, the result of the analysis does not simply pertain to a single hypothesis. The result is from comparisons made against many hypotheses as each pair-wise comparison is made.

For any particular test a p-value may be thought of as the probability of rejecting the null hypothesis when it is in fact true. For a p-value of 0.05 approximately one out of every twenty comparisons results in a false positive analysis (rejection of the null hypothesis when in fact it is true). Thus, if our experiment involves performing 100 comparisons with a p-value of 0.05, we expect five of the comparisons to be false positives. A proper statistical treatment therefore controls the false positive rate for the entire comparison set among the samples that make up the experiment.

The Statistical Analysis wizard has nine (9) steps. The steps that you use depend on your experiment and the combination of the conditions and the statistical test you select in the second step of the wizard (see Figure 114 on page 133 through Figure 117 on page 134). The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

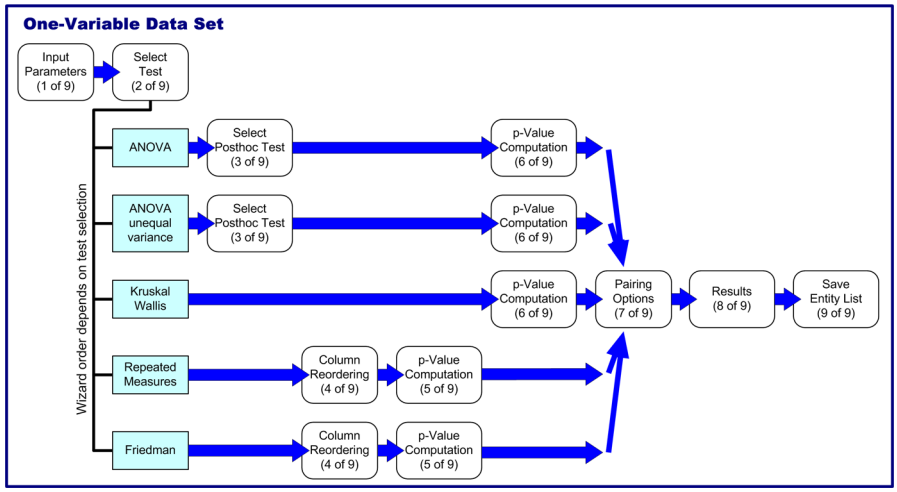


Figure 114 *Flow chart of the Statistical Analysis wizard based on the test selection for the one-variable data set. The p-value computation, step (5 of 9), may be asymptotic or permutative.*

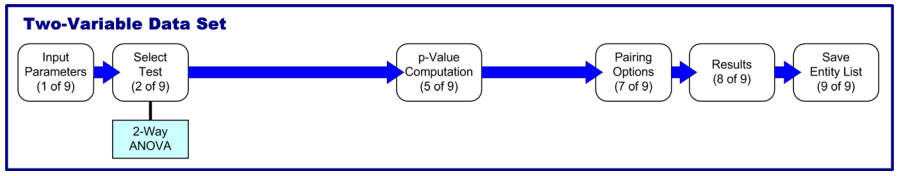


Figure 115 *Flow chart of the Statistical Analysis wizard based on the two-variable data set. The p-value computation, step (6 of 9,) is asymptotic.*

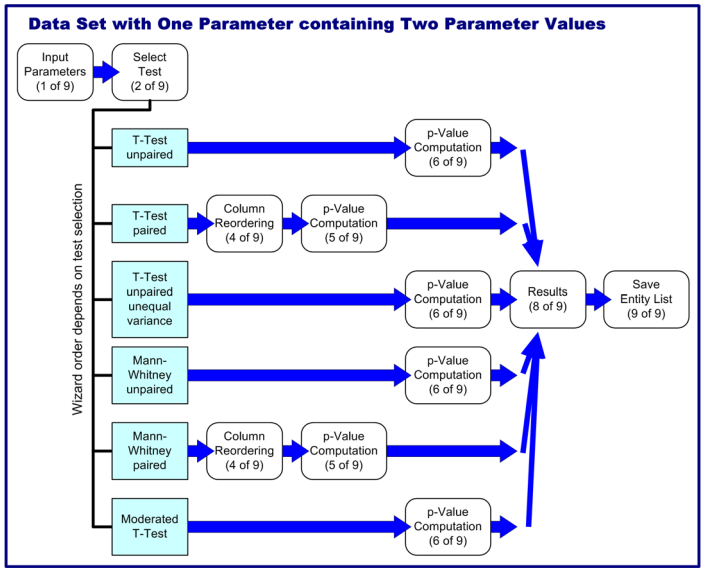


Figure 116 *Flow chart of the Statistical Analysis wizard based on a data set with one parameter containing two parameter values, and selecting two conditions in step 2.*

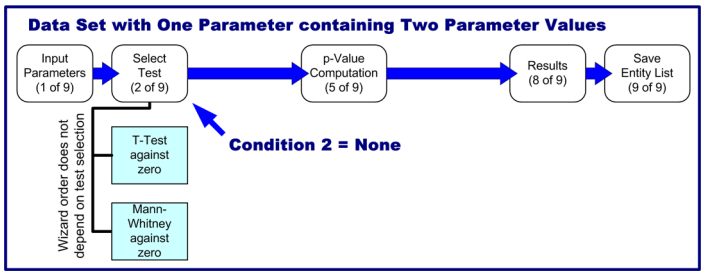


Figure 117 *Flow chart of the Statistical Analysis wizard based on a data set with one parameter containing two parameter values, and selecting only one condition in step 2.*

1. Launch the Statistical Analysis wizard in the Workflow Browser.

a Click **Statistical Analysis** in the Workflow Browser. This operation is illustrated with data from the "One-variable experiment" and the "Two-variable experiment" to provide an overview to more of the wizard options. The steps that you use depend on your experiment and the statistical test you select in the second step of the wizard (see Figure 114 and Figure 115 on page 133). Steps 4, 5, and 6, used when your data consists of a single parameter with two parameter values, are illustrated in section "Filter on Volcano Plot" on page 142.

2. Select the input parameters in **Statistical Analysis (Step 1 of 9)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c (optional) Mark **Exclude missing values from calculation of fold change and p-value**.

d Click **Next**.

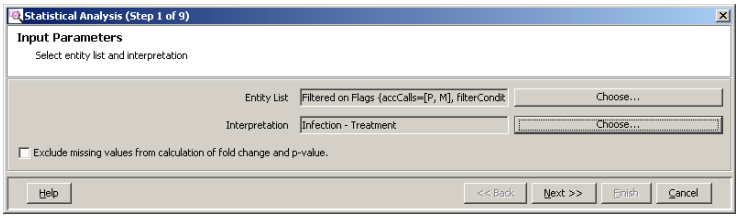


Figure 118 *Input Parameters page (Statistical Analysis (Step 1 of 9))*

3. Select the test in **Statistical Analysis (Step 2 of 9)**.

a Select **ANOVA** statistical test from the **Select Test**. The list of available statistical tests varies based on your experiment and the input parameters.

For the one-variable data set the tests available are (Figure 119):

- **ANOVA** - A statistical method that simultaneously compares the mean values between two or more attribute values of an independent variable. This test assumes that the variance in the attribute values is the same, which is typically true when there are identical sample sizes (replicates) for each attribute value.

- **ANOVA unequal variance** - An ANOVA method used when the sample sizes (replicates) for each attribute value are not identical.
- **Kruskal Wallis** - This is an ANOVA that takes additional steps to remove pre-existing individual sources of variability, such as unequal sample sizes, to produce more significance in the results.
- **Repeated measures** - An ANOVA recommended for three or more attribute values.
- **Friedman** - Recommended when the data is a collection of ranks or ratings, or alternately, when it is measured on a non-linear scale.

For the two-variable data set there is one test available (Figure 120 on page 135):

- **2-Way ANOVA** - A statistical method that simultaneously compares the mean values of the attribute values among two or more independent variables. This test assumes that the variance in the attribute values are the same, which is typically true when there are identical sample sizes (replicates) for each attribute value within each attribute (independent variable).

b Click **Next**.

Depending on your experiment and the test you selected, the next step is either "Select posthoc test in Statistical Analysis (Step 3 of 9)." on page 135, "Reorder your columns in Statistical Analysis (Step 4 of 9)." on page 136, "Select your computation and correction methods in Statistical Analysis (Step 5 of 9)." on page 136, or"Select your computation and correction methods in Statistical Analysis (Step 6 of 9)." on page 137.

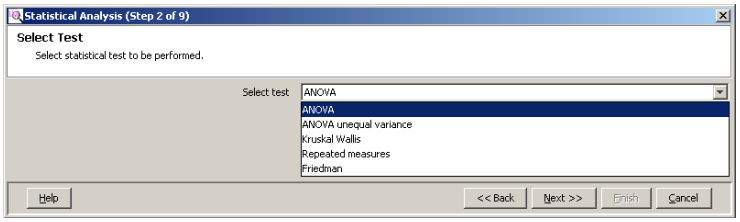


Figure 119 *Select Test page for the one-variable data set (Statistical Analysis (Step 2 of 9))*

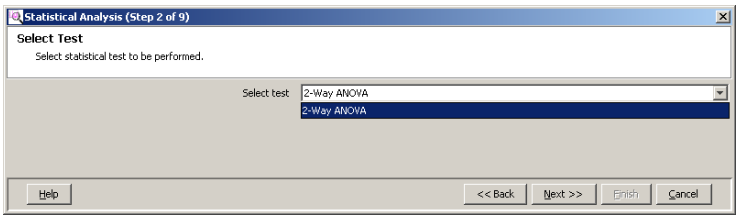


Figure 120 *Select Test page for the two-variable data set (Statistical Analysis (Step 2 of 9))*

4. Select posthoc test in **Statistical Analysis (Step 3 of 9)**.

You are guided through this step when **ANOVA** or **ANOVA unequal variance** is selected as the **Select Test (Step 2 of 9)**. The ANOVA test indicates significance in difference among the attribute values but does not indicate which attribute value means are significantly different, thus after performing an ANOVA the posthoc test is used to determine which of the attribute values are particularly different from each other.

a Select **SNK** test from the **Post Hoc** test.

The post hoc tests available are (Figure 121):

- **SNK** - A post hoc test developed to identify the location of differentiation among many attribute values. This pair-wise test is referred to as the Student-Newman-Keuls test.
- **Tukey HSD** - The most common, stringent post hoc test. The Tukey Honest Significant Difference test involves the pair-wise comparison of the differences to the Tukey critical value.

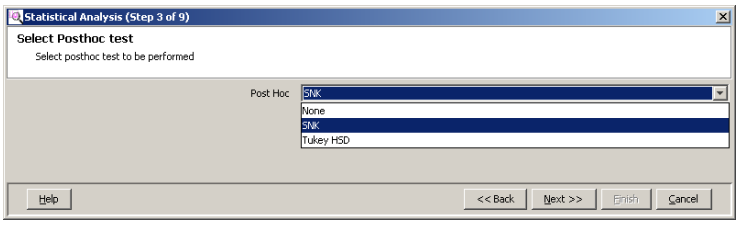


Figure 121 *Posthoc Test page (Statistical Analysis (Step 3 of 9))*

b Click **Next**.

The next step is "Select your computation and correction methods in Statistical Analysis (Step 6 of 9)." on page 137. The alternate steps you may encounter in performing Statistical Analysis are also presented following this step in case your analysis involves a different test.

---

5. Reorder your columns in **Statistical Analysis (Step 4 of 9)**.

You are guided through this step when **Repeated measures** or **Friedman** is selected as the **Select Test (Step 2 of 9).** This step is presented when the selected test involves pair-wise comparisons.

a Select and reorder the entity lists within each column as you like to carry out the paired tests. Reordering involves the **Up**, **Down**, and **Restore** buttons along the right side of the page as described in "Review and order the selected files that are imported in the MS Experiment Creation Wizard (Step 5 of 11)." on page 67.

b Click **Next**.

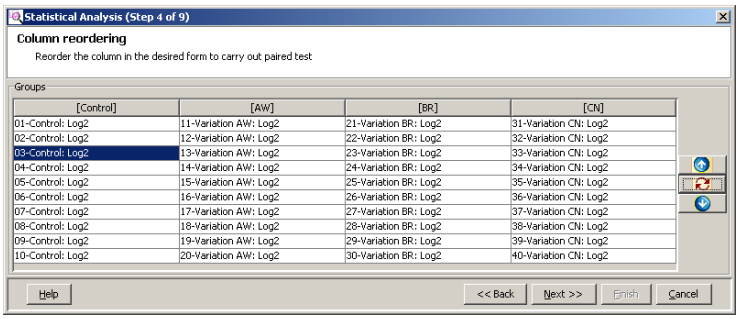


Figure 122 *Column reordering page (Statistical Analysis (Step 4 of 9)*

---

6. Select your computation and correction methods in **Statistical Analysis (Step 5 of 9)**.

You are guided through this step when **Repeated measures, Friedman**, or **2-Way ANOVA** is selected as the **Select Test (Step 2 of 9).** This step is presented when the selected test involves pair-wise comparisons.

a Select **Asymptotic** for the p-value computation algorithm. **Asymptotic** is the default and only choice based on the prior test selection.

b Select **Benjamini Hochberg FDR** (false discovery rate) for the multiple testing correction. This parameter selection is used to control the false positive rate for the entire comparison set among the samples that make up the experiment. **Benjamini Hochberg FDR** (false discovery rate) is the default selection.

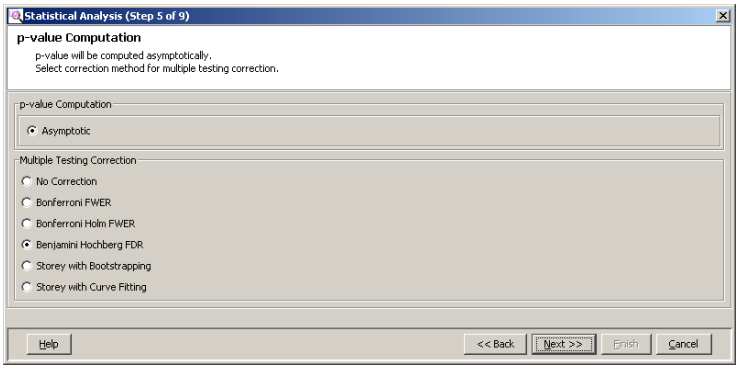


Figure 123 *p-value Computation page (Statistical Analysis (Step 5 of 9))*

c Click **Next**.

The next step is "Select your pairing options in Statistical Analysis (Step 7 of 9)." on page 138.

---

7. Select your computation and correction methods in **Statistical Analysis (Step 6 of 9)**.

You are guided through this step when **ANOVA**, **ANOVA unequal variance**, or **Kruskal Wallis** is selected as the **Select Test (Step 2 of 9).** The p-value computation is performed in cases when 1-way ANOVA tests are performed against zero - such as when the data has a single independent variable.

a Select **Asymptotic** for the **p-value Computation**. If you select **Permutative**, type the **Number of Permutations** the computation performs.

b Select **Benjamini Hochberg FDR** (false discovery rate) for the multiple testing correction. This parameter selection is used to control the false positive rate for the entire comparison set among the samples that make up the experiment. **Benjamini Hochberg FDR** (false discovery rate) is the default selection.

c Click **Next**.

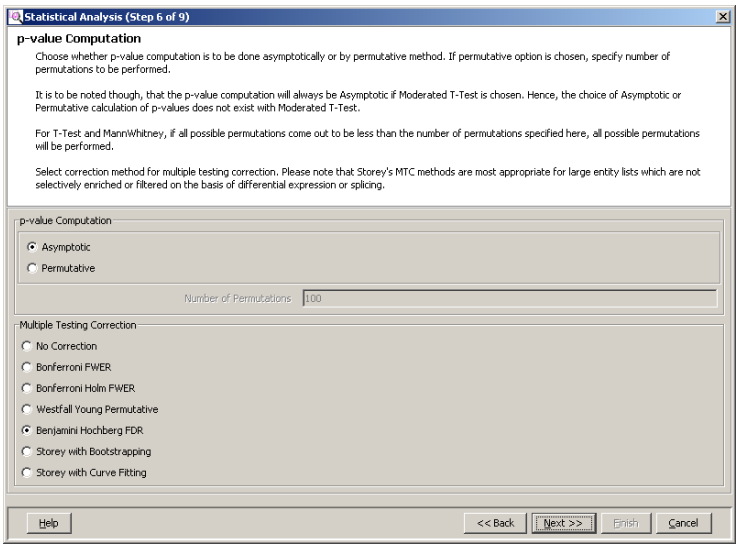


Figure 124 *p-value Computation page (Statistical Analysis (Step 6 of 9))*

8. Select your pairing options in **Statistical Analysis (Step 7 of 9)**.

Choose one or more pairs of conditions, or specifically specify the conditions you want to compare for your analysis.

a Select your **Select pairing option**.

b If you selected **All against a single condition**, select your **Select condition** from the available pairing options. The page appears as shown in Figure 125.

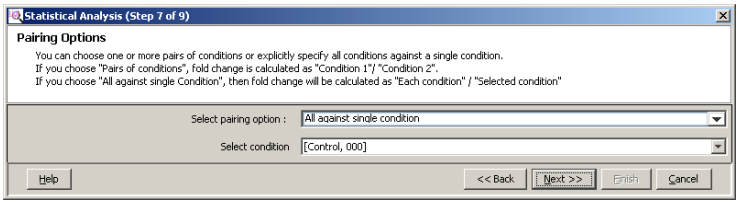


Figure 125 *Pairing Options page (Statistical Analysis (Step 7 of 9))*

c If you selected **Pairs of conditions**, select your **Condition Pairs** and set the preferred pairing order from the available pairing options. The page appears as shown in Figure 126 on page 139.

d Click **Next**.

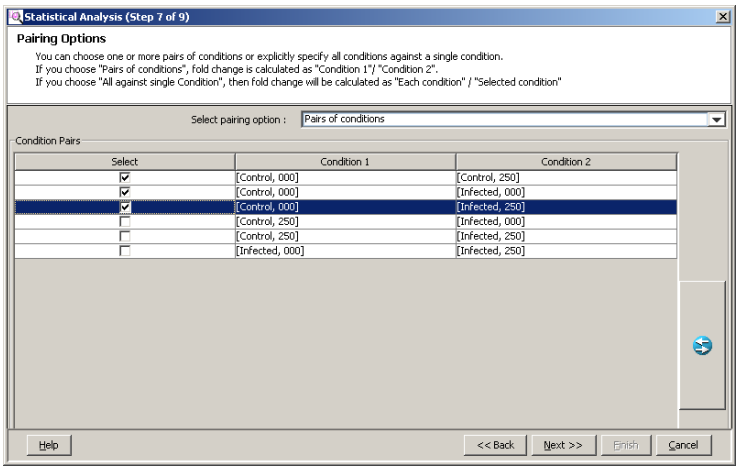


Figure 126 *Pairing Options page (Statistical Analysis (Step 7 of 9))*

9. Review your results in **Statistical Analysis (Step 8 of 9)**.

This step displays the results of your analysis. The results are displayed in multiple tiled windows and the views differ based upon the tests performed.

Test Description: This window describes the statistical test that was applied to the samples along with a summary table that organizes the results by p-value. A p-value of 0.05 is similar to stating that if the mean values for each parameter value (a condition of an independent variable) are identical, then a 5% chance or less exists of observing a difference in the parameter value mean value as large as you observed. In other words, statistical treatment of random sampling from identical populations with a p-value set at 0.05 would lead to a difference smaller than you observed in 95% of experiments and larger than you observed in 5% of the experiments.

The test description used for computing p-values, type of correction used, and P-value computation type (Asymptotic or Permutative) are reported.

Result Summary: The last row of data in the result summary shows the number of entities that would be expected to meet the significance analysis by random chance based on the p-value specified in each column heading. If the number of entities expected by chance is much smaller than those based on the corrected p-value you have realized a selection of entities that show significant difference in the mean values of the parameters.

Compound Table: Each entity that survived the filters is now presented by compound along with the p-values expected and corrected for each of the interpretation sets. Each entity is uniquely identified by its average neutral mass and retention time from across the data sets. The table shows Compound Name, p-values, corrected p-values, Fold change (Absolute), and regulation. FC Absolute means that the fold-change reported is absolute. In other words, if an entity is 2-fold up or 2-fold down, it is still called as 2.0 fold, instead of being called 2.0 fold (for up-regulation) and 0.5 (for down-regulation). Absolute essentially means that no directionality associated with the value exists. Directionality or regulation is indicated separately under the regulation column.

Venn Diagram: Display of the Venn Diagram, or other plot, depends on the samples and experiment grouping for the analysis. The entities that make up each selected

section of the Venn diagram are highlighted in the p-values spreadsheet. The Venn diagram is a graphical view of the most significant entities in each of the sample analysis. Entities in common to the analyses are depicted as overlapping section of the circles. Fewer entities in the regions of overlap are an indication that the samples support the hypothesis that a difference exists in the samples based on the experimental parameters.

Volcano plot: This plot comes up only if two groups exist in the Experiment Grouping. The entities which satisfy the default p-value cutoff (0.05) appear in red and the rest appear in grey. This plot shows the negative log10 of p-value against log (base2.0) of the fold change. Compounds with a large fold-change and low p-value are easily identifiable on this view. If no significant entities are found, you can change the corrected p-value cut-off. You can choose an alternative control group by using the **Back** button. The label at the top of the wizard shows the number of entities satisfying the given p-value.

Note: To change the plot view or export the plot to a file, click and right-click features available on the plot in a same manner similar to that presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. The graphical plot operations available are described in section 7 Data Visualization in the *Mass Profiler Professional User Manual*.

a Review your results.

b Move the slider or type in the **p-value cut-off** value. The default value is 0.05.

Move the slider **p-value cut-off** until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the p-value cut-off affects the results. A larger p-value passes a larger number of entities.

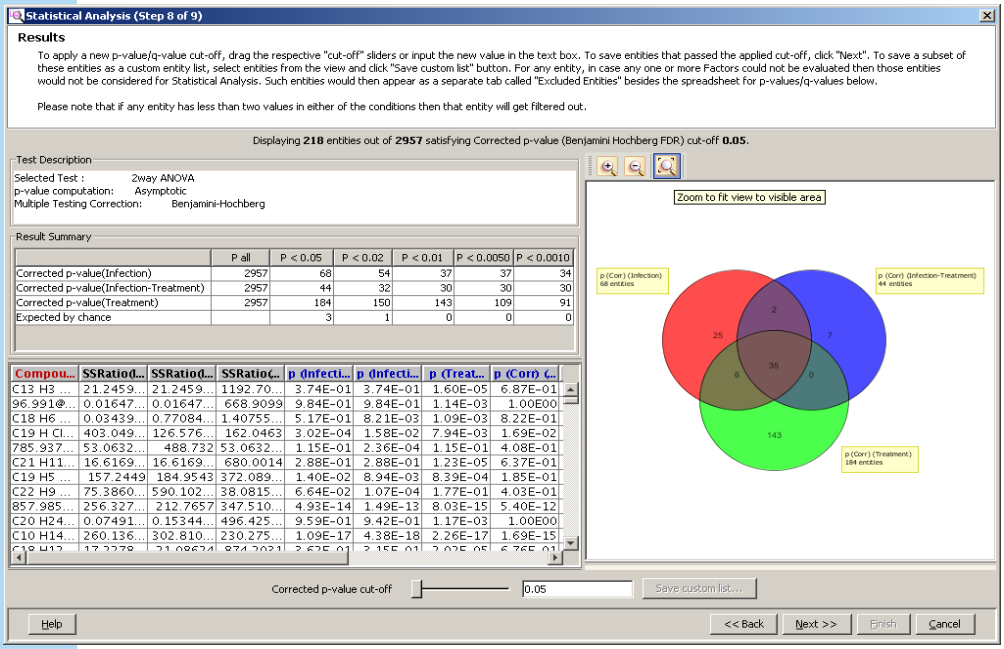


Figure 127 *Results page (Statistical Analysis (Step 8 of 9)) from the one-variable experiment*

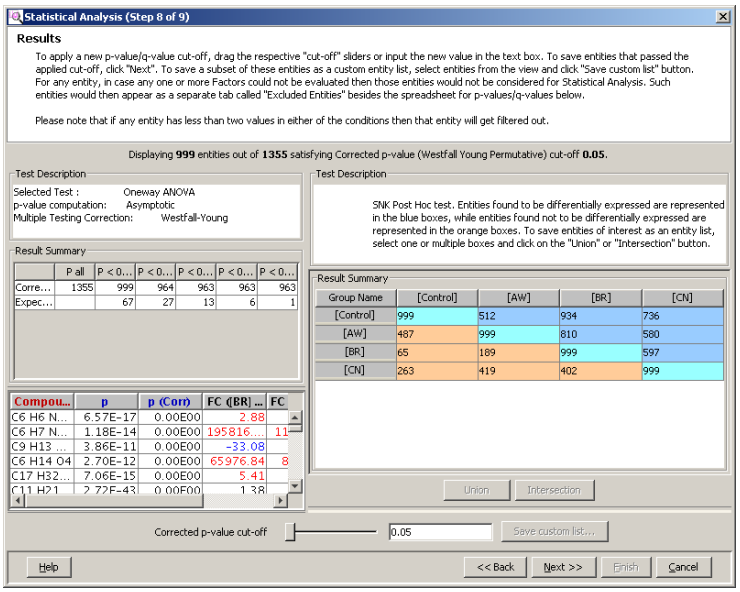


Figure 128 *Results page (Statistical Analysis (Step 8 of 9)) from the two-variable experiment*

c *(Optional)* Create a custom entity list.

1. Click one entity in the compound table, or multiple entities while pressing the **Shift** or **Ctrl** key.
2. Click **Save custom list**.
3. Click **Configure Columns** to select custom annotations.
4. Click **OK**.
5. Type in descriptive information that is stored with the saved entity list.
6. Click **OK**.
7. Repeat saving of custom entity lists as necessary.

d Click **Back**, make changes to prior parameters, and click **Next** to return to the results until you are satisfied with your analysis.

e Click **Next**.

10. Enter save entity list parameters in **Statistical Analysis (Step 9 of 9)**.

The review content and parameters you enter in the **Save Entity List** page is the same as presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

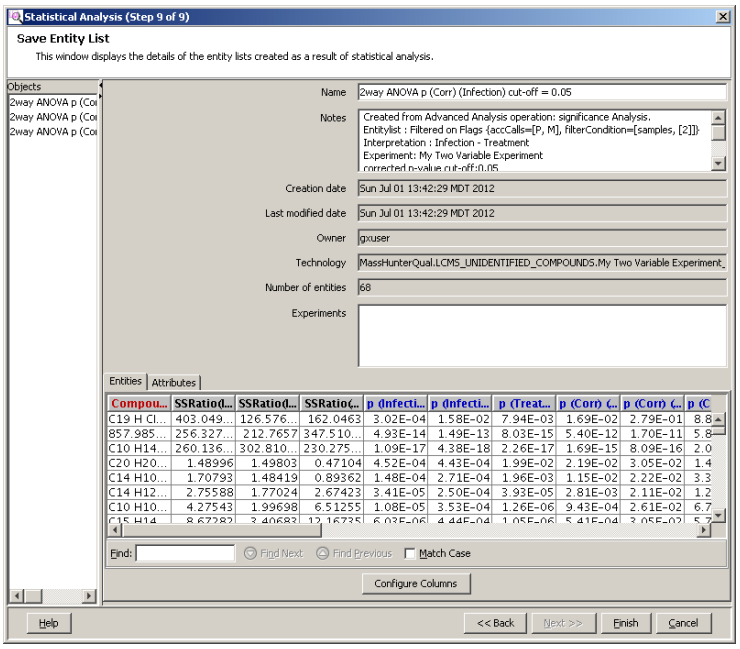


Figure 129 *Save Entity List page (Statistical Analysis (Step 8 of 8))*

Filter on Volcano Plot

A volcano plot is a log-log scatter plot where the p-value is plotted against the fold-change. A volcano plot is useful for identifying events that differ significantly between two groups of subjects. The name “volcano plot” comes from the plot's resemblance to an image of a pyroclastic volcanic eruption with the most significant points at the top of the plot as if they were spewed pieces of molten lava.

A volcano lets you visualize the relationship between fold-change (magnitude of change) and statistical significance (which takes both magnitude of change and variability into consideration).

The Filter on Volcano Plot wizard has seven (7) steps. The steps that you use depend on the statistical test you select in the second step of the wizard (see Figure 130 on page 143). The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

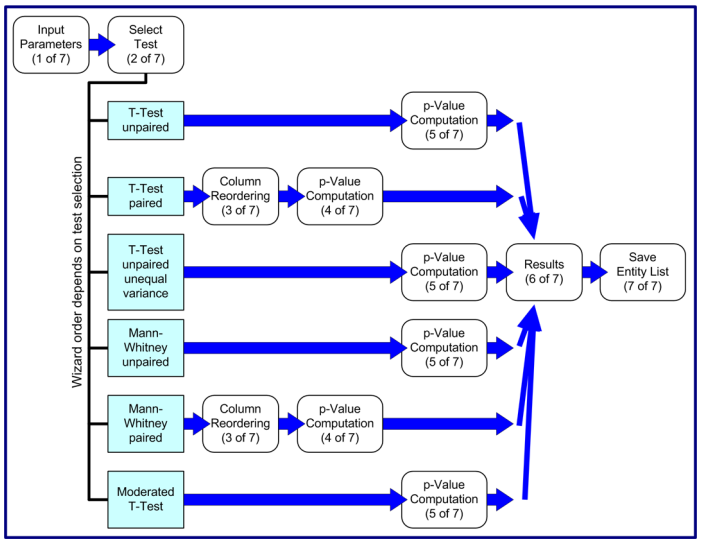


Figure 130 *Flow chart of the Filter on Volcano Plot wizard.*

1. Launch the Filter on Volcano Plot wizard in the Workflow Browser.

a Click **Filter on Volcano Plot** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment" to provide an overview of the wizard options.

2. Select the input parameters in **Filter on Volcano Plot (Step 1 of 7)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c (optional) Mark **Exclude missing values from calculation of fold change and p-value**.

d Click **Next**.

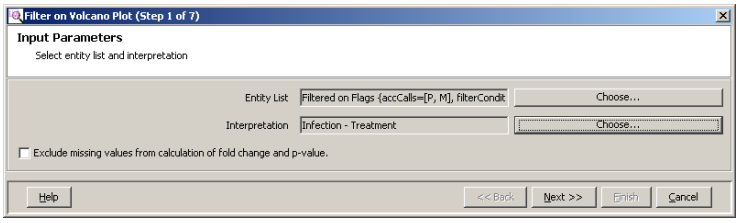


Figure 131 *Input Parameters page (Filter on Volcano Plot (Step 1 of 7))*

3. Select the test in **Filter on Volcano Plot (Step 2 of 7)**.

a Select **Condition 1** from the list of experimental attributes (parameter values) associated with the independent variable(s).

b Select **Condition 2** from the list of experimental attributes (parameter values) associated with the independent variable(s).

c Select the statistical test from the **Select Test**.

The tests available are (Figure 132):

- **T-Test unpaired** - Selected when the conditions being compared consist of attribute values from different independent variables. A t-test is a statistical method that follows the Student's t-distribution (normal distribution). The test works from the hypothesis that the difference between the mean of the conditions has a value of zero.
- **T-Test paired** - Selected when the conditions being compared are different attribute values from the same independent variable.
- **T-Test unpaired unequal variance** - Selected when the conditions being compared consist of attribute values from different independent variables and when the sample sizes (replicates) for each attribute value are not identical.
- **Mann-Whitney unpaired** - Selected when the conditions being compared consist of attribute values from different independent variables. A Mann-Whitney test is a statistical method that identifies whether one of the two conditions tends to have a larger mean without knowledge of the population distribution.
- **Mann-Whitney paired** - Selected when the conditions being compared are different attribute values from the same independent variable.
- **Moderated T-Test** - The moderated t-test is a modification of the t-test unpaired with better handling of two conditions involving (1) small differences between their mean values and also having a low variance within each condition, and (2) large differences between their mean values and a high variance within each condition.

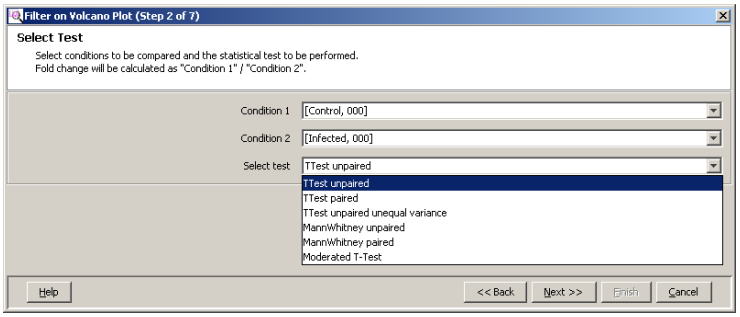


Figure 132 *Select Test page (Filter on Volcano Plot (Step 2 of 7))*

d Click **Next**.

The next step is either "Reorder your columns in Filter on Volcano Plot (Step 3 of 7)." on page 144 or"Select your computation and correction methods in Filter on Volcano Plot (Step 5 of 7)." on page 145.

---

4. Reorder your columns in **Filter on Volcano Plot (Step 3 of 7)**.

You are guided through this step when **T-Test paired** or **Mann-Whitney paired** is selected as the **Select test (Step 2 of 7).** This step is presented when the selected test involves pair-wise comparisons.

a Select and reorder the entity lists within each column as you like to carry out the paired tests. Reordering involves the **Up**, **Down**, and **Restore** buttons along the right side of the page as described in "Review and order the selected files that are imported in the MS Experiment Creation Wizard (Step 5 of 11)." on page 67.

b Click **Next**.

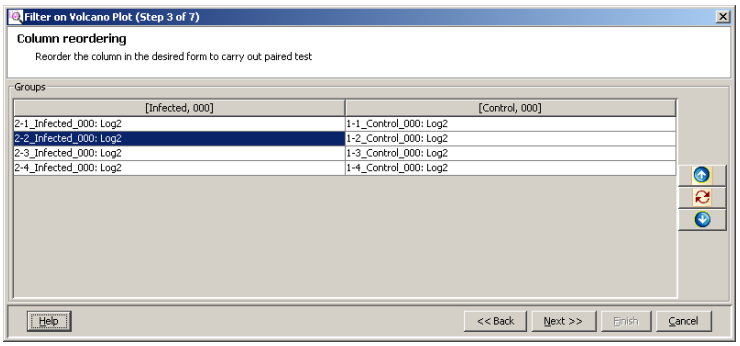


Figure 133 *Column reordering page (Filter on Volcano Plot (Step 3 of 7))*

5. Select your computation and correction methods in **Filter on Volcano Plot (Step 4 of 7)**.

You are guided through this step when **T-Test paired** or **Mann-Whitney paired** is selected as the **Select test (Step 2 of 7).** This step is presented when the selected test involves pair-wise comparisons.

a Select **Asymptotic** for the p-value computation algorithm. **Asymptotic** is the default and only choice based on the prior test selection.

b Select **Benjamini Hochberg FDR** (false discovery rate) for the multiple testing correction. This parameter selection is used to control the false positive rate for the entire comparison set among the samples that make up the experiment. **Benjamini Hochberg FDR** (false discovery rate) is the default selection.

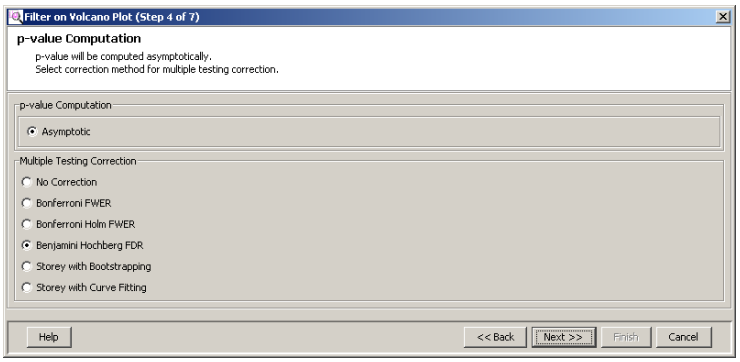


Figure 134 *p-value Computation page (Filter on Volcano Plot (Step 4 of 7))*

c Click **Next**.

The next step is "Review your results in Filter on Volcano Plot (Step 6 of 7)." on page 146.

6. Select your computation and correction methods in **Filter on Volcano Plot (Step 5 of 7)**.

You are guided through this step when **T-Test unpaired**, **T-Test unpaired unequal variance**, **Mann-Whitney unpaired**, or **Moderated T-Test** is selected as the **Select test (Step 2 of 7).** The p-value computation is performed in cases when 1-way ANOVA tests are performed against zero - such as when the data has a single independent variable.

a Select **Asymptotic** for the **p-value Computation**. If you select **Permutative**, type the **Number of Permutations** the computation performs.

b Select **Benjamini Hochberg FDR** (false discovery rate) for the multiple testing correction. This parameter selection is used to control the false positive rate for the entire comparison set among the samples that make up the experiment. **Benjamini Hochberg FDR** (false discovery rate) is the default selection.

c Click **Next**.

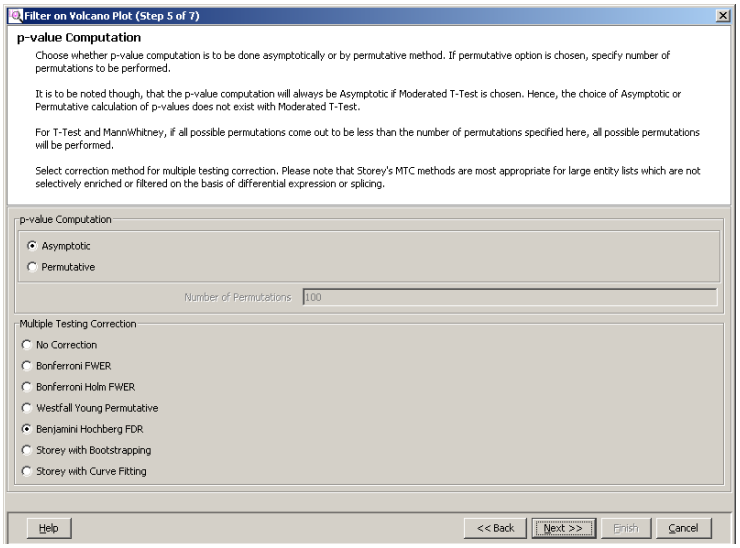


Figure 135 *p-value Computation page (Filter on Volcano Plot (Step 5 of 7))*

7. Review your results in **Filter on Volcano Plot (Step 6 of 7)**.

This step displays the results of analysis. For example on completion of T-Test unpaired. The results are displayed in multiple tiled windows and the views differ based upon the tests performed.

Test Description: This window describes the statistical test that was applied to the samples along with a summary table that organizes the results by p-value. A p-value of 0.05 is similar to stating that if the mean values for each parameter value (a condition of an independent variable) are identical, then a 5% chance or less exists of observing a difference in the parameter value mean value as large as you observed. In other words, statistical treatment of random sampling from identical populations with a p-value set at 0.05 would lead to a difference smaller than you observed in 95% of experiments and larger than you observed in 5% of the experiments.

The test description used for computing p-values, type of correction used, and P-value computation type (Asymptotic or Permutative) are reported.

Result Summary: The last row of data in the result summary shows the number of entities that would be expected to meet the significance analysis by random chance based on the p-value specified in each column heading. If the number of entities expected by chance is much smaller than those based on the corrected p-value you have realized a selection of entities that show significant difference in the mean values of the parameters.

Compound Table: Each entity that survived the filters is now presented by compound along with the p-values expected and corrected for each of the interpretation sets. Each entity is uniquely identified by its average neutral mass and retention time from across the data sets. The table shows Compound Name, p-values, cor-

rected p-values, Fold change (Absolute), and regulation. FC Absolute means that the fold-change reported is absolute. In other words, if an entity is 2-fold up or 2-fold down, it is still called as 2.0 fold, instead of being called 2.0 fold (for up-regulation) and 0.5 (for down-regulation). Absolute essentially means that no directionality associated with the value exists. Directionality or regulation is indicated separately under the regulation column.

Volcano plot: This plot comes up only if two groups exist in the Experiment Grouping. The entities which satisfy the default p-value cutoff (0.05) appear in red and the rest appear in grey. This plot shows the negative log10 of p-value against log (base2.0) of the fold change. Compounds with a large fold-change and low p-value are easily identifiable on this view. If no significant entities are found, you can change the corrected p-value cut-off. You can choose an alternative control group by using the **Back** button. The label at the top of the wizard shows the number of entities satisfying the given p-value.

You may change the plot view or export the plot to a file by using the click and right-click features available on the plot in a same manner similar to that presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. The graphical plot operations available are described in section 7 Data Visualization in the *Mass Profiler Professional User Manual*.

a Review your results.

b Move the slider or type in the **p-value cut-off** value. The default value is 0.05.

Move the slider **p-value cut-off** until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the p-value cut-off affects the results. A larger p-value passes a larger number of entities.

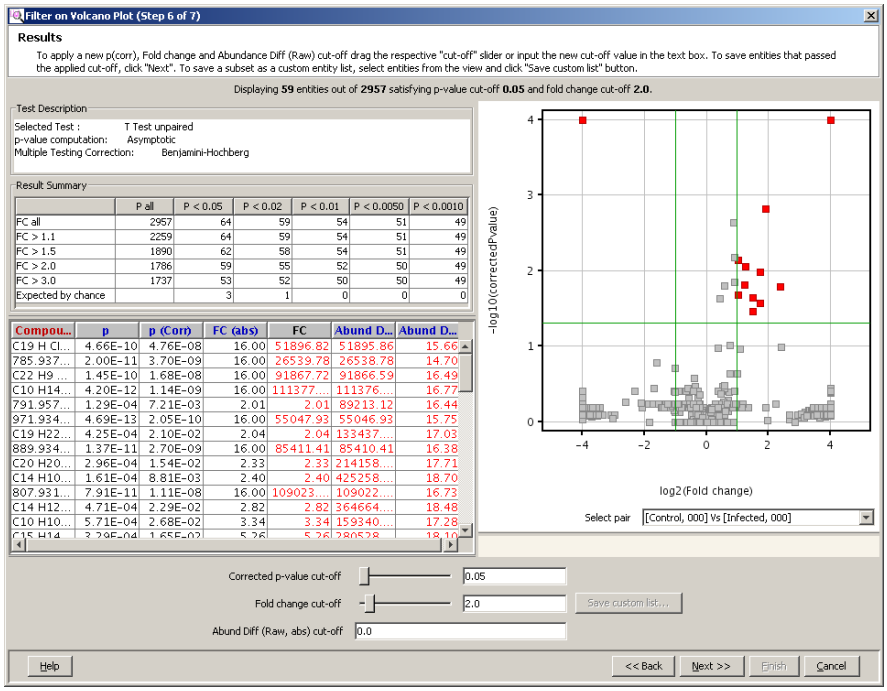


Figure 136 *Results page (Filter on Volcano Plot (Step 6 of 7))*

c Move the slider or type in the **Fold change cut-off** value. The default value is 2.0.

To re-adjust the fold change cut-off move the slider **Fold change cut-off** until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the fold change cut-off affects the results. A larger fold change cut-off value passes fewer entities.

d Type in a **Abund Diff (Raw, abs) cut-off** value. The default value is 0.0.

To calculate filtered entities using the new abundance difference value, press either the **Enter** key or the **Tab** key. Enter abundance difference cut-off values until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the abundance difference cut-off affects the results. A larger abundance difference cut-off value passes fewer entities.

e *(Optional)* Create a custom entity list as described in "Review your results in Statistical Analysis (Step 8 of 9)." on page 139.

f Click **Next**.

---

8. Enter save entity list parameters in **Filter on Volcano Plot (Step 7 of 7)**.

The review content and parameters you enter in the **Save Entity List** page is the same as presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

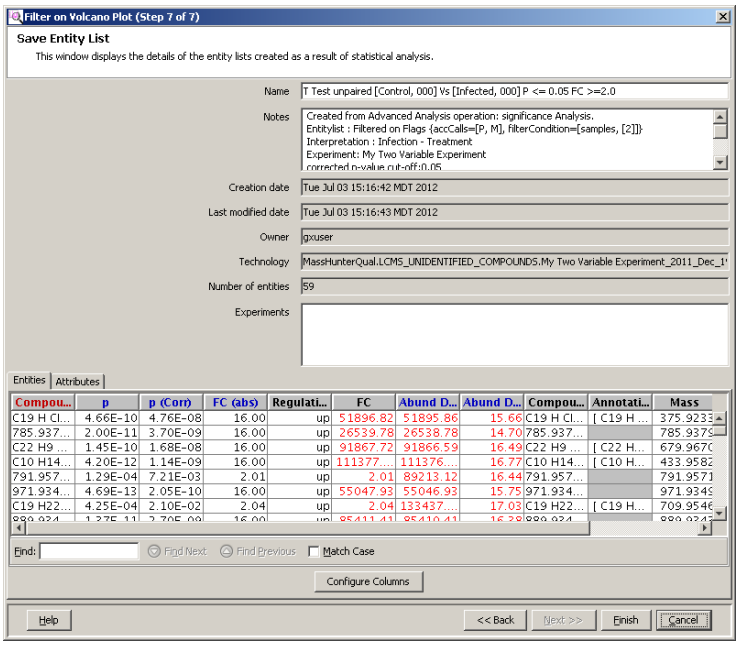


Figure 137 *Save Entity List page (Filter on Volcano Plot (Step 7 of 7))*

Fold Change

Fold change is a signed value that describes how much an entity changes from its initial to its final value. For example, when the abundance of an entity changes from a value of 60 to a value of 15 the fold change is -4. The entity experienced a four-fold decrease. Fold change is the ratio of the final value to the initial value.

Fold change analysis in metabolomics is used to identify entities with abundance ratios, or differences between a treatment and a control, that are in excess of specified cut-off or threshold value. Fold change is calculated between the conditions where Condition 1 and another condition, Condition 2, are treated as a single group. To help with terms, a condition may also be referred to as a parameter value or an attribute value.

Raw Abundance: Evaluates the absolute ratio between Condition 2 and Condition 1; Fold change= |Condition1/Condition2|.

Normalized Abundance: Evaluates the absolute difference between the normalized intensities of the conditions; Fold change= |(Condition1 - Condition2)|.

The Fold Change wizard has six (6) steps. Step 3 of 6 is skipped and not documented. The steps that you use are shown in Figure 138. The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

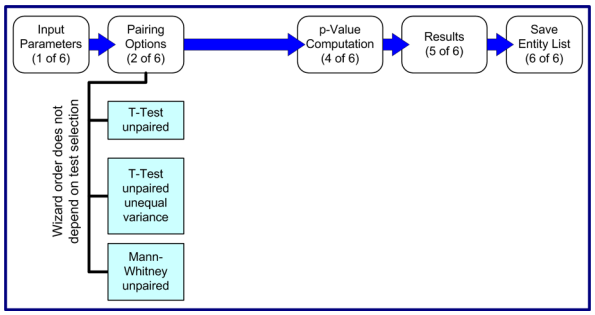


Figure 138 *Flow chart of the Fold Change wizard.*

1. Launch the Fold Change wizard in the Workflow Browser.

a Click **Fold Change** in the Workflow Browser. This operation is illustrated with data from the "One-variable experiment" to provide an overview of the wizard options.

2. Select the input parameters in **Fold Change (Step 1 of 6)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c (optional) Mark **Exclude missing values from calculation of fold change and p-value**.

d Click **Next**.

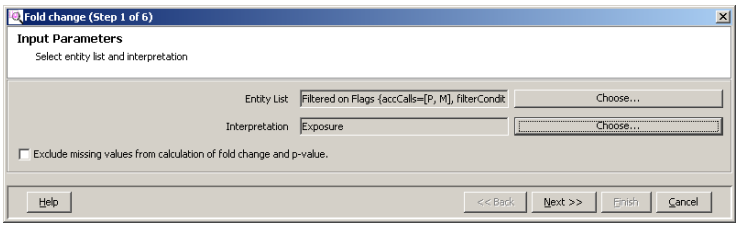


Figure 139 *Input Parameters page (Filter on Fold Change (Step 1 of 6))*

3. Select your pairing options and test in **Fold Change (Step 2 of 6)**.

Choose one or more pairs of conditions, or specifically specify the conditions you want to compare for your analysis, and the analysis test.

a Select your **Select pairing option**.

a If you selected **All against a single condition**, select your **Select condition** from the available pairing options. The page appears as shown in Figure 140.

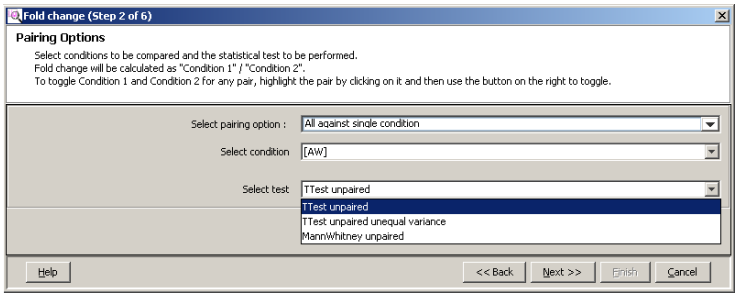


Figure 140 *Pairing Options page (Fold Change (Step 2 of 6))*

b If you selected **Pairs of conditions**, select your **Condition Pairs** and set the preferred pairing order from the available pairing options. The page appear as shown in Figure 141.

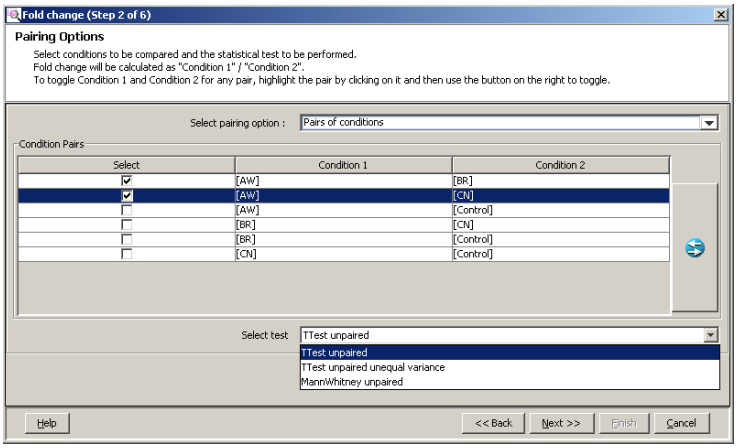


Figure 141 *Pairing Options page (Fold Change (Step 2 of 6))*

c Select the statistical test from the **Select Test**.

The tests available are:

- **T-Test unpaired** - Selected when the conditions being compared consist of attribute values from different independent variables. A t-test is a statistical

method that follows the Student's t-distribution (normal distribution). The test works from the hypothesis that the difference between the mean of the conditions has a value of zero.

- **T-Test unpaired unequal variance** - Selected when the conditions being compared consist of attribute values from different independent variables and when the sample sizes (replicates) for each attribute are not identical.
- **Mann-Whitney unpaired** - Selected when the conditions being compared consist of attribute values from different independent variables. A Mann-Whitney test is a statistical method that identifies whether one of the two conditions tends to have a larger mean without knowledge of the population distribution.

d Click **Next**.

The next step is "Select your computation and correction methods in Fold Change (Step 4 of 6)." on page 151.

4. Select your computation and correction methods in **Fold Change (Step 4 of 6)**.

You are guided through this step regardless of the test selected as the **Select test (Step 2 of 6).** The p-value computation is performed in cases when 1-way ANOVA tests are performed against zero - such as when the data has a single independent variable.

a Select **Asymptotic** for the **p-value Computation**. If you select **Permutative**, type the **Number of Permutations** the computation performs.

b Select **Benjamini Hochberg FDR** (false discovery rate) for the multiple testing correction. This parameter selection is used to control the false positive rate for the entire comparison set among the samples that make up the experiment. **Benjamini Hochberg FDR** (false discovery rate) is the default selection.

c Click **Next**.

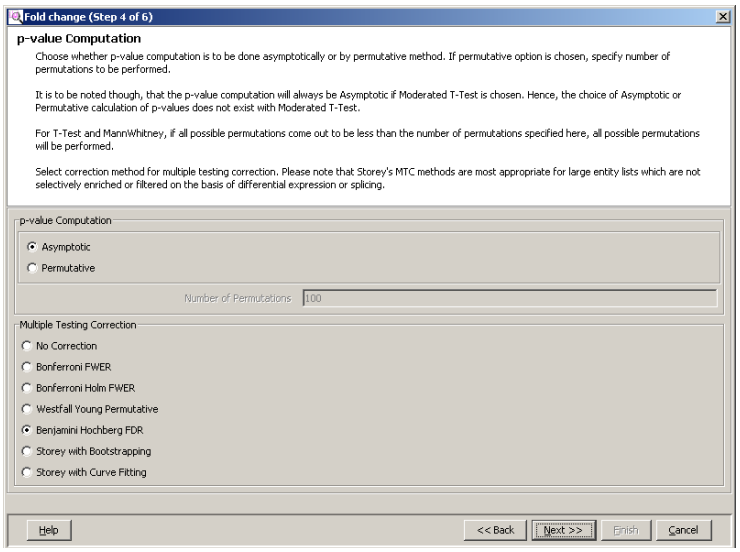


Figure 142 *p-value Computation page (Fold Change (Step 4 of 6))*

5. Review your results in **Fold Change (Step 5 of 6)**.

All of the entities passing the fold change cut-off are displayed along with their annotations. The default view opens a Profile Plot - By Group.

To change the plot view or to export the plot to a file, click and right-click features available on the plot in a manner similar to that presented in "Review the summary report in the Analysis: Significance Testing and Fold Change (Step 1 of 8) workflow." on page 82. The graphical plot operations available are described in section 7 Data Visualization in the *Mass Profiler Professional User Manual*.

a Review your results.

b Move the slider or type in the **Fold change cut-off** value. The default value is 2.0.

To re-adjust the fold change cut-off move the slider **Fold change cut-off** until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the fold change cut-off affects the results. A larger fold change cut-off value passes fewer entities.

c Type in a **Abund Diff (Raw, abs) cut-off** value. The default value is 0.0.

To calculate filtered entities using the new abundance difference value, press either the **Enter** key or the **Tab** key. Enter abundance difference cut-off values until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the abundance difference cut-off affects the results. A larger abundance difference cut-off value passes fewer entities.

d Select the **Minimum number of pairs**, the minimum number of condition pairs that the fold change cut-off must meet for each entity to pass the filter.

e *(Optional)* Create a custom entity list as described in "Review your results in Statistical Analysis (Step 8 of 9)." on page 139.

f Click **Next**.

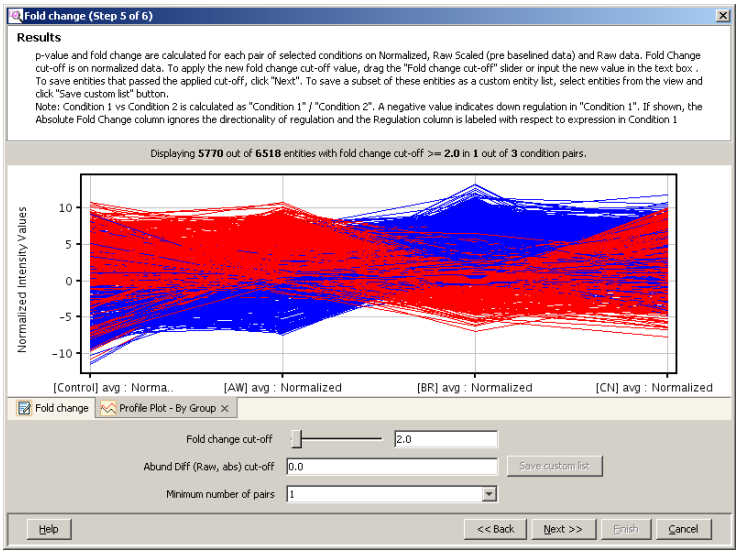


Figure 143 *Results page Profile Plot - By Group (Fold Change (Step 5 of 6))*

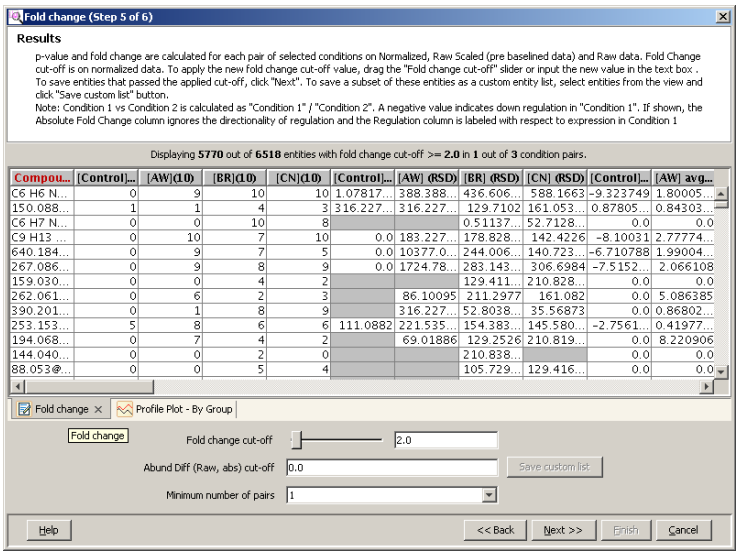


Figure 144 *Results page Fold change (Fold Change (Step 5 of 6))*

6. Enter save entity list parameters in **Fold Change (Step 6 of 6)**.

The review content and parameters you enter in the **Save Entity List** page are the same as presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

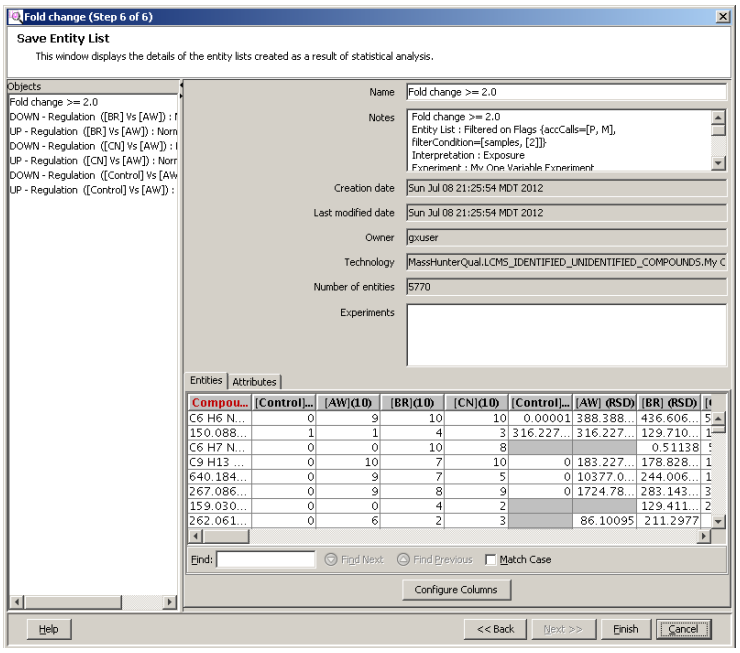


Figure 145 *Save Entity List page (Fold Change (Step 6 of 6))*

Clustering

Clustering is the organization of a set of samples or entities into subsets (clusters) that have similar features. Clustering is a method of unsupervised learning used to organize compounds or entities and conditions in the data set into subsets based on the similarity of their feature abundance profiles.

Mass Profiler Professional implements a variety of clustering methods: **K-Means**, **Hierarchical**, and **Self-Organizing Maps** (SOM), along with a variety of distance functions - **Euclidean**, **Square Euclidean**, **Manhattan**, **Chebychev**, **Differential**, **Pearson Absolute**, **Pearson Centered**, and **Pearson Uncentered**. Data is sorted on the basis of the available distance measures to group entities or conditions. Since different algorithms work on different kinds of data, these algorithms and distance measures ensure that a wide variety of data can be clustered effectively.

Interactive views such as the ClusterSet View, the Dendrogram View, and the U Matrix View are provided for visualization. These views let you drill down into subsets of data and collect individual entity lists into new entity lists for further analysis.

The entity list created is placed under Analysis in the Experiment Navigator. Clustering results are presented as:

Compound Tree: This is a dendrogram of the entities showing the relationship between the entities generated by Hierarchical Clustering.

Condition Trees: This is a dendrogram of the conditions and shows the relationship between the conditions in the experiment generated by Hierarchical Clustering.

Combined Trees: This is a two-dimensional dendrogram that results from performing Hierarchical Clustering on both entities and conditions which are grouped according to the similarity of their abundance profiles.

Classification: This is a cluster set view of entities grouped into clusters based on the similarity of their abundance profiles.

Some clustering algorithms, like Hierarchical Clustering, do not distribute data into a fixed number of clusters, but rather produce a grouping hierarchy. Most similar entities are merged together to form a cluster and this combined entity is treated as a unit thereafter until all the entities are grouped together. The result is a tree structure or a dendrogram, where the leaves represent individual entities and the internal nodes represent clusters of similar entities.

The Clustering wizard has four (4) steps. The steps that you use are shown in Figure 146 on page 155. The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

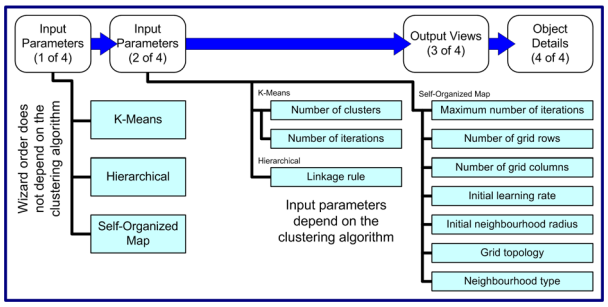


Figure 146 *Flow chart of the Clustering wizard.*

1. Launch the Clustering wizard in the Workflow Browser.

a Click **Clustering** in the Workflow Browser. This operation is illustrated with data from the "One-variable experiment" to provide an overview of the wizard options.

2. Select the input parameters in **Clustering (Step 1 of 4)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c Select **K-Means** for the **Clustering Algorithm**. Your clustering algorithm selection affects the parameters you enter on the next page. The possible values are **K-Means**, **Hierarchical**, and **Self-Organizing Maps**.

d Click **Next**.

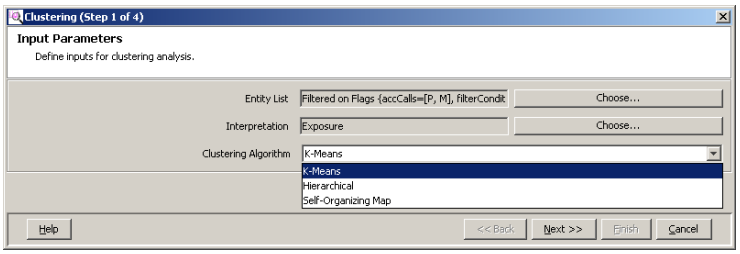


Figure 147 *Input Parameters page (Filter on Clustering (Step 1 of 4))*

3. Select and enter the input parameters in **Clustering (Step 2 of 4)**.

The parameters entered for this page depend on the clustering algorithm you selected in the prior page.

a Select **Conditions** for the **Cluster on** parameter. The possible values are **Entities**, **Conditions** or **Both entities and conditions**.

b Select **Euclidean** for the **Distance metric**. This parameter determines how the similarity of two entities or conditions is calculated. The metric influences the shape of the clusters, as some elements may be close to one another according to one metric and farther away according to another metric. The possible values are **Euclidean**, **Square Euclidean**, **Manhattan**, **Chebychev**, **Differential**, **Pearson Absolute**, **Pearson Centered**, and **Pearson Uncentered**.

c Type the **Number of clusters**. The default is 3.

d Type the **Number of iterations**. The default is 50.

If you selected **Hierarchical** for the clustering algorithm the input parameters page appears as shown in Figure 149.

If you selected **Self-Organizing Maps** for the clustering algorithm the input parameters page appears as shown in Figure 150.

e Click **Next**.

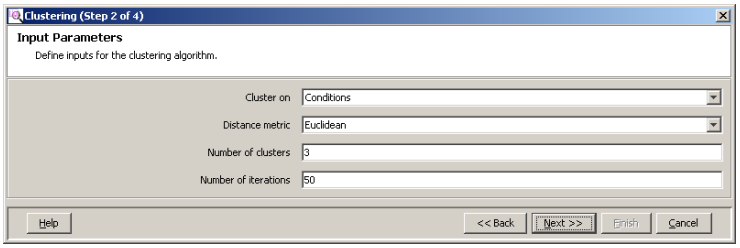


Figure 148 *Input Parameters page (Clustering (Step 2 of 4)) for Clustering Algorithm K-Means*

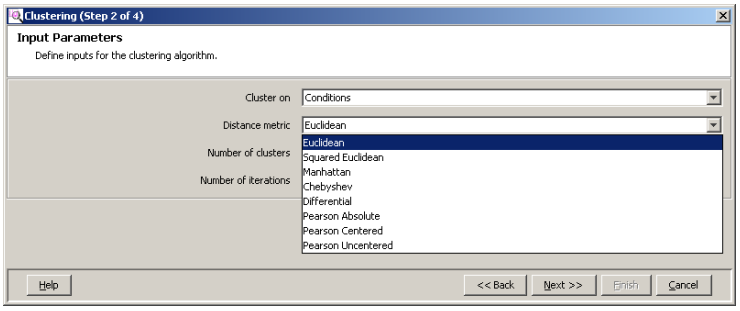


Figure 149 *Input Parameters page (Clustering (Step 2 of 4)) for Clustering Algorithm Hierarchical*

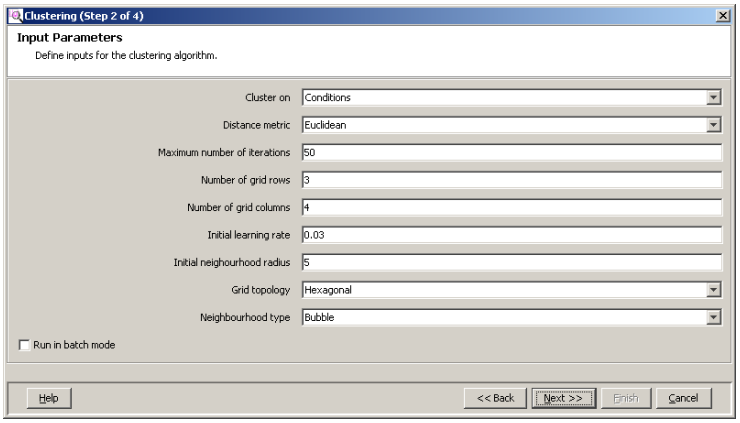


Figure 150 *Input Parameters page (Clustering (Step 2 of 4)) for Clustering Algorithm Self-Organized Map*

4. Review your results in **Output views (Step 3 of 4)**.

a Review the clustering results. The clustering results are presented as either the ClusterSet View (Figure 151), the Dendrogram View (Figure 152), or the U Matrix View (Figure 153) depending on the method and parameters selected.

To find additional information regarding the options for changing the output view press **F1** or review Section 9.3 Graphical Views of Clustering Analysis Output in the *Mass Profiler Professional User Manual*.

b If the results are not satisfactory click **Back**, change the parameters, and re-run the clustering algorithm.

c Click **Next**.

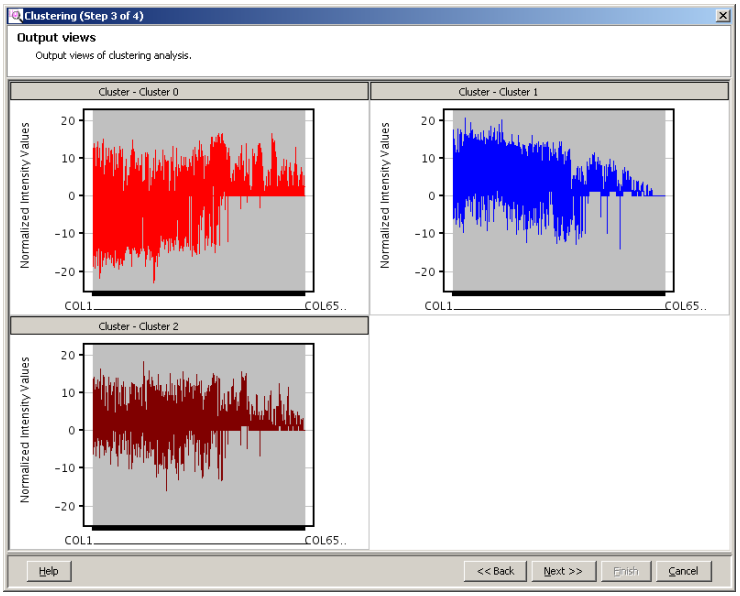


Figure 151 *Output views page (Clustering (Step 3 of 4)) - ClusterSet view*

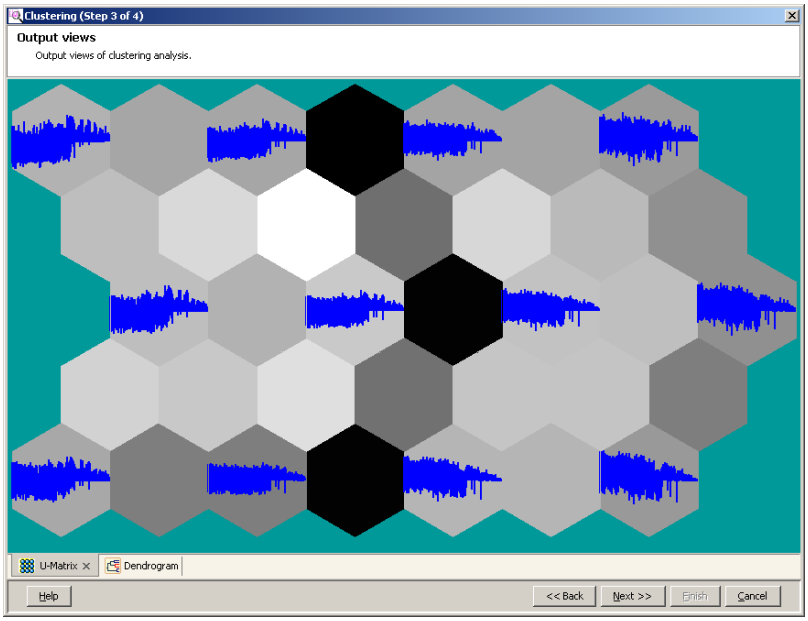


Figure 152 *Output views page (Clustering (Step 3 of 4)) - Dendrogram view*

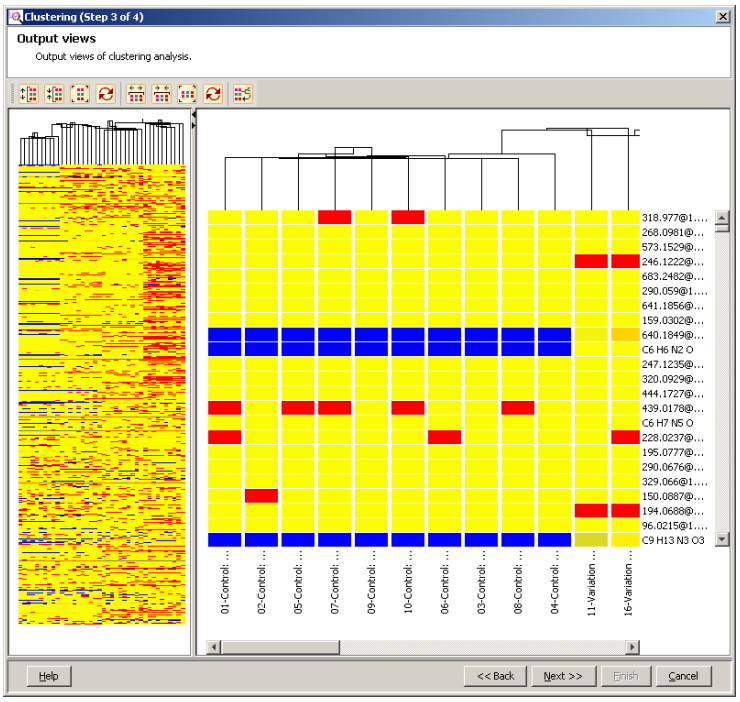


Figure 153 *Output views page (Clustering (Step 3 of 4)) - U Matrix view*

5. Enter object details in **Clustering (Step 4 of 4)**.

a Type in a descriptive **Name** and **Notes** which are stored with the saved entity list.

b Click **Finish**.

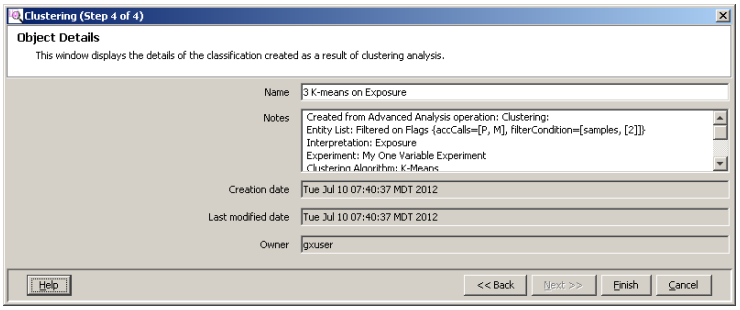


Figure 154 *Object Details page (Clustering (Step 4 of 4))*

Find Similar Entities

Find Similar Entities allows you to query a specific entity list, or the entire data set, to find entities whose abundance profile matches a selected entity of interest.

The similarity metrics available for this filter are **Euclidean** (finds similar entities based on the vector distance between two entities), **Pearson** (finds similar entities based on the trend and rate of change of trend of the chosen entity), and **Spearman** (finds similar entities based on the trend of the chosen entity).

Euclidean - Calculates the Euclidean distance where the vector elements are the columns. The square root of the sum of the square of the A and the B vectors for each element is calculated and then the distances are scaled between -1 and +1. Result = (A-B) · (A-B).

Pearson Correlation - Calculates the mean of all elements in vector X. Then it subtracts that value from each element in X and calls the resulting vector A. It does the same for Y to make a vector B. Result = A · B/(|A| · |B|)

Spearman Correlation - Orders all the elements of vector X and uses this order to assign a rank to each element of X. It makes a new vector X' where the i-th element in X' is the rank of Xi in X and then makes a vector A from X' in the same way as A was made from a in the Pearson Correlation. Similarly, it makes a vector B from Y. Result = A · B/(|A| · |B|). The advantage of using Spearman Correlation is that it reduces the effect of the outliers on the analysis.

The Find Similar Entities wizard has three (3) steps. The steps that you use are shown in Figure 155. The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

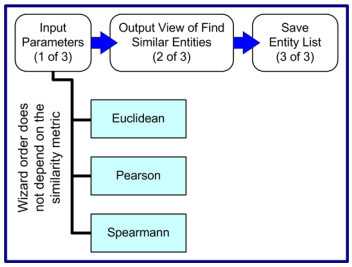


Figure 155 *Flow chart of the Find Similar Entities wizard.*

1. Launch the Find Similar Entities wizard in the Workflow Browser.

a Click **Find Similar Entities** in the Workflow Browser. This operation is illustrated with data from the "One-variable experiment" to provide an overview of the wizard options.

2. Select the input parameters in **Find Similar Entities (Step 1 of 3)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c Click **Select** to select the **Choose Query Entity** from the list of entities (See Figure 157 on page 160).

d Click **OK**.

e Select a **Similarity Metric**. The possible values are **Euclidean**, **Pearson**, and **Spearman**.

f Click **Next**.

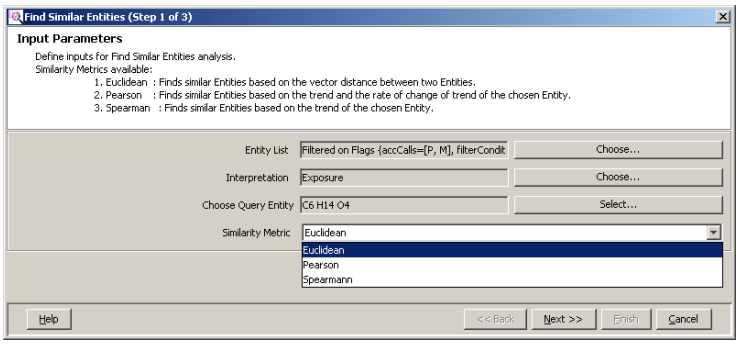


Figure 156 *Input Parameters page (Find Similar Entities (Step 1 of 3))*

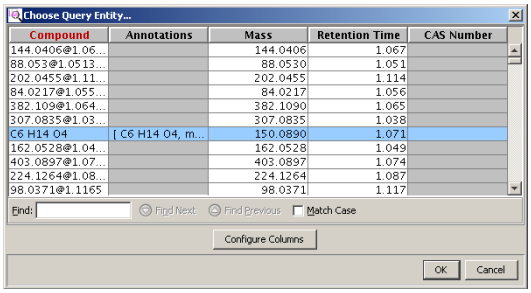


Figure 157 *Choose Query Entity dialog box*

3. Review and adjust your results in **Find Similar Entities (Step 2 of 3)**.

The abundance profile of the target entity is shown in bold and along with the profiles of the entities whose correlation coefficients are within the target profile. The number of entities that meet your criteria and the current cut-off values are shown above the profile plot.

a Move the slider or type the **Minimum** cut-off value. The default value is 0.95.

To re-adjust minimum cut-off move the slider **Minimum** cut-off value until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the cut-off affects the results. A smaller minimum cut-off value passes more entities (Figure 159 on page 161).

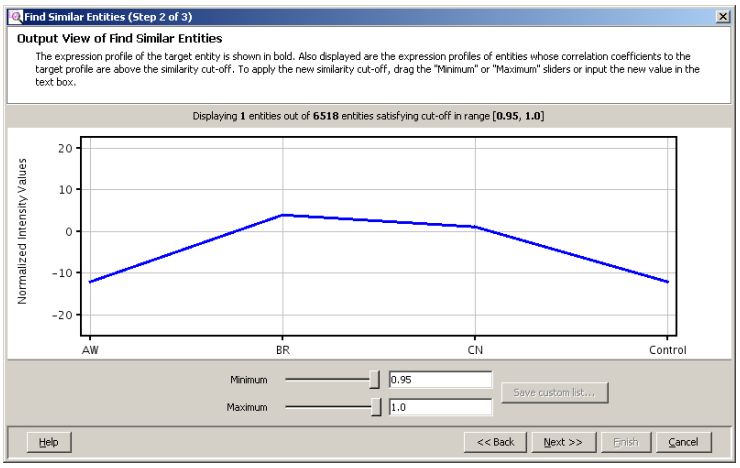


Figure 158 *Output View of Find Similar Entities page (Find Similar Entities (Step 2 of 3))*

b Move the slider or type the **Maximum** cut-off value. The default value is 1.0.

To re-adjust the maximum cut-off may move the slider **Maximum** cut-off value until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the cut-off affects the results. A smaller maximum cut-off value passes more entities (Figure 159). The Maximum cut-off value cannot be smaller than the Minimum cut-off value.

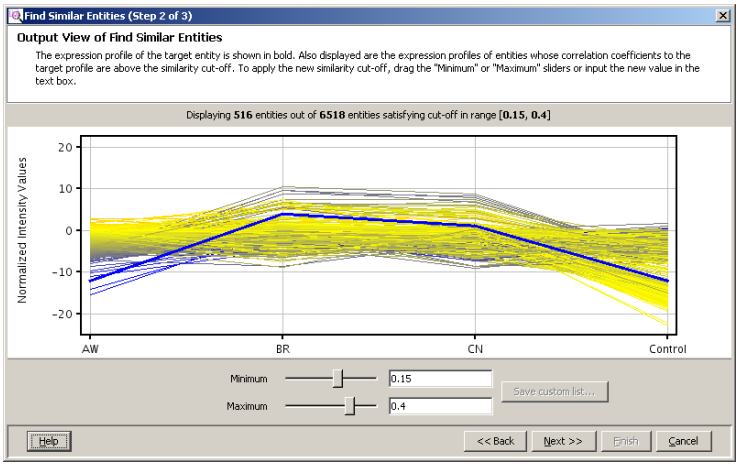


Figure 159 *Output View of Find Similar Entities page (Find Similar Entities (Step 2 of 3))*

c *(Optional)* Create a custom entity list as described in "Review your results in Statistical Analysis (Step 8 of 9)." on page 139.

d Click **Next**.

4. Enter save entity list parameters in **Find Similar Entities (Step 3 of 3)**.

The review content and parameters you enter in the **Save Entity List** page are the same as presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

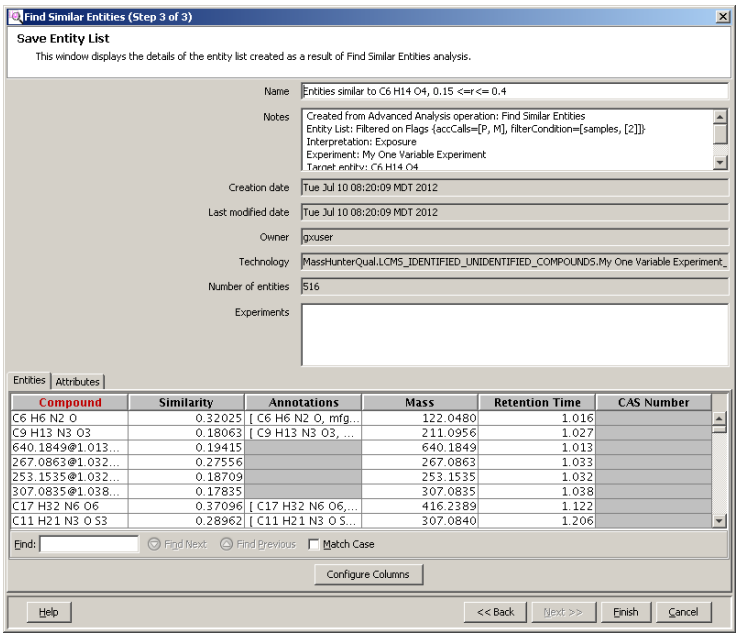


Figure 160 *Save Entity List page (Find Similar Entities (Step 3 of 3))*

Filter on Parameters

This filter allows you to find entities that show a correlation with any of the parameter values where the parameter type is specified as numeric in the experiment grouping; i.e., an independent variable that contains numeric parameter values.

The similarity metrics available for this filter are **Euclidean** (finds similar entities based on the vector distance between two entities), **Pearson** (finds similar entities based on the trend and rate of change of trend of the chosen entity), and **Spearman** (finds similar entities based on the trend of the chosen entity).

Euclidean - Calculates the Euclidean distance where the vector elements are the columns. The square root of the sum of the square of the A and the B vectors for each element is calculated and then the distances are scaled between -1 and +1. Result = (A-B) · (A-B).

Pearson Correlation - Calculates the mean of all elements in vector X. Then it subtracts that value from each element in X and calls the resulting vector A. It does the same for Y to make a vector B. Result = A · B/(|A| · |B|)

Spearman Correlation - Orders all the elements of vector X and uses this order to assign a rank to each element of X. It makes a new vector X' where the i-th element in X' is the rank of Xi in X and then makes a vector A from X' in the same way as A was made from a in the Pearson Correlation. Similarly, it makes a vector B from Y. Result = A · B/(|A| · |B|). The advantage of using Spearman Correlation is that it reduces the effect of the outliers on the analysis.

The Filter on Parameters wizard has three (3) steps. The steps that you use are shown in Figure 161. The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

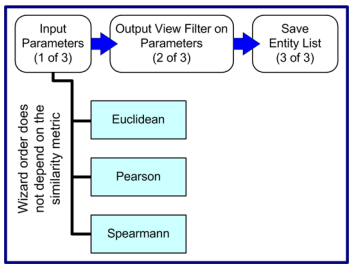


Figure 161 *Flow chart of the Filter on Parameters wizard.*

1. Launch the Filter on Parameters wizard in the Workflow Browser.

a Click **Filter on Parameters** in the Workflow Browser. This operation is illustrated with data from the "One-variable experiment" to provide an overview of the wizard options. The parameter type of the independent variable (parameter) was changed from non-numeric to numeric using the Experiment Grouping operation.

2. Select the input parameters in **Filter on Parameters (Step 1 of 3)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

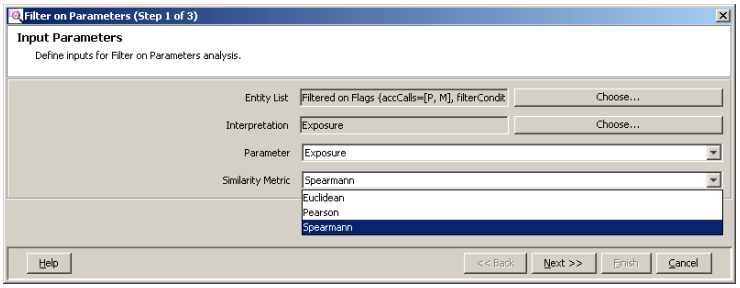


Figure 162 *Input Parameters page (Filter on Parameters (Step 1 of 3))*

c Select the **Parameter** to compare against your entity list.

d Select a **Similarity Metric**. The possible values are **Euclidean**, **Pearson**, and **Spearman**.

e Click **Next**.

3. Review and adjust your results in **Filter on Parameters (Step 2 of 3)**.

Visualize the results of your analysis in a profile plot. The abundance profile of the target parameter is shown in bold and along with the profiles of the entities whose correlation coefficients are within the target profile. The number of entities that meet your criteria and the current cut-off values are shown above the profile plot.

a Move the slider or type the **Minimum** cut-off value. The default value is 0.95.

To re-adjust minimum cut-off move the slider **Minimum** cut-off value until the results displayed are satisfactory. It is recommended that the analysis be re-run

several times to develop an understanding of how the cut-off affects the results. A smaller minimum cut-off value passes more entities.

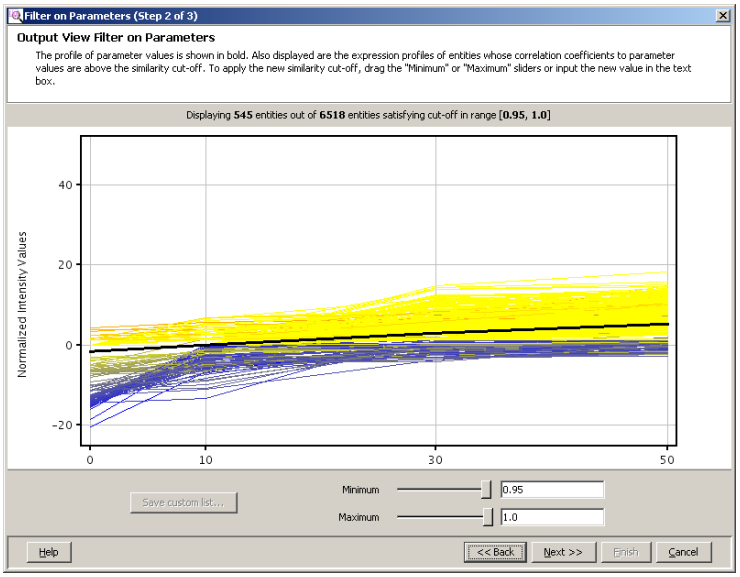


Figure 163 *Output View Filter on Parameters page (Filter on Parameters (Step 2 of 3))*

b Move the slider or type the **Maximum** cut-off value. The default value is 1.0.

To re-adjust the maximum cut-off may move the slider **Maximum** cut-off value until the results displayed are satisfactory. It is recommended that the analysis be re-run several times to develop an understanding of how the cut-off affects the results. A smaller maximum cut-off value passes more entities. The Maximum cut-off value cannot be smaller than the Minimum cut-off value.

c *(Optional)* Create a custom entity list as described in "Review your results in Statistical Analysis (Step 8 of 9)." on page 139.

d Click **Next**.

---

4. Enter save entity list parameters in **Filter on Parameters (Step 3 of 3)**.

The review content and parameters you enter in the **Save Entity List** page are the same as presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

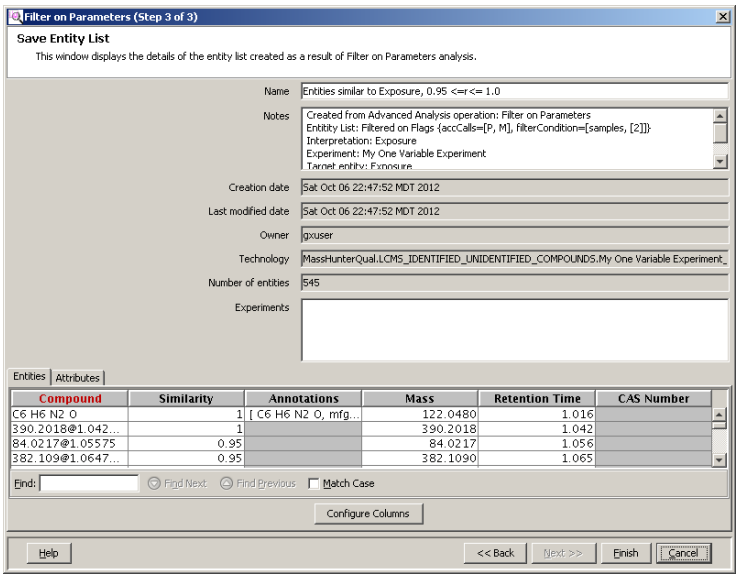


Figure 164 *Save Entity List page (Filter on Parameters (Step 3 of 3))*

Principal Component Analysis

Principal Component Analysis (PCA) facilitates viewing your data using a separation method. This is useful for data containing thousands of pieces of information (features) for each sample. When redundancy in features exists in the information PCA reduces the dimensionality of the input data to strengthen the separation.

Viewing data in 2 or 3 dimensions is easier than viewing it in multiple dimensions. Using PCA, you can overlook the less important dimensions or combine several dimensions to create a manageable number of dimensions.

PCA detects the major trends in your data by linearly combining dimensions. Each linear combination produces an Eigen vector that in short represents a fraction of the variability of the samples. The linear combinations (called Principal Axes or Components) are ordered in decreasing order of the associated Eigen value. Typically, two or three of the top few linear combinations in this ordering serve as a very good set of dimensions to project and view the data in a PCA plot. These dimensions capture most of the variability or information in the sample data.

The Principal Component Analysis wizard has four (4) steps. The steps that you use are shown in Figure 165. The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

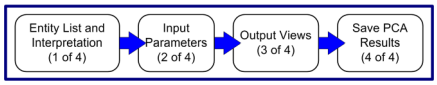


Figure 165 *Flow chart of the Principal Component Analysis wizard.*

1. Launch the Principal Component Analysis wizard in the Workflow Browser.

a Click **Principal Component Analysis** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment" to provide an overview of the wizard options.

2. Select the entity list and interpretation in **PCA (Step 1 of 4)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c Click **Next**.

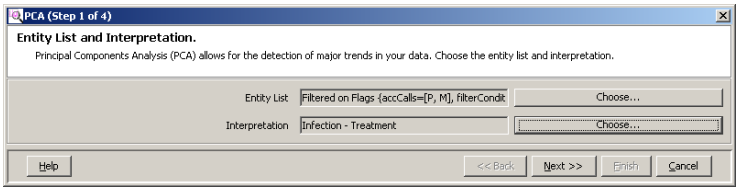


Figure 166 *Entity List and Interpretation page (PCA (Step 1 of 4))*

3. Select the input parameters in **PCA (Step 2 of 4)**.

Typically, only the first few Eigen vectors (principal components) capture most of the variation in the data. The execution speed of the PCA algorithm can be greatly enhanced only when a few Eigen vectors are computed. The pruning option determines how many Eigen vectors are computed. You can explicitly specify the exact number by selecting **Number of principal components** option.

Alternatively, you can specify that the PCA algorithm compute as many Eigen vectors as required to capture your specified **Total percentage variation** in the data. The default percentage variation is `100`.

a Select **Conditions** for the **PCA on** value. The possible values are **Entities** and **Conditions**.

b Click **Number of principal components** under the Pruning option.

c Type `4` in the entry box. The default value is `4`.

d Mark the **Scale** check box.

The normalization options let you normalize all columns to zero mean and unit standard deviation before performing PCA. This is enabled by default. It is recommended to use these normalization options if the range of values in the data columns varies widely.

e Click **Next**.

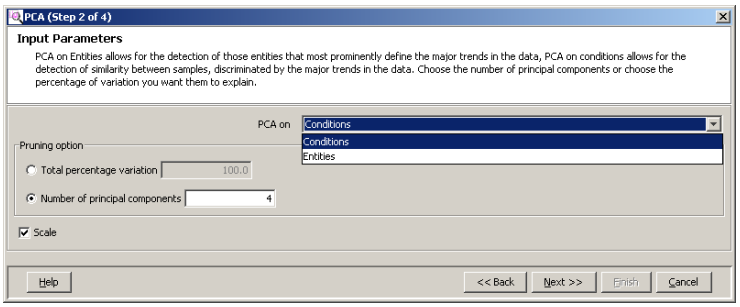


Figure 167 *Input Parameters page (PCA (Step 2 of 4))*

4. Review your results in **PCA (Step 3 of 4)**.

a Review the results of your analysis using the summary charts available in the Output views.

Eigen vectors: This chart is on the lower right of the page. The computed Eigen vectors are displayed on the X-axis and their respective percentage contribution to the sample variability is displayed on the Y-axis. The minimum number of principal axes required to capture most of the information in the data can be gauged from this plot. The red line indicates the actual variation captured by each Eigen vector. The blue line indicates the cumulative variation captured by all Eigen vectors up to that point. In the two-variable data shown in Figure 172 on page 169 100% of the variability is captured in the first three of the four Eigen vectors.

The minimum value for a PCA Eigen vector is $(1 \times 10^{-3}) /$ (total number of Principal components).

The maximum value is the square root of the maximum float value handled by your PC.

PCA Scores Tab: A scatter plot of data projected along the principal component axes. By default, the first and second PCA components capture the maximum variation of the data. If the data set has more than one parameter value, the points are colored with respect to the parameter values and help you visualize the separation of the parameter values in the data. You can select different PCA components (Eigen vectors) using the X-axis and Y-axis selections.

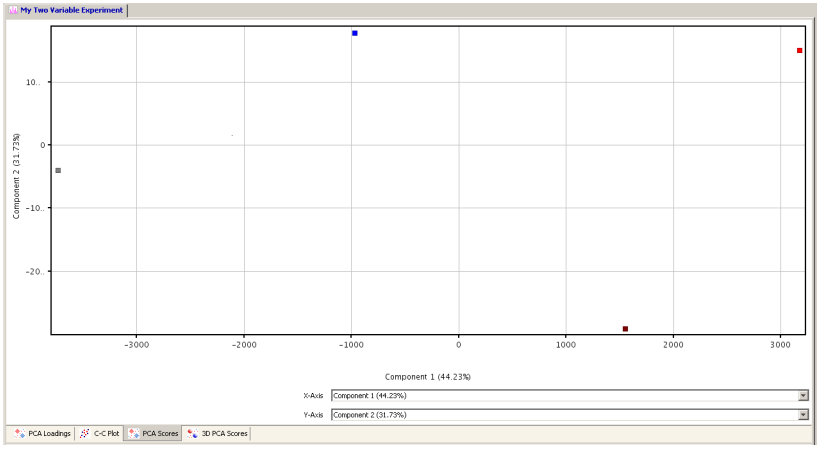


Figure 168 *General appearance of a PCA Scores Tab*

3D PCA Scores Tab: An interactive plot of the first, second, and third PCA components that capture the maximum variation of the data.

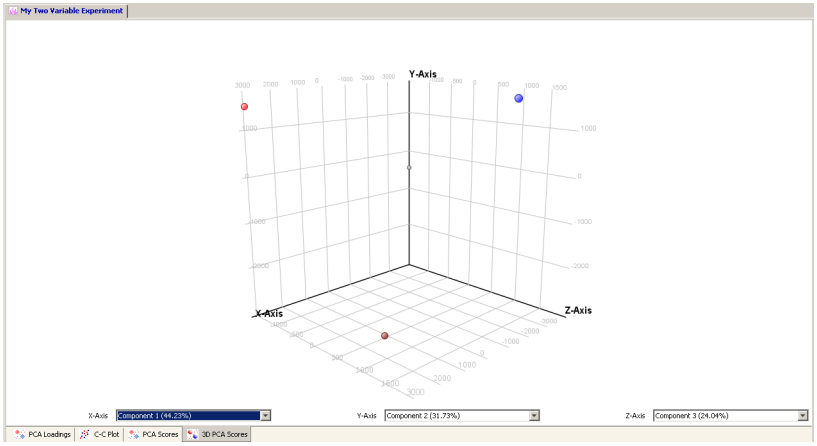


Figure 169 *General appearance of a 3D PCA Scores Tab*

PCA Loadings Tab: Each principal component is a linear combination of the selected columns. The relative contribution of each column to an Eigen vector is called its loading and is depicted in the PCA loadings plot. Each Eigen vector is plotted as a profile, and it is possible to visualize whether a certain subset of entities exist that overwhelmingly contribute (large absolute value of weight) to an important Eigen vector. The loadings plot can help identify which samples show greater variation in their entity measurements. You can select entities and save then using the **Save custom list** button.

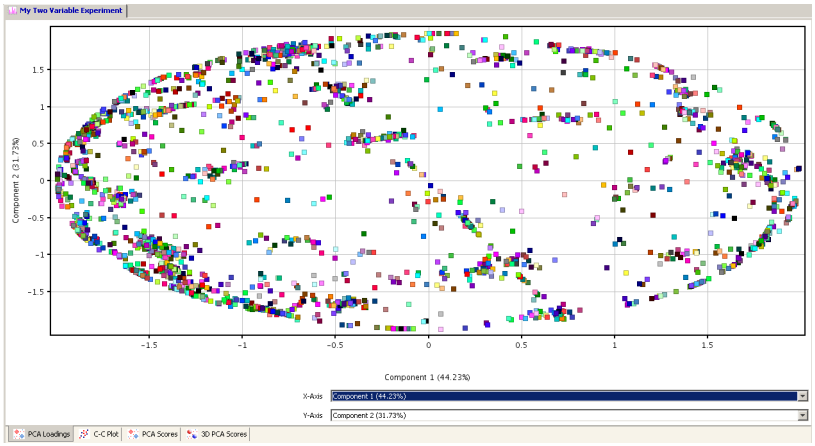


Figure 170 *General appearance of a PCA Loadings Tab*

C-C Plot Tab: This plot combines the covariance and correlation loading profiles resulting from the PCA or PLS-DA model in a scatter plot. In this plot both magnitude (covariance) and reliability (correlation) are visualized. You can select entities and save them using the **Save custom list** button.

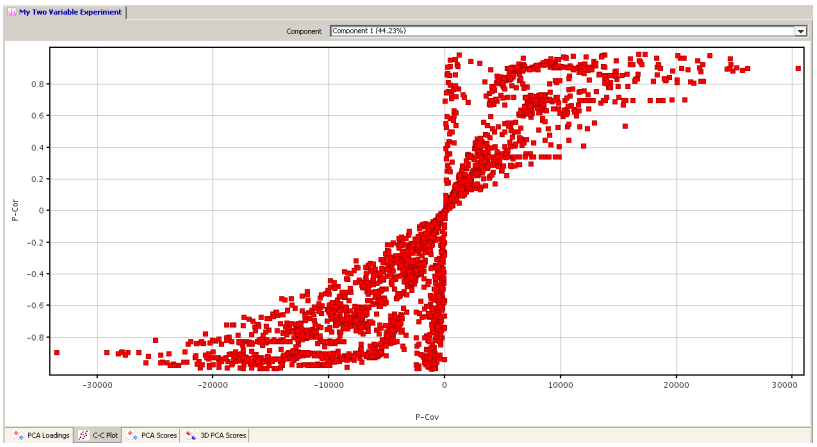


Figure 171 *General appearance of a C-C Plot Tab*

Legend: This shows the legend for the active tab.

b Click **Back**, make changes to the parameters, and click **Next** to return to this view.

c Click **Next**.

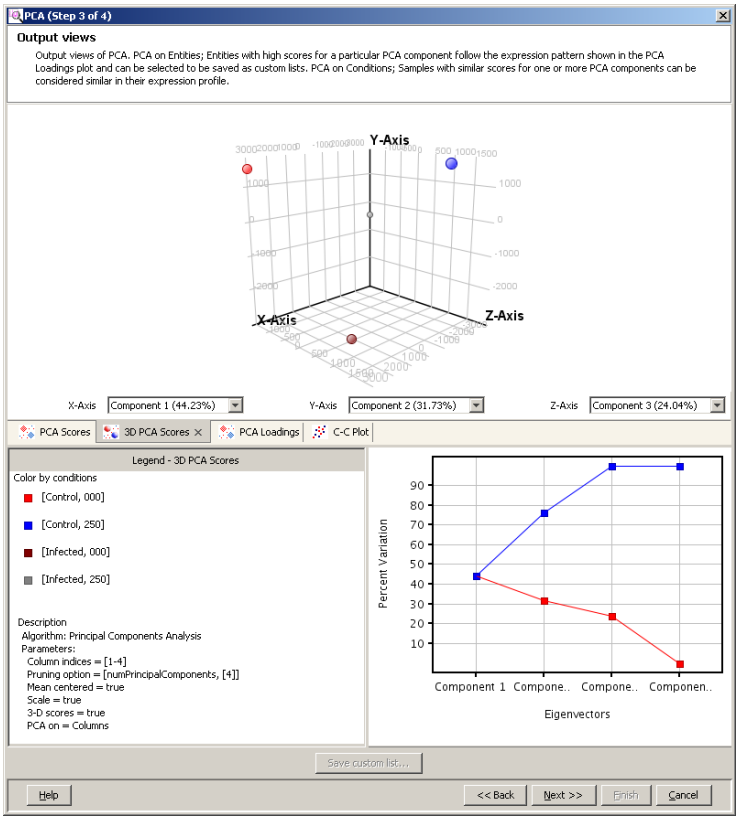


Figure 172 *Output Views page (PCA (Step 3 of 4))*

5. Enter save entity list parameters in **PCA (Step 4 of 4)**.

The review content and parameters you enter in the **Save PCA Results** page are similar to those presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

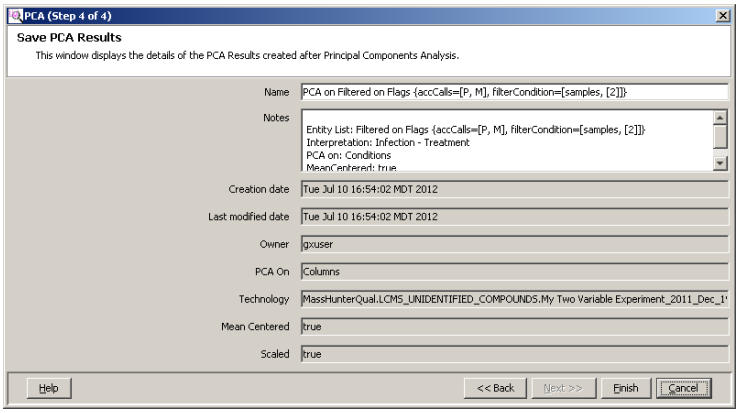


Figure 173 *Save PCA Results page (PCA (Step 4 of 4))*

Find Minimal Entities

Find Minimal Entities is a machine learning technique that analyzes your data to identify the specific features or attributes that are helpful in making a decision. Find Minimal Entities becomes relevant when you have a large number of features and there is strong intercorrelation among the features.

In Mass Profiler Professional, feature subset selection is done using either a Forward Selection Algorithm, Backward Elimination Algorithm, or a Genetic Algorithm.

The Find Minimal Entities wizard has ten (10) steps. The steps that you use depend on the selection algorithm chosen in step 2 as shown in Figure 174. The new entity list is placed in the Analysis folder within the Experiment Navigator. More than one entity list may be created from your analysis.

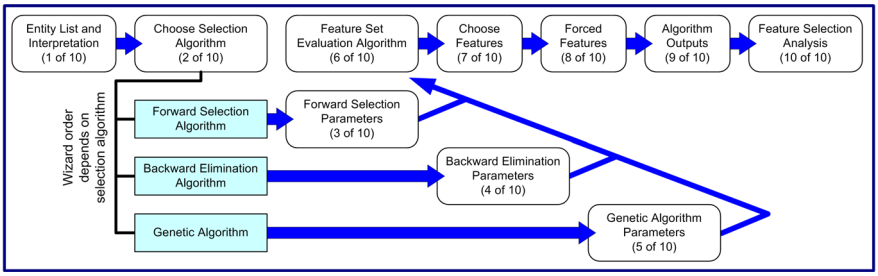


Figure 174 *Flow chart of the Find Minimal Entities wizard.*

1. Launch the Find Minimal Entities wizard in the Workflow Browser.

a Click **Find Minimal Entities** in the Workflow Browser. This operation is illustrated with data from the "Two-variable experiment" to provide an overview of the wizard options.

2. Select the entity list and interpretation in **Find Minimal Entities (Step 1 of 10)**.

a Click **Choose** to select the **Entity List**. By default, the active entity list is selected and shown in the dialog. It is recommended to select an entity list that has been at least filtered on flags.

b Click **Choose** to select the **Interpretation**. By default, the active interpretation of the experiment is selected and shown in the dialog.

c Click **Next**.

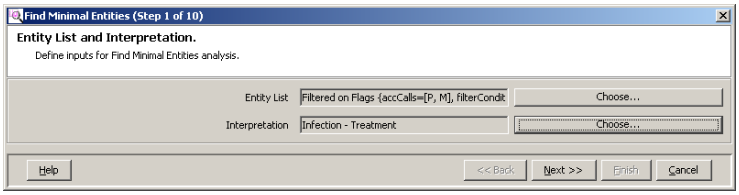


Figure 175 *Entity List and Interpretation page (Find Minimal Entities (Step 1 of 10))*

3. Select the algorithm in **Find Minimal Entities (Step 2 of 10)**.

a Select a **Choose selection algorithm**. The possible values are **Forward selection algorithm**, **Backward elimination algorithm**, and **Genetic algorithm**.

If you click **Forward selection algorithm**, then after you click **Next,** the Forward Selection Parameters page (Step 3 of 10) is displayed (page 171).

If you click **Backward elimination algorithm**, then after you click **Next,** the Backward Elimination Parameters page (Step 4 of 10) is displayed (page 172).

If you click **Genetic algorithm**, then after you click **Next**, the Genetic Algorithm Parameters page (Step 5 of 10) is displayed (page 173).

b Click **Next**.

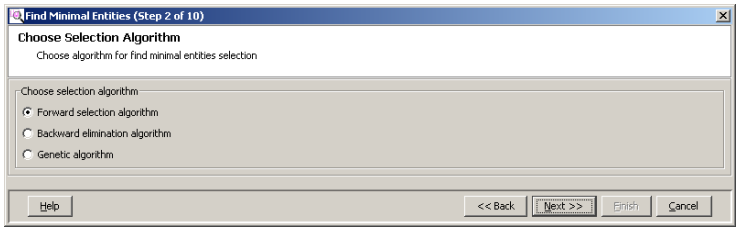


Figure 176 *Choose Selection Algorithm page (Find Minimal Entities (Step 2 of 10))*

4. Select and enter the forward selection algorithm parameters in **Find Minimal Entities (Step 3 of 10)**.

If you selected the Forward Selection Algorithm in the Choose Selection Algorithm page (Find Minimal Entities (Step 2 of 10)), then this step is where you specify the algorithm parameters.

a Type the **Target size for features**. You must enter the target size of the features. By default the value shown is one tenth of the total entities. This value must equal the number of features you select in "Choose your features in Find Minimal Entities (Step 7 of 10)." on page 177.

b Select an **Evaluation metric**. The possible values are **Overall Accuracy** and **Min Class Accuracy**.

c Type the **Target metric value**. The default value is 100.

d Select an **Evaluation algorithm**. Select one of the evaluation algorithms to build a class. The possible values are **Naive Bayes**, **Neural Network**, **Support Vector Machine**, and **Axis Parallel Decision Tree**. See section 6.5 Class Prediction in the *Mass Profiler Professional User Manual* for more information about these classification algorithms.

e Select an **Evaluation metric type**. The possible values are **Validation Accuracy** and **Train Accuracy**.

f Click **Next**.

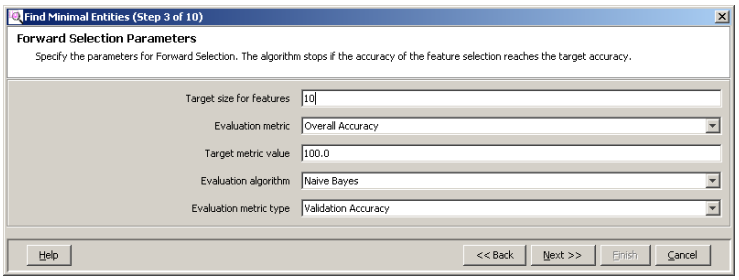


Figure 177 *Forward Selection Parameters page (Find Minimal Entities (Step 3 of 10))*

5. Select and enter the backward elimination algorithm parameters in **Find Minimal Entities (Step 4 of 10)**.

If you selected the Backward Elimination Algorithm in the Choose Selection Algorithm page (Find Minimal Entities (Step 2 of 10)), then this step is where you specify the algorithm parameters. The parameters are the same as in Forward Selection Algorithm.

Backward elimination begins with a full set of features and removes one feature per cycle. The resulting set, with one less feature than the earlier one is evaluated, and the algorithm selects the highest performing candidate (again subject to a specified validation metric). If an empty set is reached or the subsequent removal of any feature only deteriorates the current performance, the search is stopped.

The goal of backward elimination is to consider the contribution of all features initially and then try to remove the most irrelevant features, leaving a smaller and more predictive subset. Since backward elimination starts with all of the features present in the set, it is more computationally intensive than the forward selection algorithm.

a Type the **Target size for features**. You must enter the target size of the features. By default the value shown is one tenth of the total entities. This value must equal the number of features you select in "Choose your features in Find Minimal Entities (Step 7 of 10)." on page 177.

b Select an **Evaluation metric**. The possible values are **Overall Accuracy** and **Min Class Accuracy**.

c Type the **Target metric value**. The default value is 100.

d Select an **Evaluation algorithm**. Select one of the evaluation algorithms to build a class. The possible values are **Naive Bayes**, **Neural Network**, **Support Vector Machine**, and **Axis Parallel Decision Tree**. See section 6.5 Class Prediction in the *Mass Profiler Professional User Manual* for more information about these classification algorithms.

e Select an **Evaluation metric type**. The possible values are **Validation Accuracy** and **Train Accuracy**.

f Click **Next**.

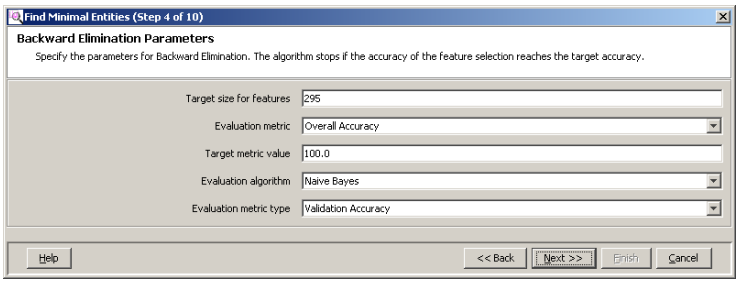


Figure 178 *Backward Elimination Parameters page (Find Minimal Entities (Step 4 of 10))*

---

6. Select and enter the genetic algorithm parameters in **Find Minimal Entities (Step 5 of 10)**.

If you selected the Genetic Algorithm in the Choose Selection Algorithm page (Find Minimal Entities (Step 2 of 10)), then this step is where you specify the algorithm parameters. More information about this algorithm may be reviewed in the Genetic Algorithm Parameters section of the *Mass Profiler Professional User Manual*.

a Type the **Population size**. The default is 25.

b Type the **Number of generations**. The default is 10.

c Type the **Mutation rate**. The default is 1.

d Type the **Target size (max) for features**. You must enter the target size of the features. By default the value shown is one tenth of the total entities. This value must equal the number of features you select in "Choose your features in Find Minimal Entities (Step 7 of 10)." on page 177.

e Select a **Fitness metric**. The possible values are **Overall Accuracy** and **Min. Class Accuracy**.

f Type the **Target fitness value**. The default value is 100.

g Select an **Evaluation algorithm**. Select one of the evaluation algorithms to build a class. The possible values are **Naive Bayes**, **Neural Network**, **Support Vector Machines**, and **Axis Parallel Decision Tree**. See section 6.5 Class Prediction in the *Mass Profiler Professional User Manual* for more information about these classification algorithms.

h Select the **Fitness metric type**. The possible values are **Validation Accuracy** and **Train Accuracy**.

i Click **Next**.

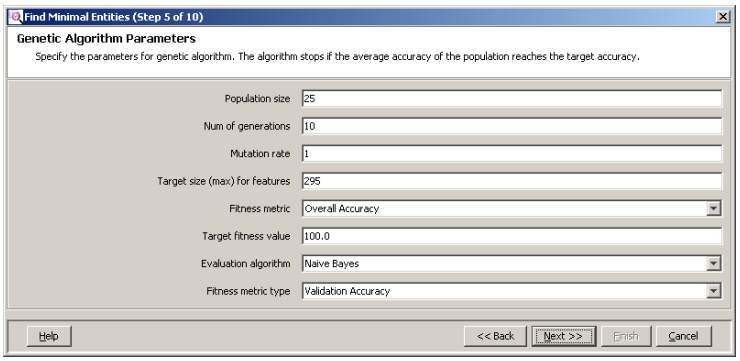


Figure 179 *Genetic Algorithm Parameters page (Find Minimal Entities (Step 5 of 10))*

7. Enter the feature set evaluation algorithm parameters in **Find Minimal Entities (Step 6 of 10)**.

The evaluation parameters you enter on this page vary depending on the evaluation algorithm selected.

Naive Bayes parameters

If you selected **Naive Bayes** enter the following parameters.

a Select a **Distribution type**.

b Select a **Validation type**. The possible values are **N-Fold** and **Leave One Out**.

c Type the **Number of folds**. The default value is 3.

d Type the **Number of repeats**. The default value is 10.

e Click **Next**.

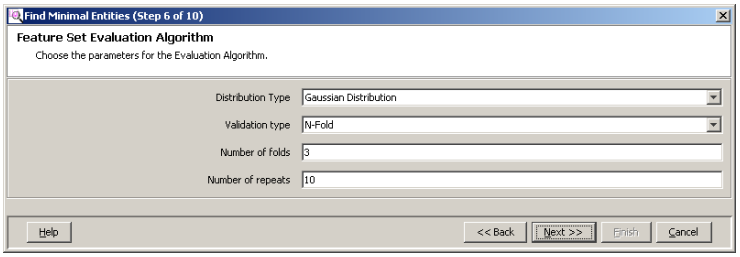


Figure 180 *Feature Set Evaluation Algorithm page (Find Minimal Entities (Step 6 of 10)) - Naive Bayes*

Neural Network parameters

If you selected **Neural Network** enter the following parameters.

a Type the **Number of iterations**.

b Type the **Learning rate**.

c Type the **Momentum**.

d Select a **Number of layers**. The default value is 3.

e *Optional*. Click **Set Neurons** and then **OK** when you completed typing in your values.

f Select a **Validation type**. The possible values are **N-Fold** and **Leave One Out**.

g Type the **Number of folds**. The default value is 3.

h Type the **Number of repeats**. The default value is 10.

i Click **Next**.

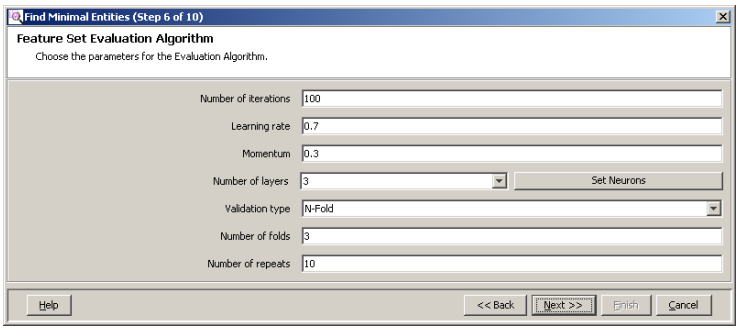


Figure 181 *Feature Set Evaluation Algorithm page (Find Minimal Entities (Step 6 of 10)) - Neural Network*

Support Vector Machines parameters

If you selected **Support Vector Machines** enter the following parameters.

a Select a **Kernel type**. The possible values are **Linear**, **Polynomial**, and **Gaussian**.

b Type the **Maximum number of iterations**. The default value is 100000.

c Type the **Cost**. The default value is 100.

d Type the **Ratio**. The default value is 1.0.

e If you selected a Kernel type Linear skip to the Validation type.

f If you selected a Kernel type Polynomial type in Kernel parameter1, Kernel parameter 2, and Exponent values then skip to the Validation type.

g If you selected a Kernel type Gaussian type in Sigma value.

h Select a **Validation type**. The possible values are **N-Fold** and **Leave One Out**.

i Type the **Number of repeats**. The default value is 10.

j Click **Next**.

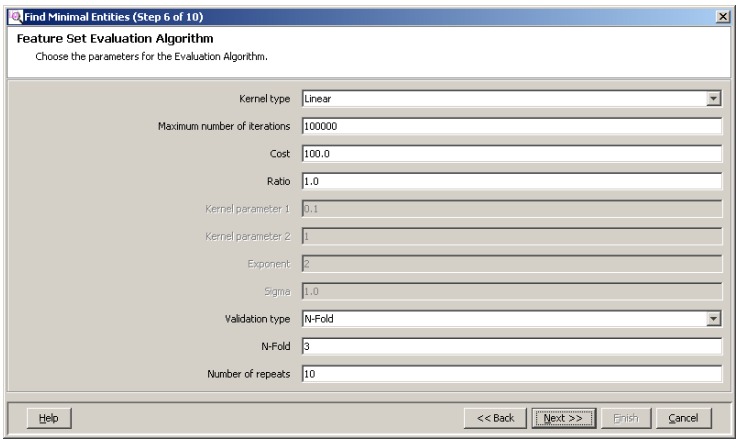


Figure 182 *Feature Set Evaluation Algorithm page (Find Minimal Entities (Step 6 of 10)) - Support Vector Machines*

Axis Parallel Decision Tree parameters

If you selected **Axis Parallel Decision Tree** enter the following parameters.

a Select a **Pruning method**. The possible values are **Minimum Error**, **Pessimistic Error**, and **None**.

b Select a **Goodness function**. The possible values are **Gini** and **Information Gain**.

c Type the **Leaf impurity**. The default value is 1.0.

d Select the **Leaf impurity type**. The possible values are **Global** and **Local**.

e Select a **Validation type**. The possible values are **N-Fold** and **Leave One Out**.

f Type the **Number of folds**. The default value is 3.

g Type the **Number of repeats**. The default value is 10.

h Type the **Attribute Fraction at nodes**. The default value is 1.0.

i Click **Next**.

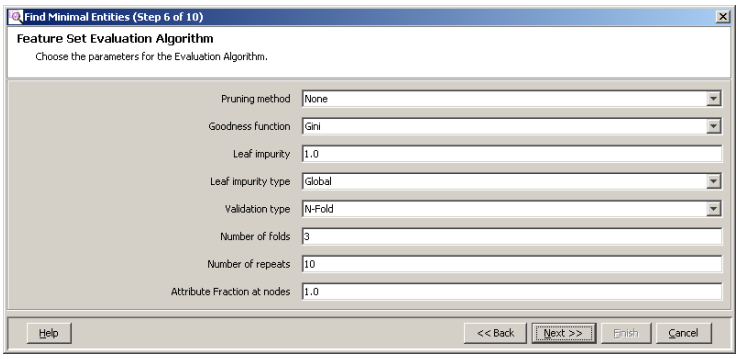


Figure 183 *Feature Set Evaluation Algorithm page (Find Minimal Entities (Step 6 of 10)) - Axis Parallel Decision Tree*

8. Choose your features in **Find Minimal Entities (Step 7 of 10)**.

Select features from the available items list and click on the movement direction arrow to place them in the **Selected items** list. You can also search for an item from the **Match by item** drop-down list. By default, all of the features are selected.

a Select ten (10) features as an example and to reduce the analysis time.

b Click **Next**.

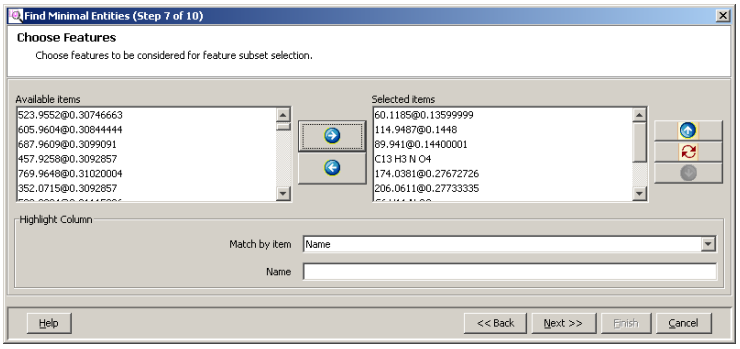


Figure 184 *Choose Features page (Find Minimal Entities (Step 7 of 10))*

9. Choose forced features in **Find Minimal Entities (Step 8 of 10)**.

Select features from the available items list and click on the movement direction arrow to place them in the **Selected items** list. You can also search for an item from the **Match by item** drop-down list. By default, none of the feature are selected.

a Select two (2) features from your prior selection as an example. The available features are only those selected in the previous step.

b Click **Next**.

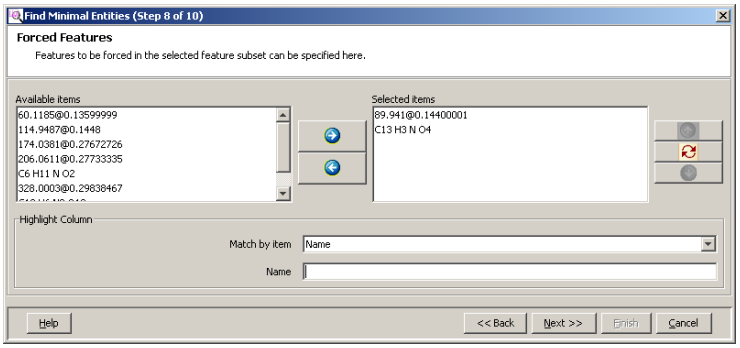


Figure 185 *Forced Features page (Find Minimal Entities (Step 8 of 10))*

10. Review the algorithm outputs in **Find Minimal Entities (Step 9 of 10)**.

The output of the algorithm used in the find minimal mass process is displayed in a report spreadsheet as well as a **Feature vs. Accuracy** plot. In the plot the drop-down list of X and Y axis lets you select different combination of the parameters and view the plot.

The spreadsheet displays the columns, **Number of descriptors, Train Overall Accuracy, Train Accuracy in condition 1, Train Accuracy in condition 2, Minimum of Train Accuracy, Validation Overall Accuracy, Validation Accuracy in condition 1, Validation Accuracy in condition 2,** and **Minimum of Validation Accuracy**. This also consists of options of Save Custom Feature Set and Run Model.

a To open and save the entities in the entity list select the a descriptor row and click the **Save Custom Feature Sets** icon.

b To make a prediction with help of training accuracies select a descriptor row and click the **Run Model** icon. The model displays the prediction result in several tabs: Confusion Matrix, Lorenz Curve, Model Formula, and Validation Report.

c Review your results.

d Click **Next**.

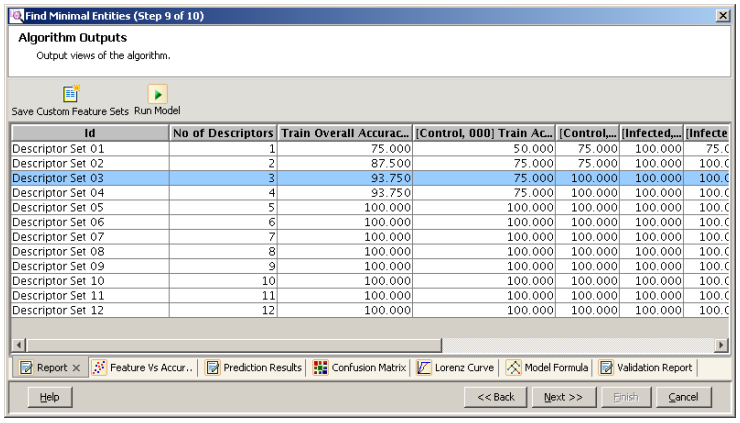


Figure 186 *Algorithm Outputs page (Find Minimal Entities (Step 9 of 10))*

11. Enter feature selection analysis parameters in **Find Minimal Entities (Step 10 of 10)**.

The review content and parameters you enter in the **Save PCA Results** page are similar to those presented in the Significance Testing and Fold Change workflow.

a Follow the steps presented in "The Save Entity List page" on page 124.

b Click **Finish**.

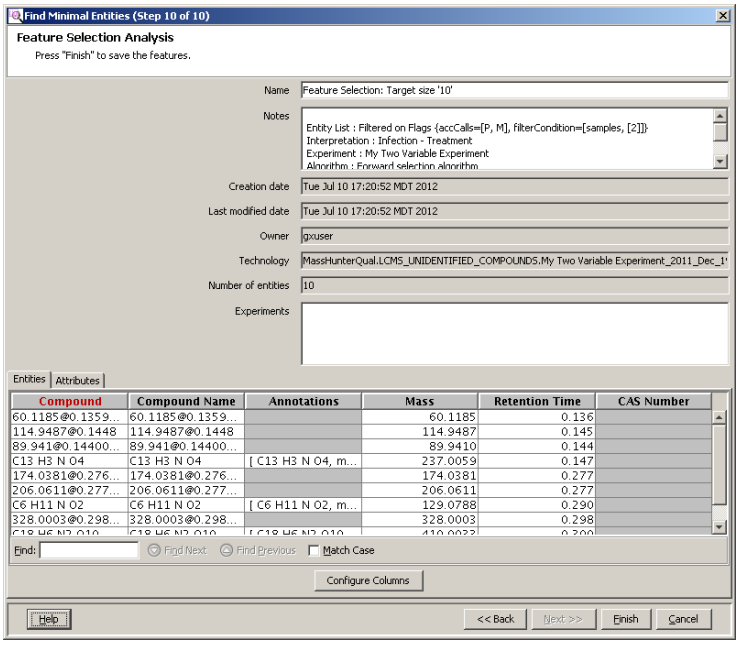
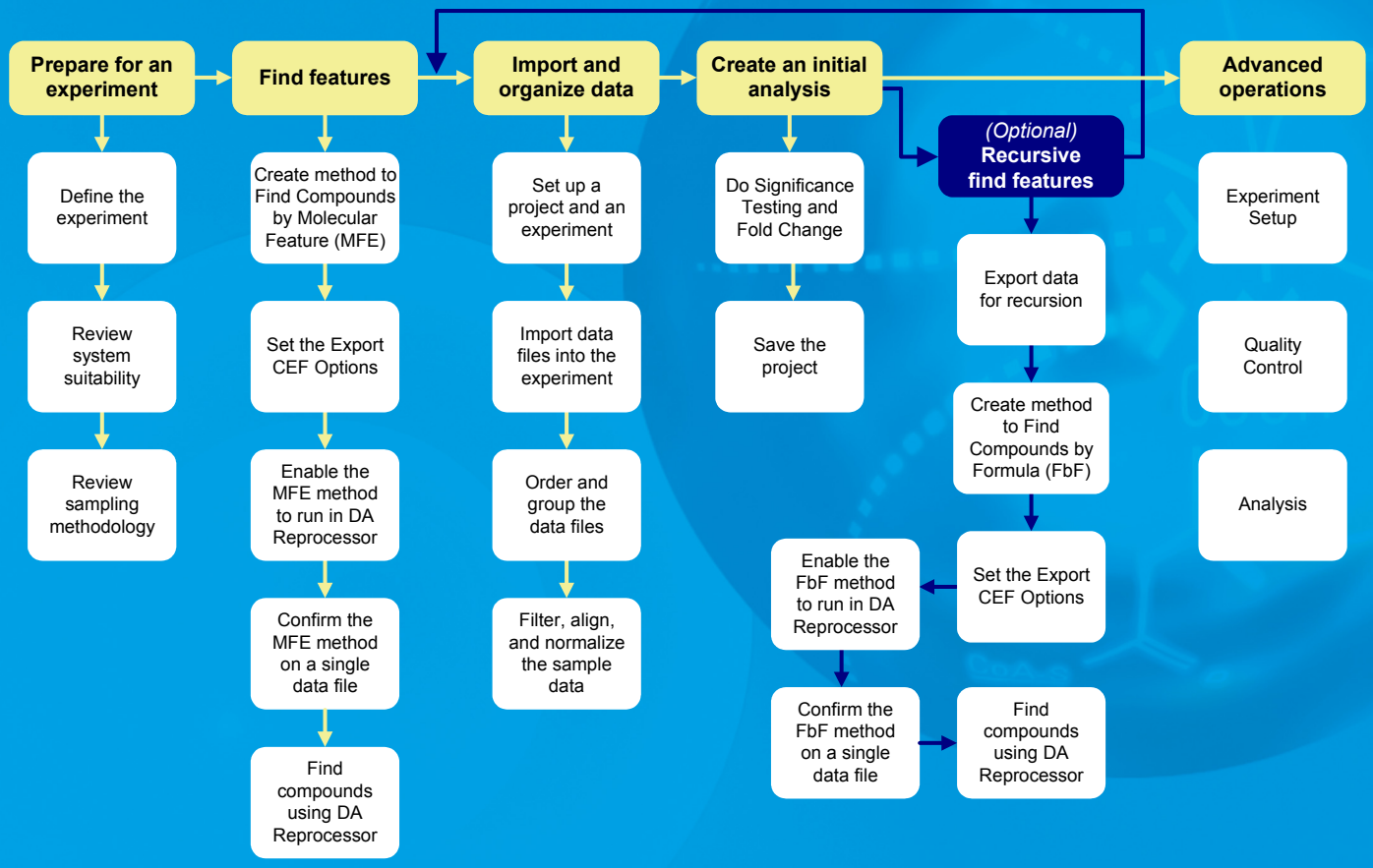


Figure 187 *Feature Selection Analysis page (Find Minimal Entities (Step 10 of 10))*

Reference information

This chapter consists of definitions and references. The definitions section includes a list of terms and their definitions as used in this workflow. The references section includes citations to Agilent publications that help you use Agilent products and perform your metabolomics analyses.



Agilent Technologies

Definitions

This section contains a list of terms and their definitions as used in this workflow. Review of the terms and definitions presented in this section helps you understand the Agilent software wizards and the metabolomics workflow.

Alignment

Adjustment of the chromatographic retention time of eluting components to improve the correlation among data sets, based on the elution of specific component(s) that are (1) naturally present in each sample or (2) deliberately added to the sample through spiking the sample with a known compound or set of compounds that does not interfere with the sample.

AMDIS

Acronym for automated mass spectral deconvolution and identification system developed by NIST (http://www.amdis.net).

Amino acid

Biologically significant molecules that contain a core carbon positioned between a carboxyl and amine group in addition to an organic substituent. Dual carboxyl and amine functionalities facilitate the formation of peptides and proteins.

ANOVA

Abbreviation for analysis of variance which is a statistical method that simultaneously compares the means between two or more attributes or parameters of a data set. ANOVA is used to determine if a statistical difference exists between the means of two or more data sets and thereby prove or disprove the hypothesis. See also t-Test.

Attribute

Another term for an independent variable. Referred to as a parameter and is assigned a parameter name during the various steps of the metabolomic data analysis.

Attribute value

Another term for one of several values within an attribute for which exist correlating samples. Referred to as a condition or a parameter value and given an assigned value during the various steps of the metabolomic data analysis.

Baselining

A technique used to view and compare data that involves converting the original data values to values that are expressed as changes relative to a calculated statistical value derived from the data. The calculated statistical value is referred to as the baseline.

Bayesian

A term used to refer to statistical techniques named after the Reverend Thomas Bayes (ca. 1702 - 1761).

Bayesian inference

The use of statistical reasoning, instead of direct facts, to calculate the probability that a hypothesis may be true. Also known as Bayesian statistics.

Bioinformatics

The use of computers, statistics, and informational techniques to increase the understanding of biological processes.

Biomarker

An organic molecule whose presence and concentration in a biological sample indicates a normal or altered function of higher level biological activity.

Carbohydrate

An organic molecule consisting entirely of carbon, hydrogen, and oxygen that is important to living organisms.

CEF file

A binary file format called a compound exchange file (CEF) that is used to exchange data between Agilent software. In the metabolomics workflow CEF files are used to share molecular features between MassHunter Qualitative Analysis and Mass Profiler Professional.

Cell

The fundamental unit of an organism consisting of several sets of biochemical functions within an enclosing membrane. Animals and plants are made of one or more cells that combine to form tissues and perform living functions.

Census

Collection of a sample from every member of a population.

Cheminformatics

The use of computers and informational techniques (such as analysis, classification, manipulation, storage, and retrieval) to analyze and solve problems in the field of chemistry.

Chemometrics

A science employing mathematical and analytical processes to extract information from chemical data sets. The processes involve interactive applications of techniques employed in disciplines such as multivariate statistics, applied mathematics, and computer science to obtain meaningful information from complex data sets. Chemometrics is typically used to obtain meaningful information from data derived from chemistry, biochemistry and chemical engineering. Agilent Mass Profiler Professional is designed to employ chemometrics processes to GC/MS and LC/MS data sets to obtain useful information.

Child

A subset of information that is created by an algorithm from an original set of information. An entity list created using Mass Profiler Professional is a child. An original entity list is referred to as the parent of one or more child entity lists.

Co-elution

When compounds elute from a chromatographic column at nominally the same time making the assignment of the observed ions to each compound difficult.

Composite spectrum

A compound spectrum generated to represent the molecular feature that includes more than one ion, isotope, or adduct (not just M + H) and is used by Mass Profiler Professional for recursive analysis and ID Browser.

Compound

A metabolite that may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.

Condition	Another term for one of several values within a parameter for which exist correlating samples. Condition may also be referred to as a parameter value during the various steps of the metabolomic data analysis. See also attribute value.
Data	Information in a form suitable for storing and processing by a computer that represent the qualitative or quantitative attributes of a subject. Examples include GC/MS and LC/MS data consisting fundamentally of time, ion m/z, and ion abundance from a chemical sample.
Data processing	Conversion of data into meaningful information. Computers are employed to enable rapid recording and handling of large amounts of data, i.e. Agilent MassHunter Workstation and Agilent Mass Profiler Professional.
Data reduction	See reduction.
Deconvolution	The technique of reconstructing individual mass and mass spectral data from co-eluting compounds.
Dependent variable	An element in a data set that can only be observed as a result of the influence from the variation of an independent variable. For example, a pharmaceutical compound structure and quantity may be controlled as two independent variables while the metabolite profile presents a host of small-molecule products that make up the dependent variables of a study.
Determinate	Having exact and definite limits on an analytical result that provide a conclusive degree of correlation of the subject to the specimen.
Element	A metabolite that may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.
Endogenous	Pertaining to cause, development, or origination from within an organism.
Entity	A metabolite that may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.
Experiment	Data acquired in an attempt to understand causality where tests or analyses are defined and performed on an organism to discover something that is not yet known, to demonstrate as proof of something that is known, or to find out whether something is effective.
Externality	A quality, attribute, or state that originates and/or is established independently from the specimen under evaluation.

Extraction	The process of retrieving a deliberate subset of data from a larger data set whereby the subset of the data preserves the meaningful information as opposed to the redundant and less meaningful information. Also known as data extraction.
Feature	Independent, distinct characteristic of a phenomena and data under observation. Features are an important part of the identification of patterns - pattern recognition - within data whether processed by a human or by artificial intelligence, such as Agilent MassHunter Workstation and Agilent Mass Profiler Professional. In metabolomics analysis a feature is a metabolite and may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.
Feature extraction	The reduction of data size and complexity through the removal of redundant and non-specific data by using the important variables (features) associated with the data. Careful feature extraction yields a smaller data set that is more easily processed without any compromise in the information quality. This is part of the principal component analysis process employed by Agilent Mass Profiler Professional.
Feature selection	The identification of important, or non-important, variables and the variable relationships in a data set using both analytical and a priori knowledge about the data. This is part of the principal component analysis process employed by Agilent Mass Profiler Professional.
Filter	The process of establishing criteria by which entities are removed (filtered) from further analysis during the metabolomics workflow.
Filter by flag	A flag is a term used to denote a quality of an entity within a sample. A flag indicates if the entity was detected in each sample as follows: Present means the entity was detected, Absent means the entity was not detected, and Marginal means the signal for the entity was saturated
Hypothesis	A proposition made to explain certain facts and tentatively accepted to provide a basis for further investigation. A proposed explanation for observable phenomena may or may not be supported by the analytical data. Statistical data analysis is performed to quantify the probability that the hypothesis is true. Also known as the scientific hypothesis.
Hypothetical	A statement based on, involving, or having the nature of a hypothesis for the purposes of serving as an example and not necessarily based on an actuality.
ID Browser	Agilent software that automatically annotates the entity list with the compound names and adds them to any of the various visualization and pathway analysis tools.
Identified compound	Chromatographic components that have an assigned, exact identity, such as compound name and molecular formula, based on prior assessment or comparison with a database. See also Unidentified Compound.

Independent variable

An essential element, constituent, attribute, or quality in a data set that is deliberately controlled in an experiment. For example, a pharmaceutical compound structure and quantity may be controlled as two independent variables while the metabolite profile presents a host of independent small molecule products that make up the dependent variables of a study. An independent variable may be referred to as a parameter and is assigned a parameter name during the various steps of the metabolomic data analysis.

Inorganic compound

Non carbon and non biological origin compounds such as minerals and salts.

Interpretation

Expression of your data in entity lists after grouping your samples, applying filters, and performing statistical correlation methods. When you open an experiment, the "All Samples" interpretation is active. You can click on another interpretation to activate it.

Lipidomics

Identification and quantification of cellular lipids from an organism in a specified biological situation. The study of lipids is a subset of metabolomics.

Mass variation

Using the mass to charge (m/z) resolution to improve compound identification. Compounds with nearly identical and identical chromatographic behavior are deconvoluted by adjusting the m/z range for extracting ion chromatograms.

Mean

The numerical result of dividing the sum of the data values by the number of individual data observations.

Metabolism

The chemical reactions and physical processes whereby living organisms convert ingested compounds into other compounds, structures, energy and waste.

Metabolite

Small organic molecules that are intermediate compounds and products produced as part of metabolism. Metabolites are important modulators, substrates, byproducts, and building blocks of many different biological processes. In order to distinguish metabolites from lager biological molecules, known as macromolecules such as proteins, DNA and others, metabolites are typically under 1000 Da. A metabolite may be individually referred to as a compound, molecular feature, element, or entity during the various steps of the metabolomic data analysis.

Metabolome

The complete set of small-molecule metabolites that may be found within a biological sample. Small molecules are typically in the range of 50 to 600 Da.

Metabolomics

The process of identification and quantification of all metabolites of an organism in a specified biological situation. The study of the metabolites of an organism presents a chemical "fingerprint" of the organism under the specific situation. See metabonomics for the study of the change in the metabolites in response to externalities.

Metabonomics	The metabolic response to externalities such as drugs, environmental factors, and disease. The study of metabonomics by the medical community may lead to more efficient drug discovery and to individualized patient treatment. Meaningful information learned from the metabolite response can be used for clinical diagnostics or for understanding the onset and progression of human diseases. See metabolomic for the identification and quantitation of metabolites.
Normalization	A technique used to adjust the ion intensity of mass spectral data from an absolute value based on the signal measured at the detector to a relative intensity of 0 to 100 percent based on the signal of either (1) the ion of the greatest intensity or (2) a specific ion in the mass spectrum.
Null hypothesis	The default position taken by the hypothesis that no effect or correlation of the independent variables exists with respect to the measurements taken from the samples.
Observation	Data acquired in an attempt to understand causality where no ability exists to (1) control how subjects are sampled and/or (2) control the exposure each sample group receives.
One-hit wonder	An entity that appears in only one sample, is absent from the replicate samples, and does not provide any utility for statistical analysis. Entities that are one-hit wonders may be filtered using Filter by Flags.
Organic compound	Carbon-based compounds, often with biological origin.
Organism	A group of biochemical systems that function together as a whole thereby creating an individual living entity such as an animal, plant, or microorganism. Individual living entities may be multicellular or unicellular. See also specimen.
p-value	The probability of obtaining a statistical result that is comparable to or greater in magnitude than the result that was actually observed, assuming that the null hypothesis is true. The null hypothesis is stated that no correlation exists between the independent variables and the measurements taken from the samples. Rejection of the null hypothesis is typically made when the p-value is less than 0.05 or 0.01. A p-value of 0.05 or 0.01 may be restated as a 5% or 1% chance of rejecting the null hypothesis when it is true. When the null hypothesis is rejected, the result is said to be statistically significant meaning that a correlation exists between the independent variables and the measurements as specified in the hypothesis.
Parameter	Another term for an independent variable. Referred to as a parameter or parameter name and is assigned a parameter name during the various steps of the metabolomic data analysis. See also condition and attribute.
Parameter value	Another term for one of several values within a parameter for which exist correlating samples. Parameter value may also be referred to as a condition during the various steps of the metabolomic data analysis. See also attribute value.

Parent

The original set of information that is processed by an algorithm to create one or more subsets of information. A subset entity list is referred to as the child of a parent entity list.

Peptide

Linear chain of amino acids that is shorter than a protein. The length of a peptide is sufficiently short that it is easily made synthetically from the constituent amino acids.

Peptide bond

The covalent bond formed by the reaction of a carboxyl group with an amine group between two molecules, e.g. between amino acids.

Permutation

Any of the total number of subsets that may be formed by the combination of individual parameters among the independent variables. For example the number of permutations of A and B in variable Φ in combination with X, Y, and Z in variable θ equals six (6 = 2 x 3) and may be represented as AX, AY, AZ, BX, BY, and BZ. Note that the combinations of parameters within a variable are not relevant such as AB, XY, XZ, and YZ.

Polarity

The condition of an effect as being positive or negative, additive or subtractive, with respect to some point of reference, such as with respect to the concentration of a metabolite.

Polymer

A molecule formed by the covalent bonding of a repeating molecular group to form a larger molecule.

Pooled sample

When the amount of available biological material is very small samples may be combined into a single sample (pooled) and then split into different aliquots for multiple analyses. By pooling the sample, sufficient material exists to obtain replicate analyses of each sample where formerly there was insufficient material to obtain replicate analytical results. The trade-off loss of information about the biological variation that was formerly present in each unique sample is offset by a gain in statistical significance of the results.

Principal component

Transformed data into axes, or principal components, so that the patterns between the axes most closely describe the relationships between the data. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components often may be viewed, and interpreted, most readily in graphical axes with additional dimensions represented by color and/or shape representing the key elements (independent variables) of the hypothesis. This is part of the principal component analysis process employed by Agilent Mass Profiler Professional.

Principal component analysis

The mathematical process by which data containing a number of potentially correlated variables is transformed into a data set in relation to a smaller number of variables called principal components which account for the most variability in the data. The result of the data transformation leads to the identification of the best explana-

	tion of the variance in the data, e.g. identification of the meaningful information. Also known as PCA.
Protein	Linear chain of amino acids whose amino acid order and three-dimensional structure are essential to living organisms. Also know as a polypeptide.
Proteomics	The study of the structure and function of proteins occurring in living organisms. Proteins are assemblies of amino acids (polypeptides) based on information encoded in the genes of an organism and are the main components of the physiological metabolic pathways of the organism.
Quality	A feature, attribute, and/or characteristic element whose presence, absence, or inability to be properly ascertained due to instrumental factors, is factored into whether a sample is or is not representative of the larger specimen.
Recursive	Reapplying the same algorithm to a subset of a previous result in order to generate an improved result.
Recursive finding	A three-step process in the metabolomics workflow that improves the accuracy of finding statistically significant features in sample data files. Step 1: Find untargeted compounds by molecular feature in MassHunter Qualitative Analysis. Step 2: Filter the molecular features in Mass Profiler Professional. Step 3: Find targeted compounds by formula in MassHunter Qualitative Analysis. Importing the most significant features identified using Mass Profiler Professional back into MassHunter Qualitative Analysis as targeted features improves the accuracy in finding these features from the original sample data files.
Reduction	The process whereby the number of variables in a data set is decreased to improve computation time and information quality. For example, an extracted ion chromatogram obtained from GC/MS and LC/MS data files. Reduction provides smaller, viewable and interpretable data sets by employing feature selection and feature extraction. Also know as dimension reduction and data reduction. This is part of the principal component analysis process employed by Agilent Mass Profiler Professional.
Regression analysis	Mathematical techniques for analyzing data to identify the relationship between dependent and independent variables present in the data. Information is gained from the estimation, regression, or the sign and proportionality of the effects of the independent variables on the dependent variables. This is part of the principal component analysis process employed by Agilent Mass Profiler Professional. Also known as regression.
Replicate	Collecting multiple identical samples from a population so that when the samples are evaluated a value is obtained that more closely approximates the true value.

Sample

A part, piece, or item that is taken from a specimen and understood as being representative of the larger specimen (e.g., blood sample, cell culture, body fluid, aliquot) or population. An analysis may be derived from samples taken at a particular geographical location, taken at a specific period of time during an experiment, or taken before or after a specific treatment. A small number of specimens used to represent a whole class or group.

Sample class prediction

A workflow used to build a model and classify samples from mass spectrometry data. Class prediction is a supervised learning method and involves three steps: validation, training, and prediction. The algorithm learns from samples (training set) with known functional class and builds a prediction model to classify new samples (test set) of unknown class.

Specimen

An individual organism, e.g., a person, animal, plant, or other organism, of a class or group that is used as a representative of a whole class or group.

Spike

The specific and quantitative addition of one or more compounds to a sample.

Standard

A chemical or mixture of chemicals selected for use as a basis of comparing the quality of analytical results or for use to measure and compensate the precise offset or drift incurred over a set of analyses.

Standard deviation

A measure of variability among a set of data that is equal to the square root of the arithmetic average of the squares of the deviations from the mean. A low standard deviation value indicates that the individual data tend to be very close to the mean, whereas a high standard deviation indicates that the data is spread out over a larger range of values from the mean.

State

A set of circumstances or attributes characterizing a biological organism at a given time. A few sample attributes may include temperature, time, pH, nutrition, geography, stress, disease, and controlled exposure.

Statistics

The mathematical process employed in manipulating numerical data from scientific experiments to derive meaningful information. This is part of the principal component analysis process employed by Agilent Mass Profiler Professional.

Subject

A chemical or biological sample taken from a specimen, or a whole specimen, that undergoes a treatment, experiment, or an analysis for the purposes of further understanding.

Survey

Collection of samples from less than the entire population in order to estimate the population attributes.

t-Test

A statistical test to determine whether the mean of the data differs significantly from that expected if the samples followed a normal distribution in the population.

The test may also be used to assess statistical significance between the means of two normally distributed data sets. See also ANOVA.

Unidentified compound

Chromatographic components that are only uniquely denoted by their mass and retention times and which have not been assigned an exact identity, such as compound name and molecular formula. Unidentified compounds are typically produced by feature finding and deconvolution algorithms. See also Identified Compound.

Variable

An element in a data set that assumes changing values, e.g. values that are not constant over the entire data set. The two types of variables are independent and dependent.

Volume

The area of the extracted compound chromatogram (ECC). The ECC is formed from the sum of the individual ion abundances within the compound spectrum at each retention time in the specified time window. The compound volume generated by MFE is used by Mass Profiler Professional to make quantitative comparisons.

Wizard

A sequence of dialog boxes presented by Mass Profiler Professional that guides you through well-defined steps to enter information, organize data, and perform analyses.

References

This section consists of citations to Agilent manuals, primers, application notes, presentations, product brochures, technical overviews, training videos, and software that help you use Agilent products and perform your metabolomics analyses.

Manuals

- Agilent Metabolomics Workflow - Discovery Workflow Overview
  (Agilent publication 5990-7069EN, Revision B, October 2012)
- Agilent Mass Profiler Professional - User Manual
  (Agilent publication, January 2012)
- Agilent G3835AA MassHunter Mass Profiler Professional - Quick Start Guide
  (Agilent publication, G3835-90009, Revision A, November 2012)
- Agilent G3835AA MassHunter Mass Profiler Professional - Familiarization Guide
  (Agilent publication, G3835-90010, Revision A, November 2012)
- Agilent G3835AA MassHunter Mass Profiler Professional - Application Guide
  (Agilent publication, G3835-90011, Revision A, November 2012)
- Agilent MassHunter Workstation Software Qualitative Analysis - Familiarization Guide
  (Agilent publication G3336-90018, Revision A, September 2011)
- Agilent MassHunter Workstation Software Quantitative Analysis - Familiarization Guide
  (Agilent publication G3335-90108, First Edition, June 2011)

Primers

- Proteomics: Biomarker Discovery and Validation
  (Agilent publication 5990-5357EN, February 11, 2010)
- Metabolomics: Approaches Using Mass Spectrometry
  (Agilent publication 5990-4314EN, October 27, 2009)

Application Notes

- Multi-omic Analysis with Agilent's GeneSpring 11.5 Analysis Platform
  (Agilent publication 5990-7505EN, March 25, 2011)
- An LC/MS Metabolomics Discovery Workflow for Malaria-Infected Red Blood Cells Using Mass Profiler Professional Software and LC-Triple Quadrupole MRM Confirmation
  (Agilent publication 5990-6790EN, November 19, 2010)
- Profiling Approach for Biomarker Discovery using an Agilent HPLC-Chip Coupled with an Accurate-Mass Q-TOF LC/MS
  (Agilent publication 5990-4404EN, October 20, 2009)
- Metabolite Identification in Blood Plasma Using GC/MS and the Agilent Fiehn GC/MS Metabolomics RTL Library
  (Agilent publication 5990-3638EN, April 1, 2009)
- Metabolomic Profiling of Bacterial Leaf Blight in Rice
  (Agilent publication 5989-6234EN, February 14, 2007)

Presentations

- Multi-omics Analysis Software for Targeted Identification of Key Biological Pathways
  (Agilent publication n/a, May 3, 2012)
- Metabolomics LCMS Approach to: Identifying Red Wines according to their variety and Investigating Malaria infected red blood cells
  (Agilent publication n/a, November 3, 2010)
- Small Molecule Metabolomics
  (Agilent publication n/a, November 3, 2010)
- Presentation: Metabolome Analysis from Sample Prep through Data Analysis
  (Agilent publication n/a, November 3, 2010)

Product Brochures

- Emerging Insights: Agilent Solutions for Metabolomics
  (Agilent publication 5990-6048EN, April 30, 2012)
- Agilent Mass Profiler Professional Software - Discover the Difference in your Data
  (Agilent publication 5990-4164EN, April 27, 2012)
- Confidently Better Bioinformatics Solutions
  (Agilent publication 5990-9905EN, February 2, 2012)
- Integrated Biology from Agilent: The Future is Emerging
  (Agilent publication 5990-6047EN, September 1, 2010)
- Agilent Fiehn GC/MS Metabolomics RTL Library
  (Agilent publication 5989-8310EN, December 5, 2008)
- Agilent METLIN Personal Metabolite Database
  (Agilent publication 5989-7712EN, December 31, 2007)
- Agilent Metabolomics Laboratory: The breadth of tools you need for successful metabolomics research
  (Agilent publication 5989-5472EN, January 31, 2007)

www.agilent.com

Revision B, October 2012

5990-7067EN

Agilent Technologies