
# Progenesis LC-MS User Guide

## Analysis workflow guidelines

## for version 4.0

# Contents

## Introduction

This user guide takes you through a complete analysis of 6 LC-MS runs with 2 groups (3 replicate runs per group) using the unique Progenesis LC-MS workflow. It starts with LC-MS data file loading then Alignment, followed by Analysis that creates a list of interesting features (peptides) which are explored within Progenesis Stats using multivariate statistical methods then onto Protein identity and Reporting.

To allow ease of use the tutorial is designed to start with the restoration of an Archived experiment where the data files have already been loaded. However, the document covers all the stages in the LC-MS workflow, therefore if you are using your own data files please refer to Appendix 1 (page 55) then start at page 6.

## How to use this document

You can print this user guide to help you work hands-on with the software. The complete user guide takes about 50 minutes and is divided into two sections. This means you can perform the first half focused on LC-MS run alignment and analysis then complete the second half of analysis exploring comparative differences and Protein identity at a convenient time. If you experience any problems or require assistance, please contact us at support@nonlinear.com

## How can I analyse my own runs using LC-MS?

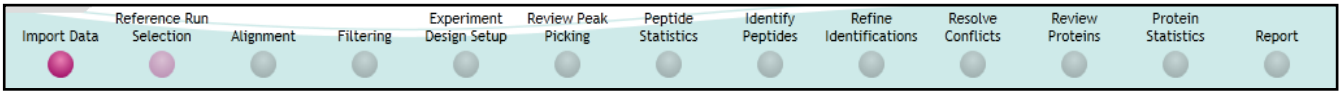
You can freely explore the quality of your LC-MS data using Data Import and then licence your own LC-MS runs using this evaluation copy of Progenesis LC-MS. Instructions on how to do this are included in a section at the end of the user guide document. Alternatively if you would like to arrange a demonstration in your own laboratory contact support@nonlinear.com and we will help you.

## LC-MS Data used in this user guide

NLD would like to thank Dr Robert Parker and Prof Haroun Shah at the Health Protection Agency, London, UK for providing the example data used in this user guide as well as invaluable discussion on the handling of the data.

# Workflow approach to LC-MS run analysis

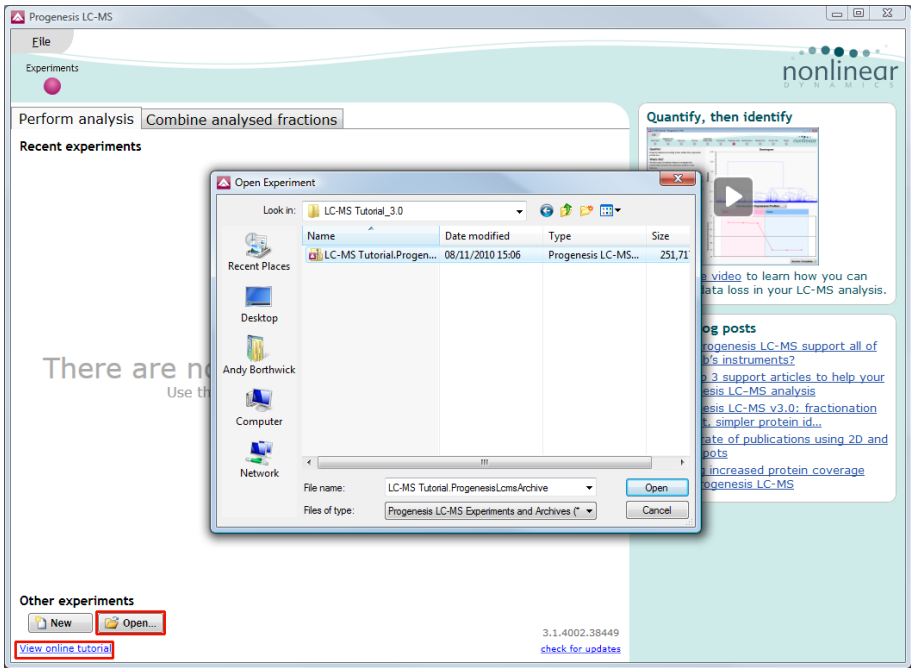
Progenesis LC-MS adopts an intuitive **Workflow** approach to performing comparative LC-MS data analysis. The following user guide describes the various stages of this workflow (see below) focusing mainly on the stages from Alignment to Report.



# Restoring the LC-MS Tutorial

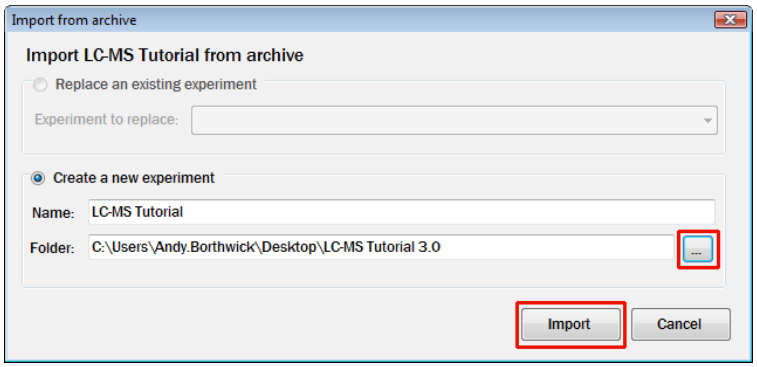
Open Progenesis LC-MS and download the Compressed (.zip) Tutorial Archive file from the the 'View online tutorial' link shown below, placing it in a **new folder** on your desktop. Before restoring the tutorial in the software **you must** first right click on the (.zip) file and extract it to the same folder.

Now you can restore the uncompressed LC-MS tutorial archive file. To do this, first locate the LC-MS Tutorial Archive file using the **Open** button and press Open.

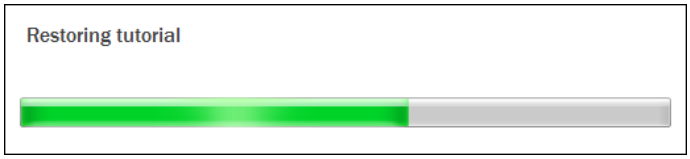



This opens the 'Import from archive' dialog.

Select the **Create a new experiment** option and select the folder in which you placed the archive, using the icon (to the right) .




Then press **Import.**

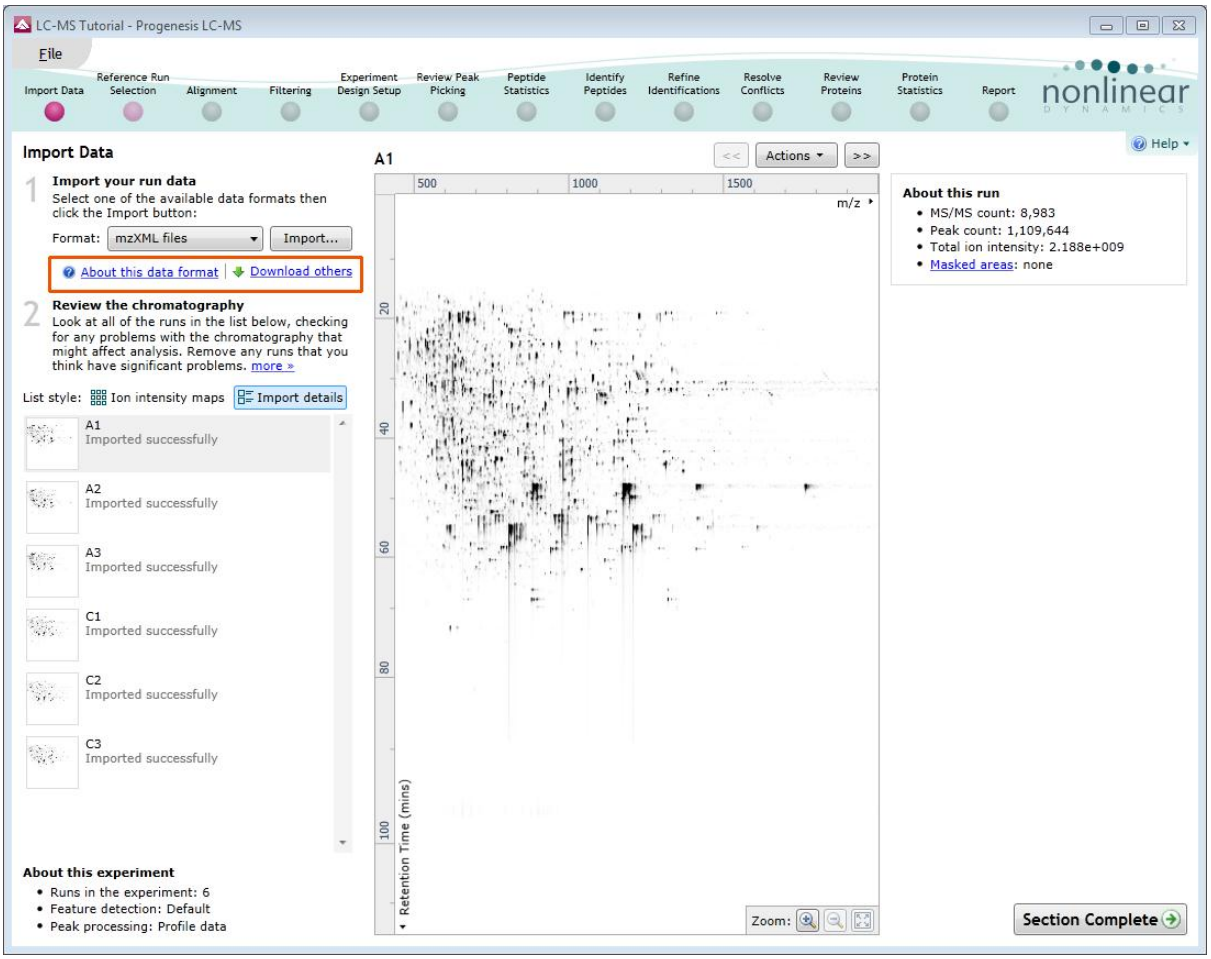



**Note**: use the **Replace an existing experiment** option if you want to over-write an existing version of the tutorial.

***Tip***: at each stage in the software there are links to more information and help on the website.

# Stage 1: Import Data and QC review of LC-MS data set

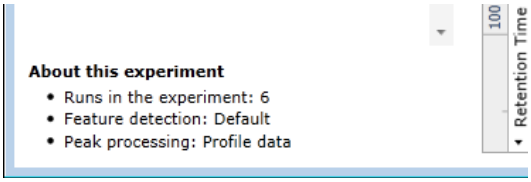
The LC-MS tutorial will now open at the LC-MS **Import Data** stage (see below).



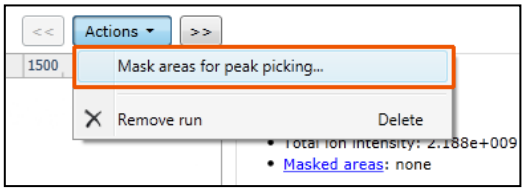

Each data file appears as a 2D representation of the run. At this stage you will be warned if any of the data files have been 'centroided' during the data acquisition and conversion process.

**Note**: as each data file is loaded the progress is reported in the **Import Data** list. The dialog below the Run reports on the QC of the imported Data files. In this case 'No problems found' with this data file.

**Note**: the **'Data Processing Methods',** selected when the experiment was created, are reported on the bottom left of the application (see Appendix 1, page 55).




***Tip***: the **'Exclude areas from selected run'** facility allows you to examine and exclude areas (usually early and/or late in the LC dimension (Retention Time)) that appear excessively noisy due to capture of data during column regeneration (see Appendix 2, page 57). This is not required for this data set.

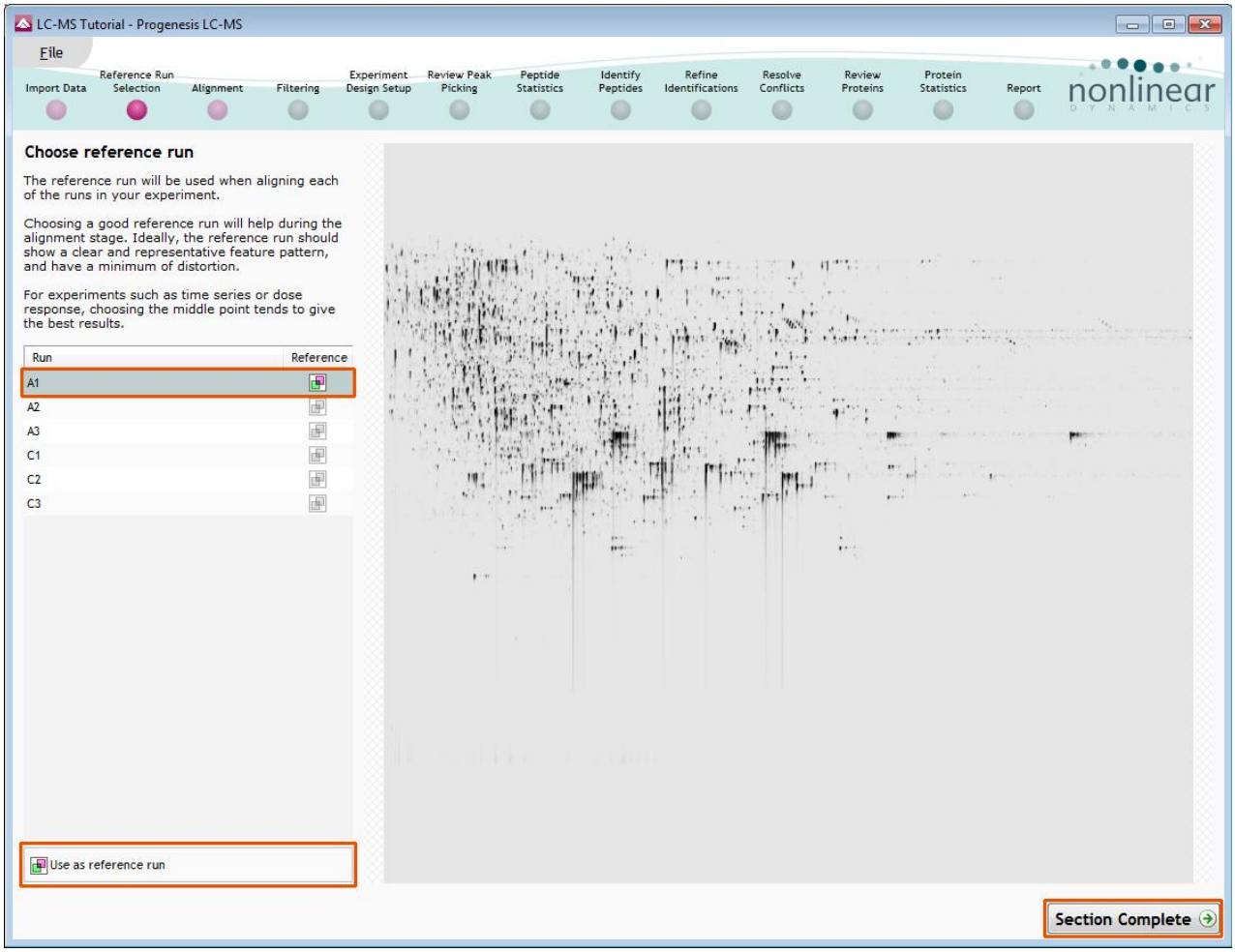



**Note**: use the **Remove Run** to remove run(s) from the current experiment.

Once all the files have been imported move to the next stage in the workflow by clicking **Section Complete.**

## Stage 2: Reference Run selection

This stage in the analysis workflow allows you to review and select the most appropriate Reference LC-MS run to align all the other runs to.



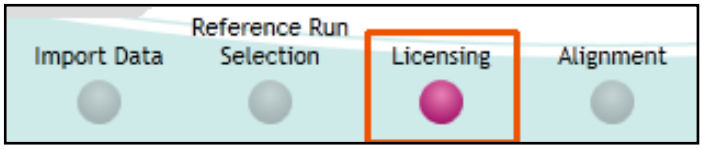

To select a Reference run either click on the run in the list and then click **Use as reference run** or double click on the run in the list.

***Tip***: choice of a Reference run is usually dependant on the visual quality of the Chromatography avoiding runs that show visual gaps in the chromatography.

Now move to the next stage in the workflow by clicking **Section Complete.**

## Stage 3: Licensing

This stage in the analysis workflow will **only** appear if you are using 'Unlicensed' data files to evaluate the software and have no dongle attached.




For details on how to use Licensing go to Appendix 3 (page 58)

**If you are using the tutorial archive, this page will not appear as the data files are licensed.**

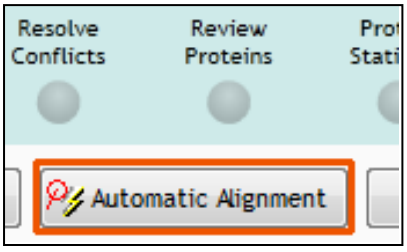
# Stage 4: Alignment

At this stage Progenesis LC-MS Alignment opens displaying your data.


### Generation of alignment vectors

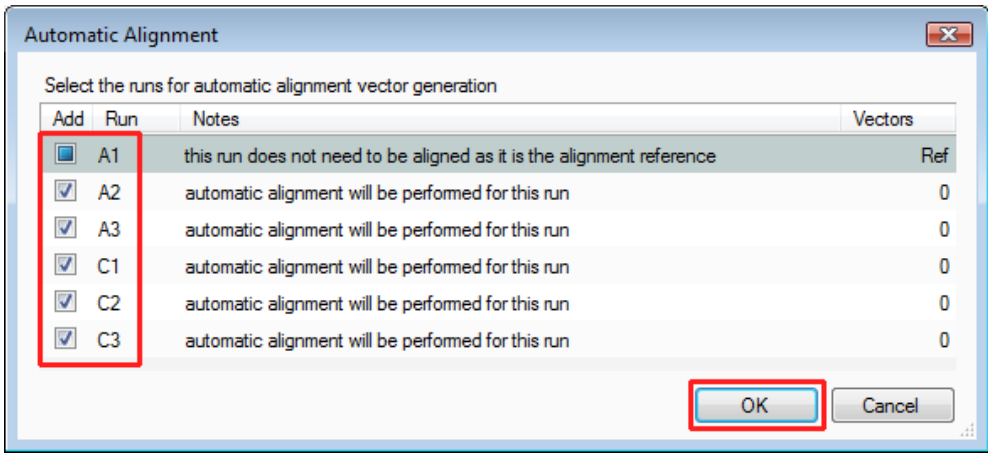
The alignment of LC-MS runs is required in the LC (retention time) direction, this is key to correcting for the variable elution of peptides during the chromatographic separation.

The Alignment algorithm will generate ‘Automatic’ vectors, in the retention time direction for each run, to enable the alignment of all the LC-MS runs to the ‘Reference Run’.




The alignment vectors are generated automatically for all the LC-MS runs by using the ‘Automatic vector wizard’ accessed by clicking on **Automatic Alignment** on the top tool bar.

Select (tick) the runs you require to generate vectors for and click **OK**.




The following pages in this user guide explain in more detail the views and functions of the Alignment stage in the Progenesis LC-MS Alignment, focusing on the Program layout.

These pages act as a useful guide and reference to the Alignment Stage that you can return to after having generated the Alignment vectors automatically.

## Taking a detailed approach to alignment

In some cases, where the misalignment is severe, using a combination of a ‘few’ manually placed vectors on each run and then using the Automatic vector wizard to generate the rest of the vectors for each run can give better results.

For more details on manual assistance of Alignment refer to Appendix 4 page 59

## Layout of Alignment

To familiarize you with Progenesis LC-MS Alignment, this section describes the various graphical features used in the alignment of the LC-MS runs

To setup the display so that it looks similar to the one below:

- Click on the features shown in the current focus (orange rectangle) in Window C, this will update windows A,B and D as shown below.
- In window A **click and hold** the left mouse button on a green feature.
- If the green and magenta features (immediately above) have not aligned automatically then **drag** the green feature over the magenta feature and **release** the mouse button.
- The view will 'bounce' back and a red vector, starting in the green feature and finishing in the circled magenta feature will now appear as shown below in window A.

The experiment structure is displayed on the left of the screen in the **Run** panel.


The **Runs:** panel shows the run that is currently being aligned in green, and the run it is being aligned to in magenta. This is the reference run you chose at the previous stage, in this case **A1.**

| Run | Include? | Vectors |
|---|---|---|
| A1 | ✓ | Ref |
| A2 | | 1 |
| A3 | | 0 |
| C1 | | 0 |
| C2 | | 0 |
| C3 | | 0 |

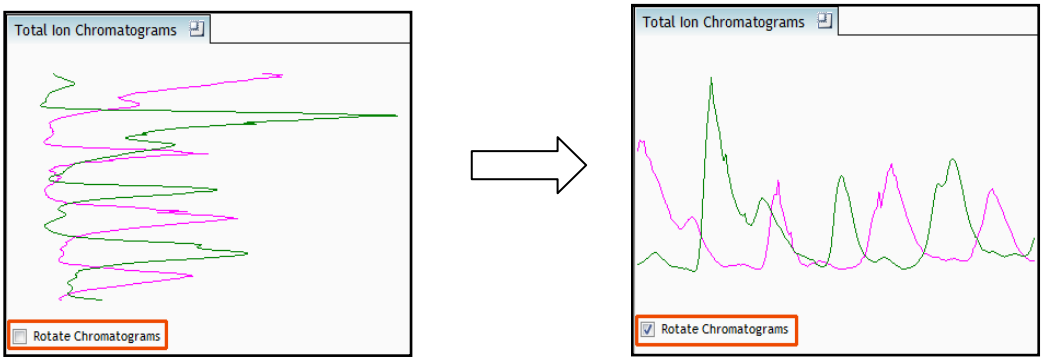
**Vector Editing (Window A)**: is the main alignment area and displays the area defined by the current **focus** rectangle shown in Window C. The current run is displayed in green and the chosen reference run is displayed in magenta. Here is where you place the alignment vectors.

**Transition (Window B)**: uses an **alpha blend** to animate between the current and reference runs. Before the runs are aligned, the features appear to move up and down. Once correctly aligned, they will appear to pulse. During the process of adding vectors, this view will change to show a zoomed view of the area being aligned to help accurate placement.

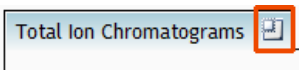
**Whole Run (Window C)**: shows the **focus** for the other windows. When you click on the view the orange rectangle will move to the selected area. The focus can be moved systematically across the view using the cursor keys. The focus area size can be altered using the controls in the bottom left of the screen or by clicking and dragging out a new area with the mouse.

**Total Ion Chromatograms (Window D)**: shows the current **total ion** chromatogram (green) overlaid on the Reference chromatogram (magenta). As the features are aligned in the **Vector Editing** view the chromatograms become aligned. The retention time range displayed is the vertical dimension of the Focus Grid currently displayed in the **Whole Run** view (Window C).

**Note**: the orientation of the TIC view can be changed according to individual preference




This view assists in the verification of the feature alignment.

**Note:** the icon to the right of the 'Window' titles expands  the view

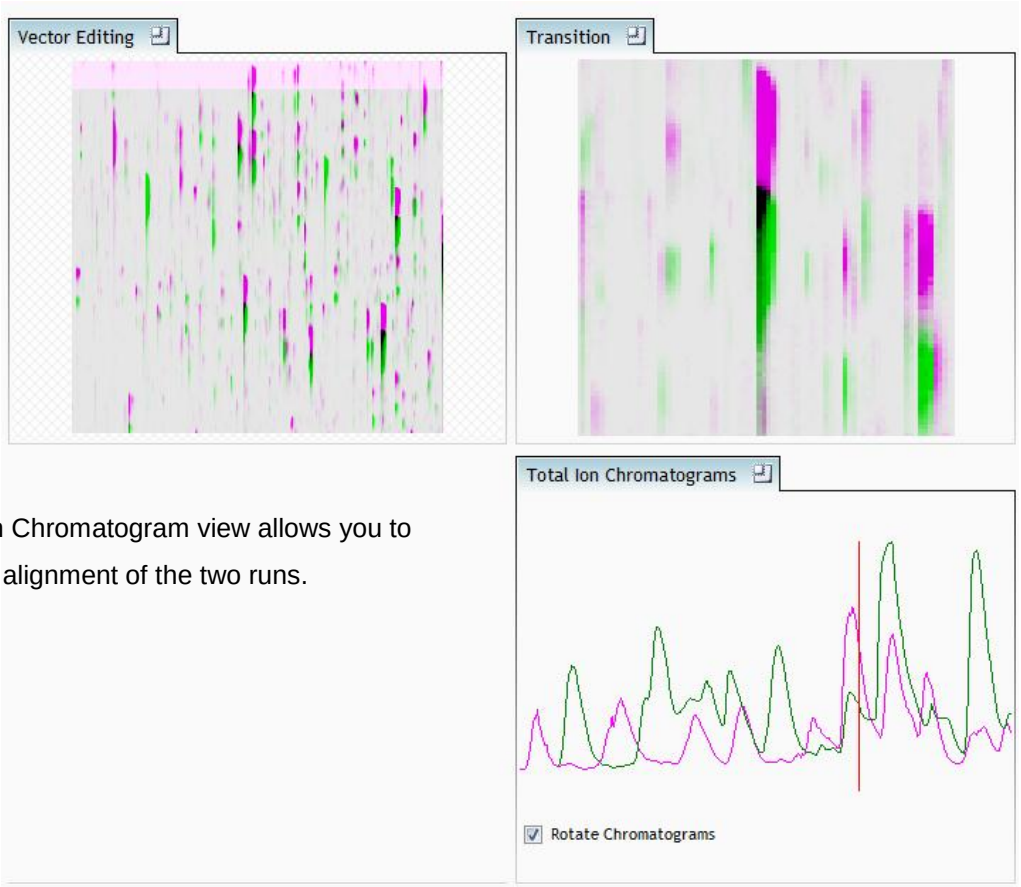
## Reviewing generation of alignment vectors

After applying **Automatic alignment** the number of vectors will be updated on the **Runs** panel and the vectors will appear (in blue) on the view.

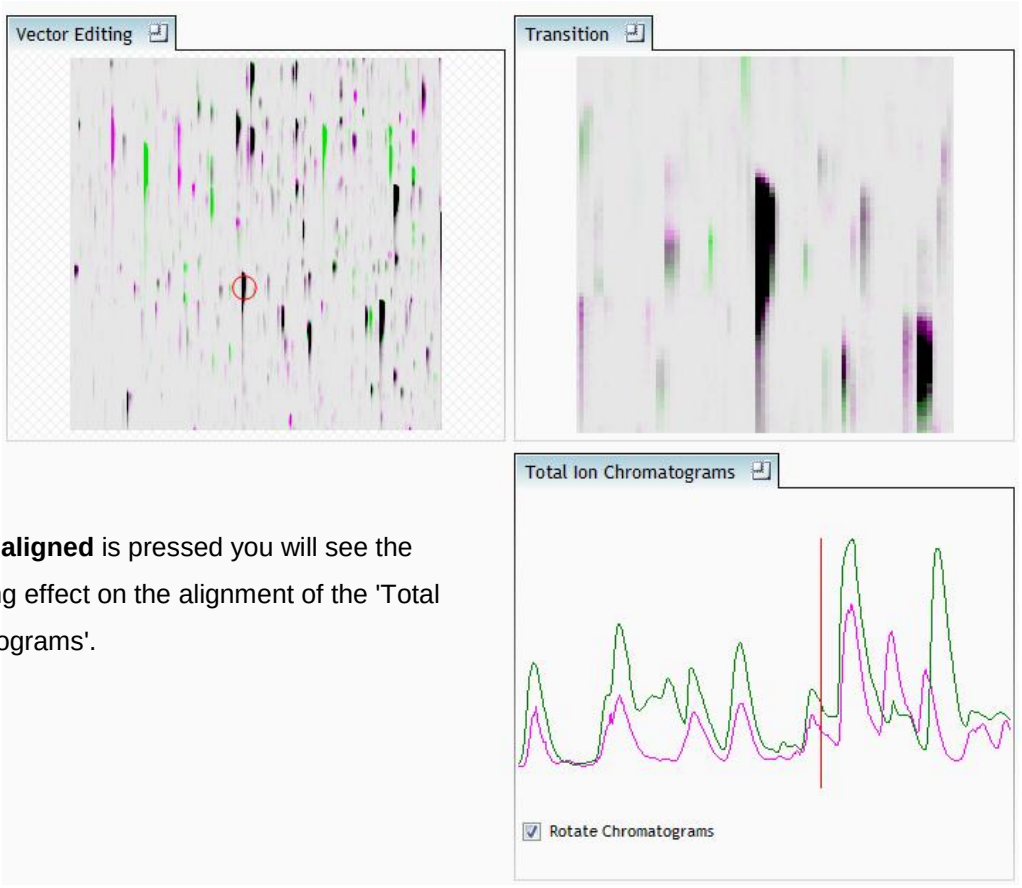
If the alignment has worked well then in Windows A and C the grid lines (option under **View** menu) should show minimal distortion, Window B (Transition) will show features pulsing slightly but not moving up and down.


At this point, you should check the automatically placed (blue) vectors. This will be easier with a larger grid size. Make sure the grid size is set to 4 using the **'Focus grid size'** control at the bottom left of the window.

In each square, you can, if required edit the vectors to improve the run alignment (for more information refer to Appendix 4 (page 59).




The Total Ion Chromatogram view allows you to further verify alignment of the two runs.



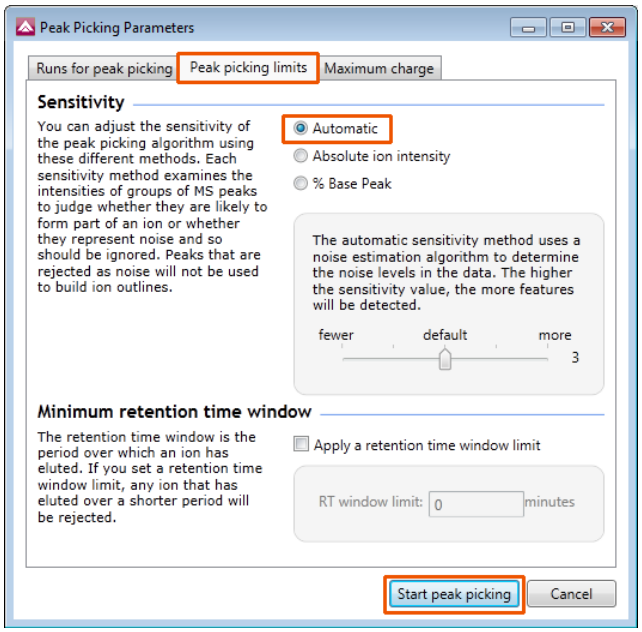

When **show aligned** is pressed you will see the corresponding effect on the alignment of the 'Total Ion Chromatograms'.

# Stage 5A: Filtering

Now that you have reviewed your aligned Runs, you are ready to analyse them. Move to the **Filtering** stage, by either clicking on **Section Complete** (bottom right) or on Filtering on the workflow.


### Peak Picking Parameters

The Peak Picking Parameters dialog opens, showing all the runs in the current experiment and a tick against each run. This is the default setting, where the peak picking algorithm uses information from all of the runs to contribute to the pattern of feature outlines.


***Tip***: It may be appropriate **only** to pick peaks that are present in a limited number of your runs. In which case un-tick the runs that you do **NOT** want to contribute to the feature detection pattern. This may be important when one or more of the runs appear noisy due to non-optimal chromatography or sample handling.

**Note**: features outlines will be added to 'un-ticked' runs; although these runs will not contribute to the peak picking pattern.

***Tip***: depending on run quality, a suggested minimum number of ticked runs should include at least one replicate of each experimental condition.

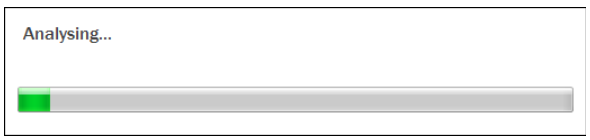
The sensitivity of the detection can be controlled by adjusting settings under the **Peak picking limits** tab.

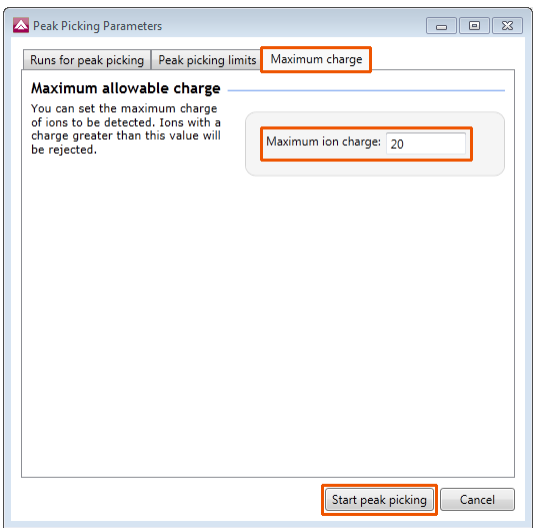
For the runs in this user guide, we will use the default settings for the **Automatic** method.

The third tab allows you to set the **maximum charge** of the ions, which will be detected. The default setting is a charge state of 20.

Press **Start Peak Picking** to start the detection process.

During the few minutes that the automatic analysis requires, a progress bar will appear telling you that it is Analysing.



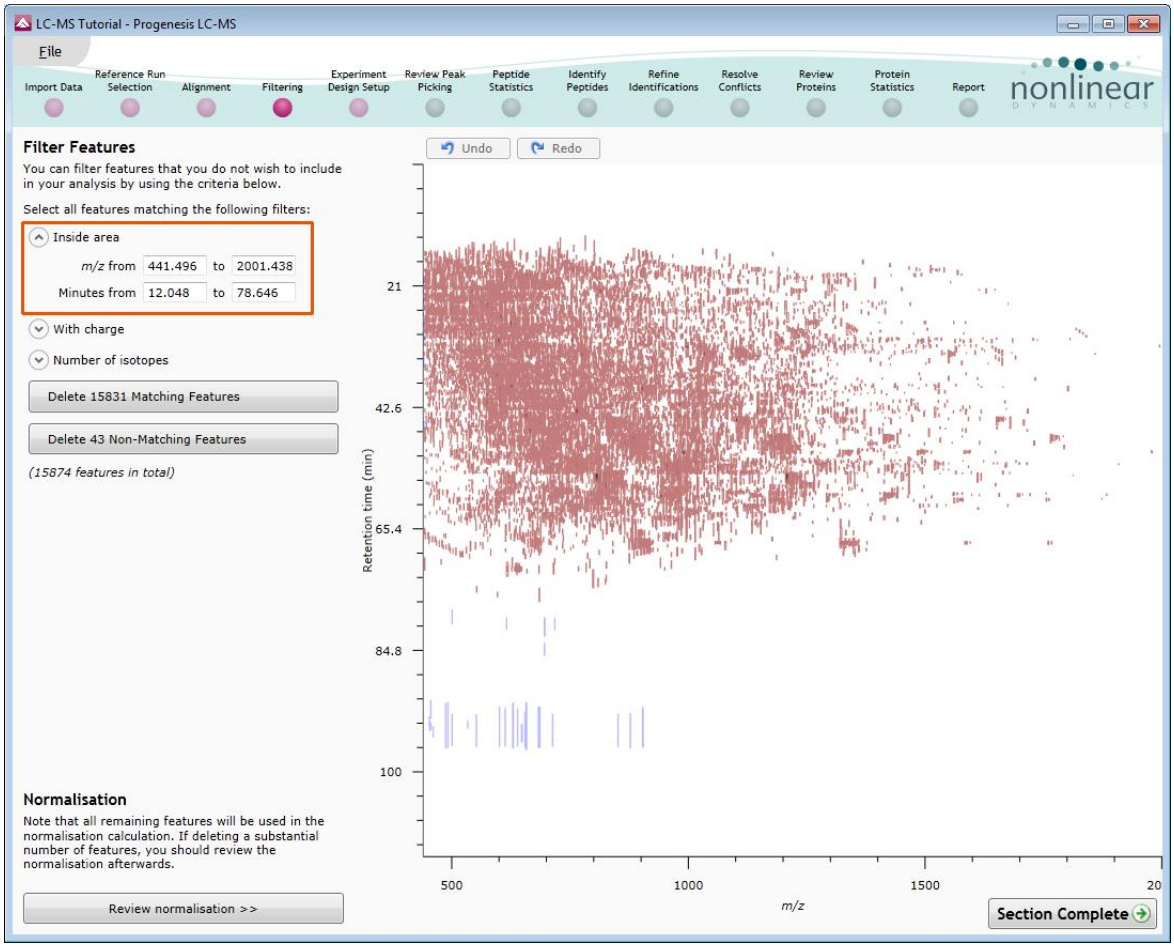
On completion of analysis, the Filtering stage will open displaying the number of features detected, in this example 15884.

If required you can remove features based on position, charge state, number of isotopes or combinations of these feature properties.

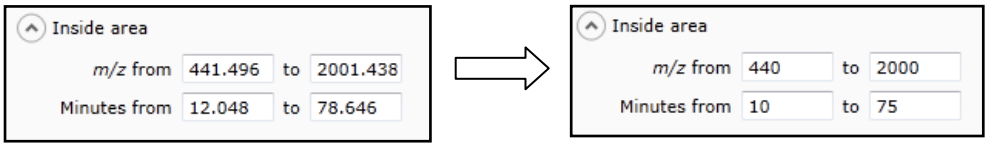

For example, to delete features with early and late ‘Retention times’ drag out an area as shown.

All features contained within the masked area are selected.

As you release the mouse button, the ranges for the masked area will appear on the top left.




***Tip***: the limits can be adjusted by entering the required values in the boxes.



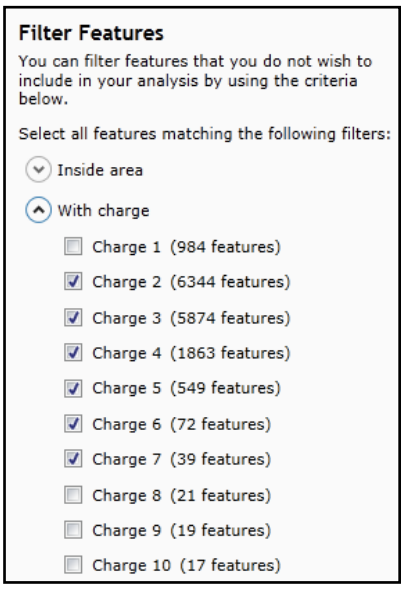

To remove the features outside of the selected area, (in this case 33), press the **Delete 33 Non-Matching Features** button.

In addition to setting limits for 'Retention time and m/z', features can also be selected based on charge or the number of isotopes present. This allows you to refine the selection through a combination of feature properties.

For example, when **With charge** is selected the number of features present at each charge state is displayed, these can be selected accordingly.

Area limits, charge state and number of isotopes can be combined to refine the feature selection.

***Tip***: when filtering on one property of the feature i.e. charge state, make sure you have 'collapsed' the other filters (see right)



**For this user guide**, we will filter the area as shown above and also delete a further 1110 features with a charge state of 1 or 8 and above by ticking the various options.

Hence all features with a charge state of 1 or 8 and above will appear blue on the main view.

To remove these features press **Delete 1110 Non Matching Features.**

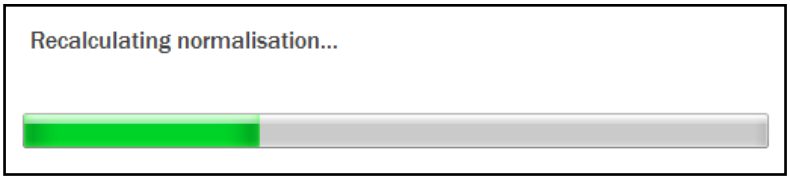
You can use the **Undo** button to bring back deleted features, however, when you move to the next section you will lose the capacity to undo the filter. Before moving on from filtering you should review the normalisation of the experiment.

***Tip***: When you have reached the filtering stage, it is good practice to close the experiment and save an archive. This can be used to restore the unfiltered state if the filtering you have performed is too 'stringent'.

## Stage 5B: Reviewing Normalisation

Normalisation review is accessed from the button at the bottom right corner of the filtering page


If you have filtered out a number of features from the original detection pattern then the normalisation will update.

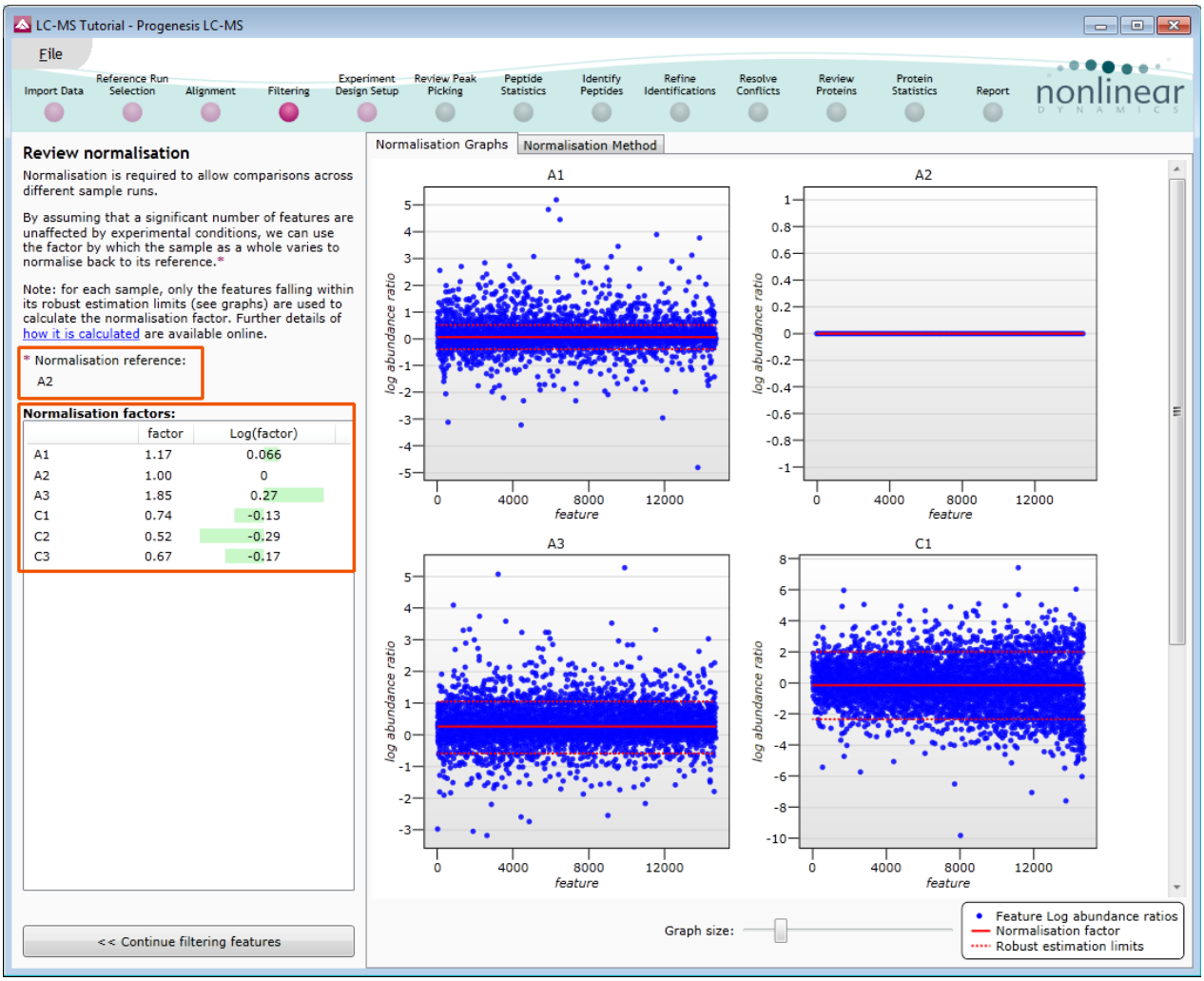



The **Review Normalisation** page will open displaying plots for the normalisation of all the features on each run.

This page in the workflow **does not** allow you to alter the Normalisation of your data but provides you with individual views for each run showing the data points used in the calculation of the normalisation factor for the run.

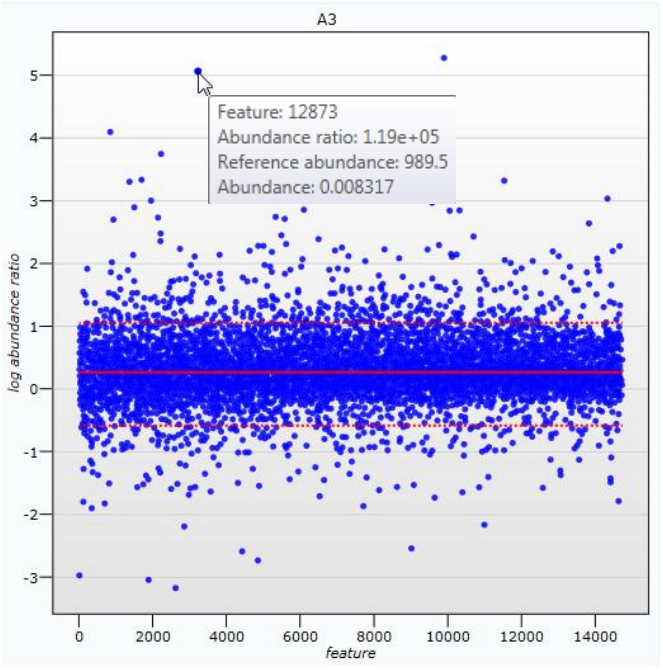
Alternatively, if you do not believe normalisation is necessary, you can opt to use un-normalised feature abundances for the rest of the analysis.

Normalisation factors are reported in the table to the left of the plots.




**Calculation of Normalisation Factor:**

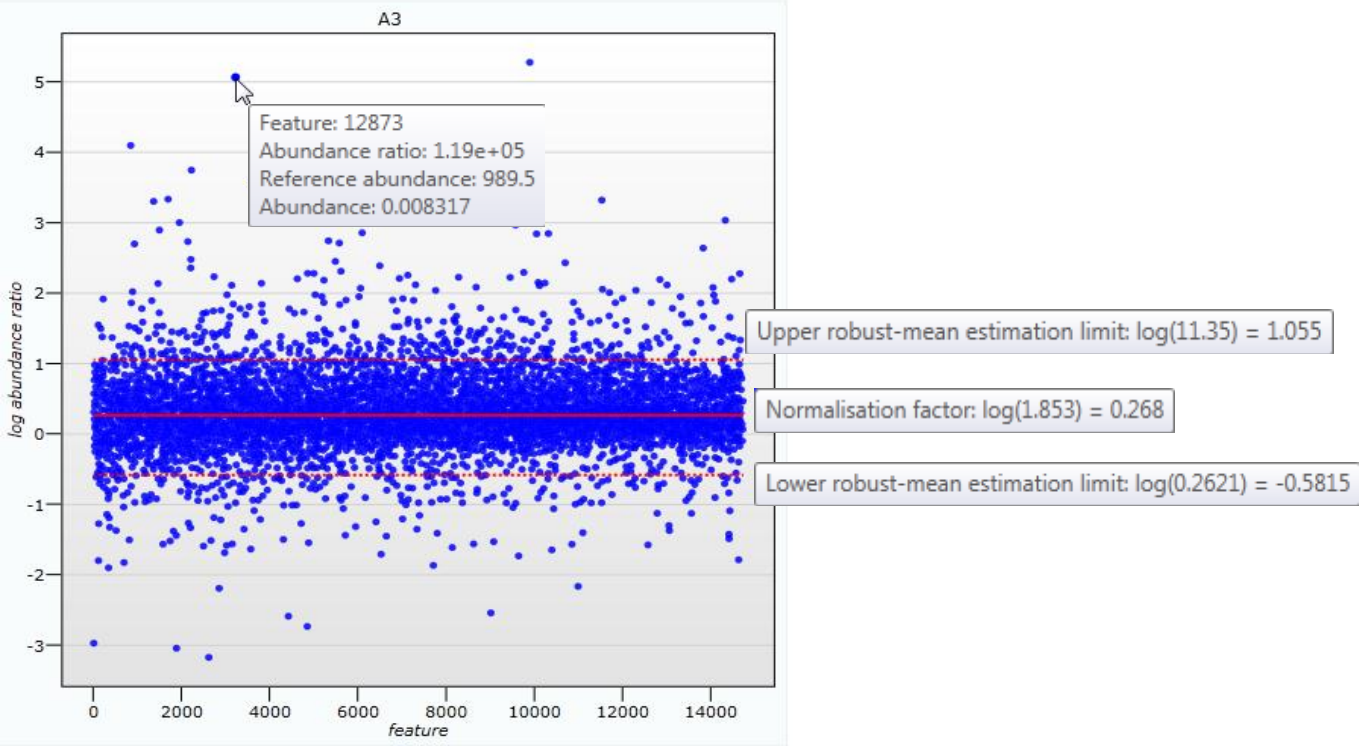
Progenesis LC-MS will automatically select one of the runs that is 'least different' from all the other runs in the data set to be the 'Normalising reference'. The run used, is shown above the table of Normalisation factors.



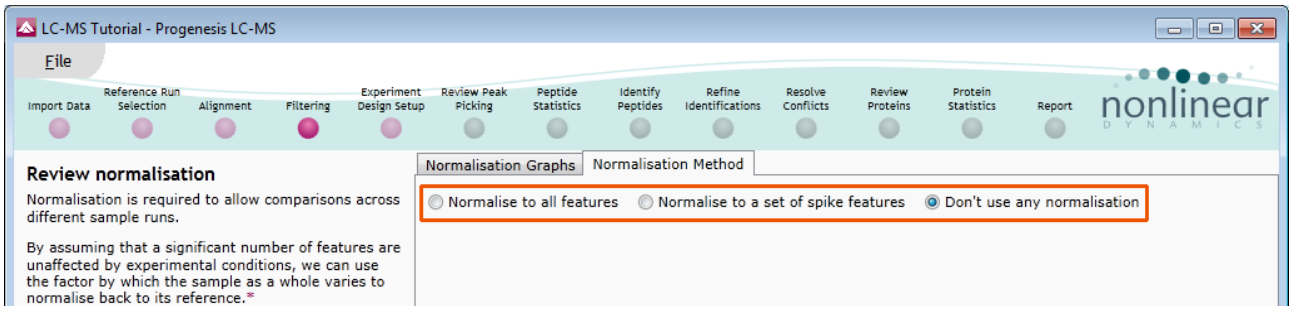

For each sample run, each blue dot shows the log of the abundance ratio for a different feature (normalisation target abundance/run abundance).

The details for individual features can be viewed as you hold the cursor over the dots on the plot.

On the graph the features are shown ordered by ascending mean abundance. The normalisation factor is then calculated by finding the mean of the log abundance ratios of the features that fall within the 'robust estimated limits' (dotted red lines). Features outside these limits are considered to be outliers and therefore will not affect the normalisation.



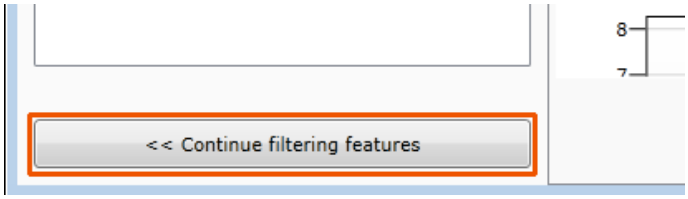

Finally, if you do **not** wish to work with normalised data then you can **use the raw abundances** by switching off the normalisation.




**Note**: once you have identified the spike features, you can then apply the **Normalise to a set of spike features** by using this option to locate and select the features.

For this experiment, you should leave the **Normalise to all features** option selected.

Now return to filtering by clicking on the button on the bottom left of the screen




For this example, we **DO NOT** do any additional Filtering so click on **Section complete**.

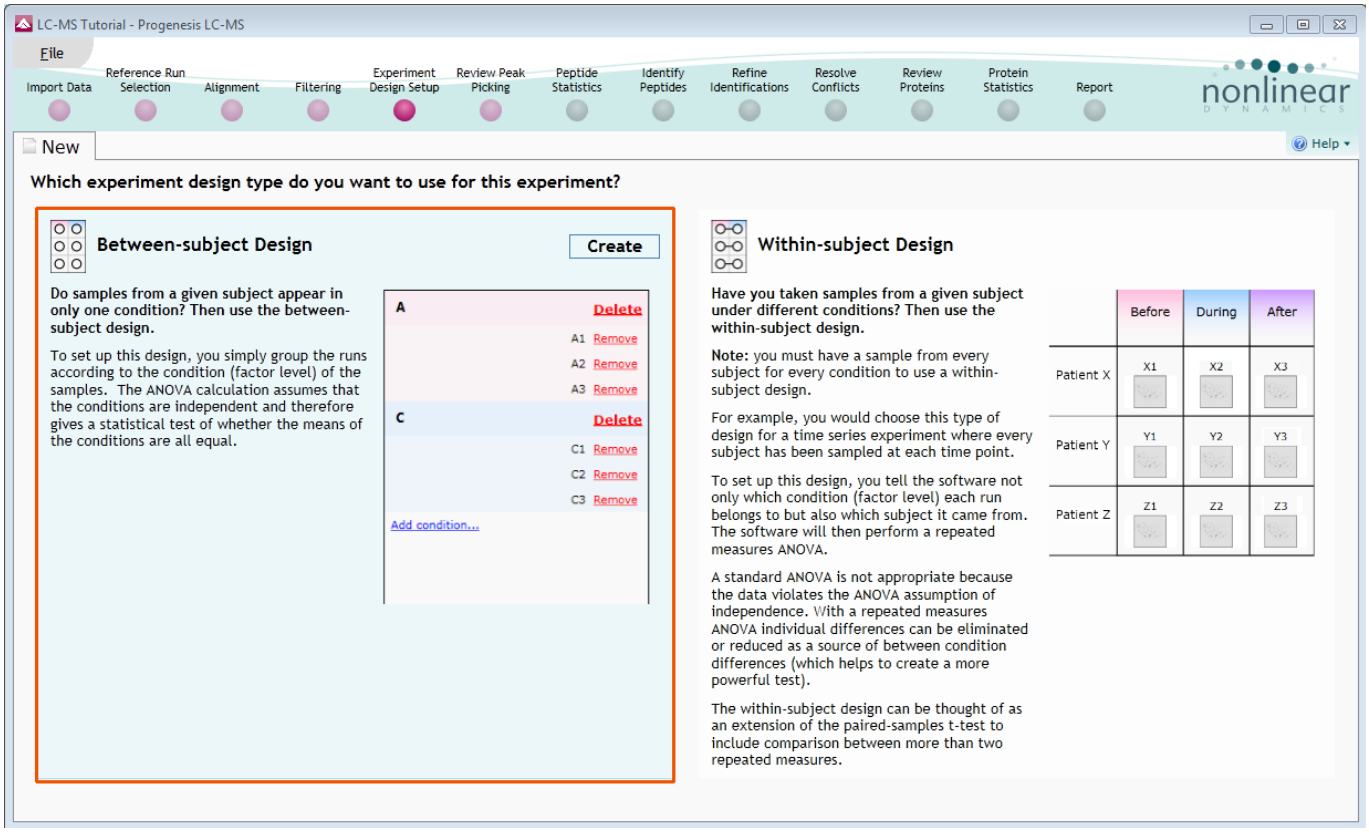
**Note**: if you do any extra filtering then **Normalisation recalculates** as you move to the next stage in the Workflow**.**

# Stage 6: Experiment Design Setup for Analysed Runs

At this stage in the workflow you can setup one or more experimental designs for your data.

There are two basic types of experimental designs:

**Between-subject design**: here samples from any given subject appear in only one condition. (i.e. control versus various drug treatments).The ANOVA calculation assumes that the conditions are independent and applies the statistical test that assumes the means of the conditions are equal.

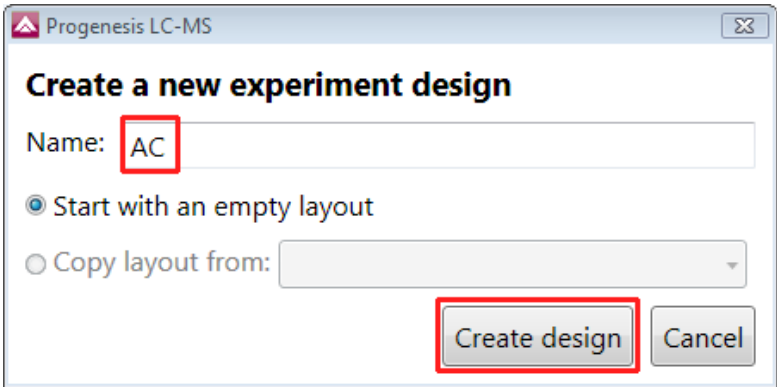



**Within-subject design**: here samples have been taken from a given subject under different conditions (i.e. the same subject has been sampled over a period of time or after one or more treatments). Here a standard ANOVA is not appropriate as the data violates the ANOVA assumption of independance. Therefore by using a *repeated measures* ANOVA, individual differences can be eliminated or reduced as a source of between condition differences. This within-subject design can be thought of as a extension of the paired samples t-test, including comparison between more than two repeated measures.

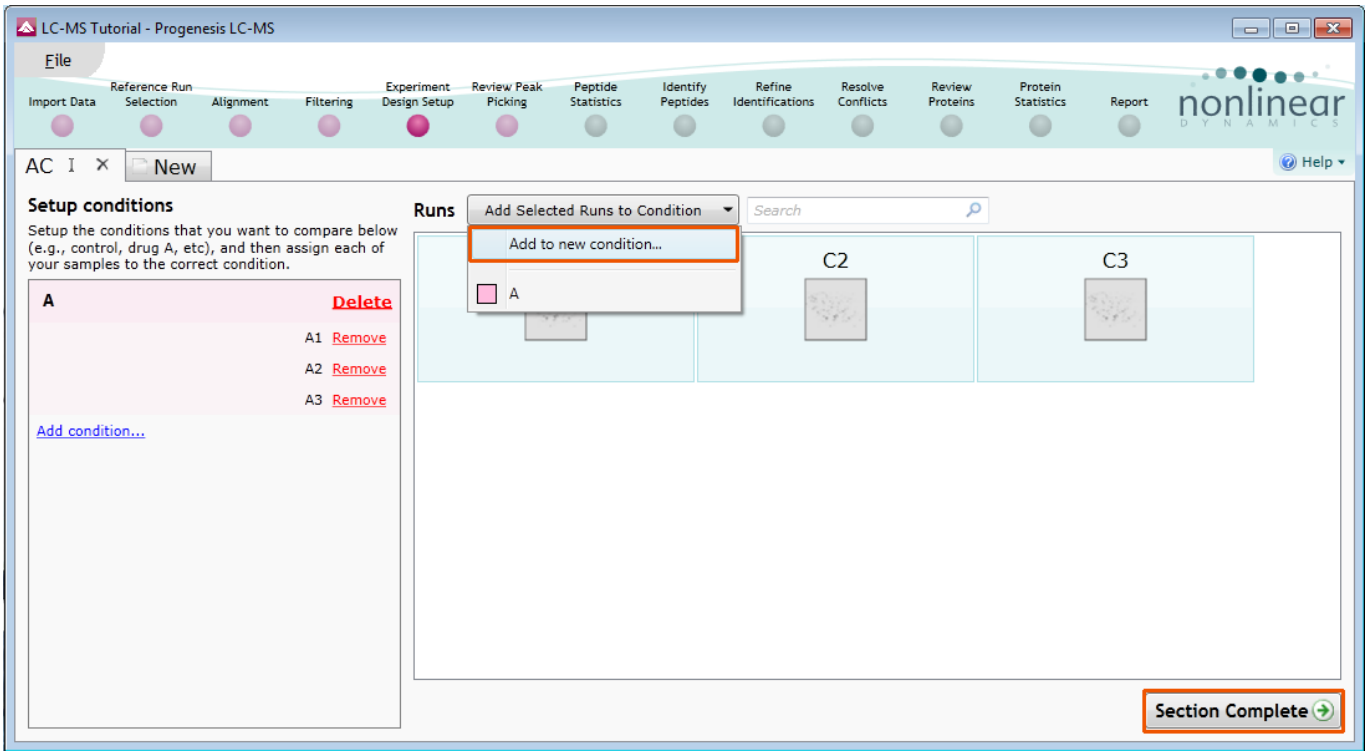
Additional information on how to apply the Within-subject Design is in **Appendix 5** page 64

This experiment contains 2 conditions: A and C and uses the **Between-subject design** to group the analysed runs to reflect the Biological conditions in the original study.

To create a new **Between-subject Design** hold the cursor over this option and click to open the dialog.



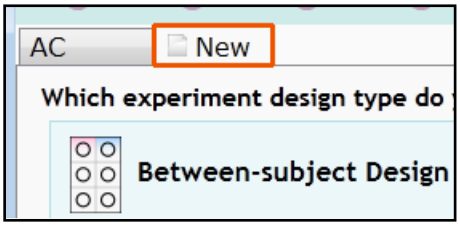

Give the new experimental design a name and then click **Create design**.



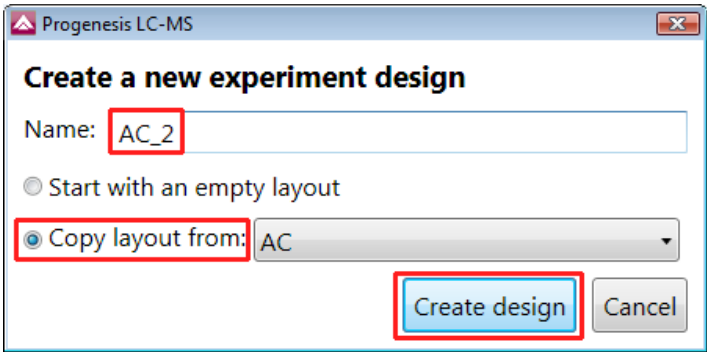

## To create a new condition

1. Select the runs for the condition by clicking on the required icon in the **Runs** panel, as shown.
2. Press the ‘black triangle’ next to the **Add Selected Runs to Condition** button on the main toolbar.
3. Select **Add to new condition**… from the drop down menu.
4. A new condition will appear in the **Conditions** panel on the left.
5. Rename the condition (e.g. C) by over typing the default name
6. Repeat steps 1 to 5 until all the runs are grouped into conditions.

To create another Design, for example comparing only two replicates for each condition, A and C, click on the **New** tab and click on Between-subject Design again.

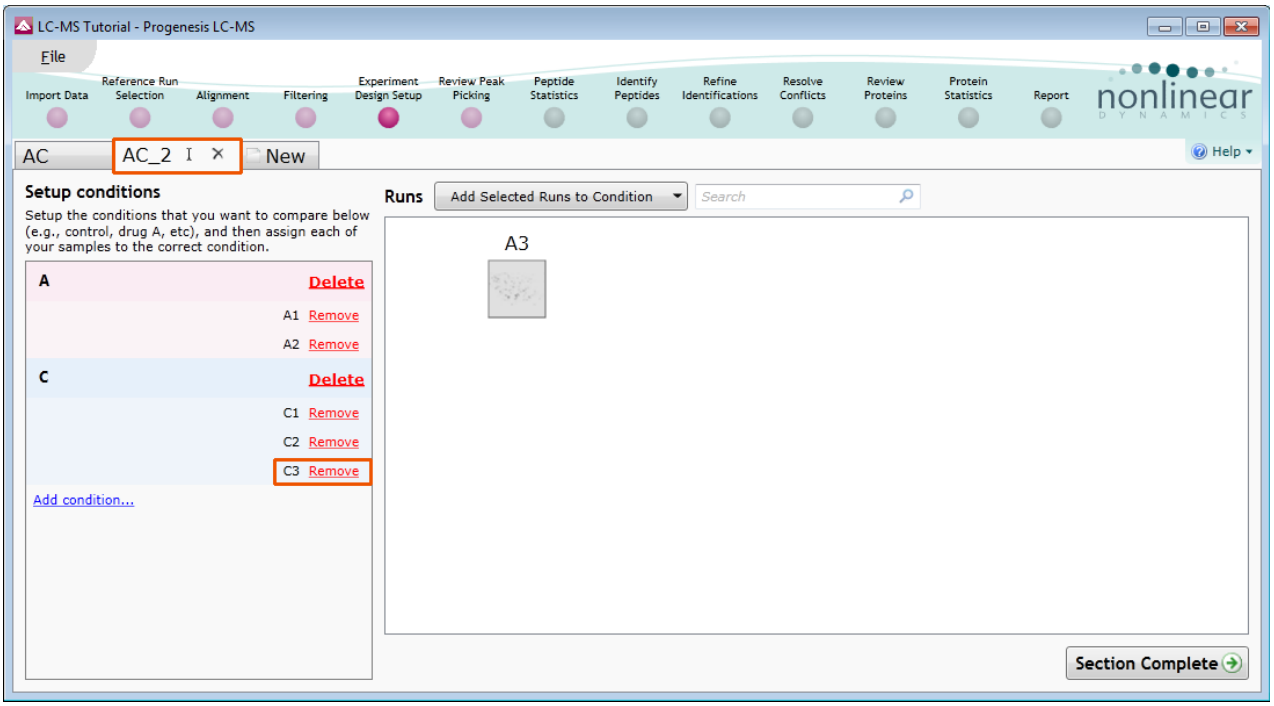



Give the new design a name, then tick the **Copy layout from** option and select the **AC** design.




When **Create design** is pressed the new tab refreshes to allow you to adjust the conditions.

Use the Delete link on the Conditions panel to remove replicates and/or conditions that are not required in this particular design.




On deleting each replicate the runs will reappear in the Runs window.

**Note**: both designs are available as separate tabs.

To move to the next stage in the workflow click **Section Complete**.

# Stage 7: Validation, review and editing of results

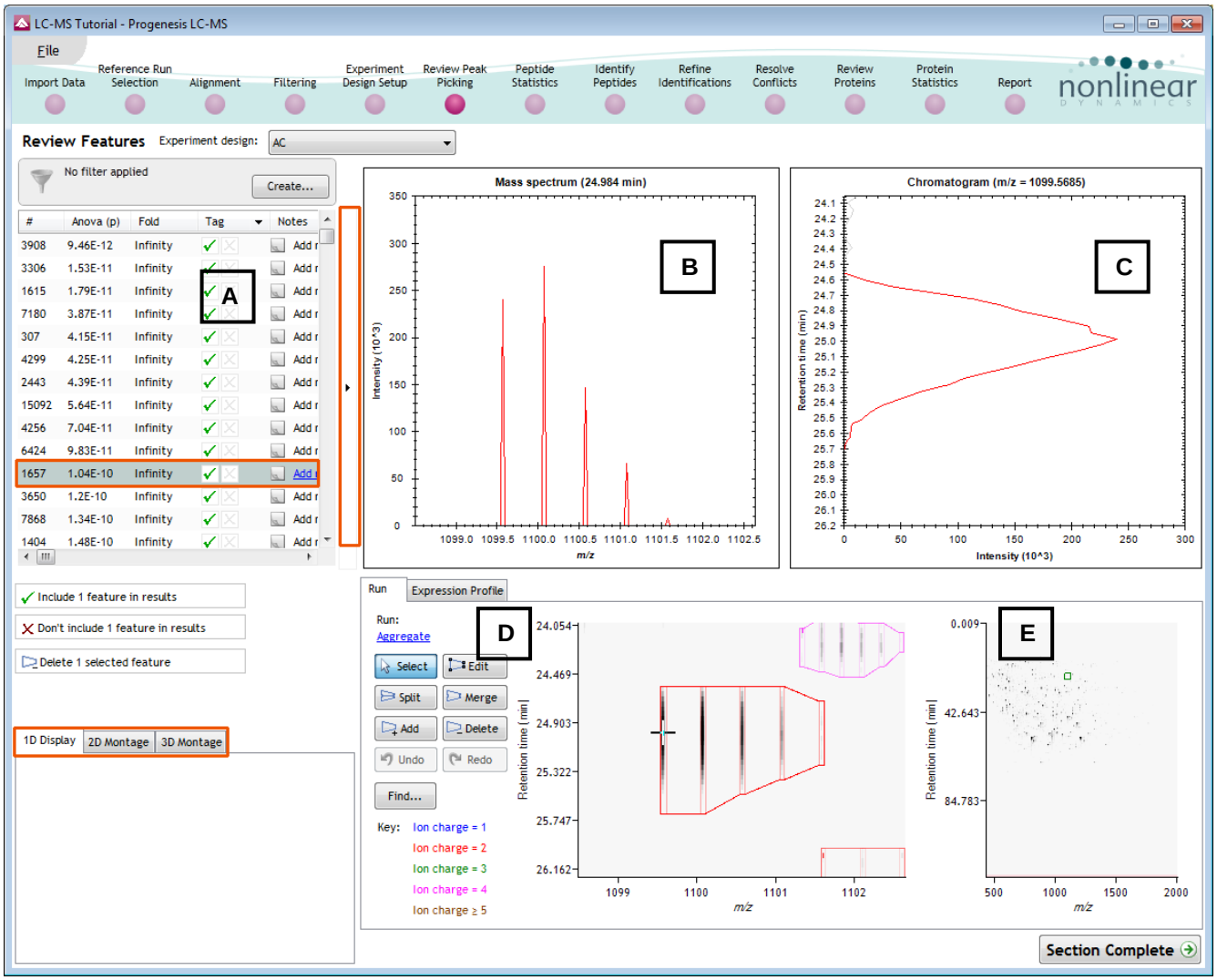
The purpose of this stage in the Workflow is to review the list of features using the visual tools provided and edit features if required.

The review stage has 4 display modes: 1D, 2D, 3D and Feature Details controlled by the tabs on the bottom left of the display and the expander bar to the right of the table. Each display has multiple views to allow comparative exploration of the detected features on the aligned LC-MS runs.

## Exploring analysed data using the Data displays

**Window A**: shows the list of features ranked by the p value for the one way **Anova** using the current grouping.

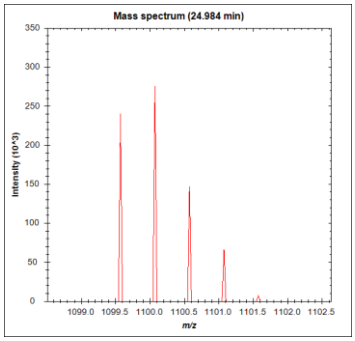
**Note**: a value of 'Infinity' in the **Fold** column indicates 'Presence/Absence'



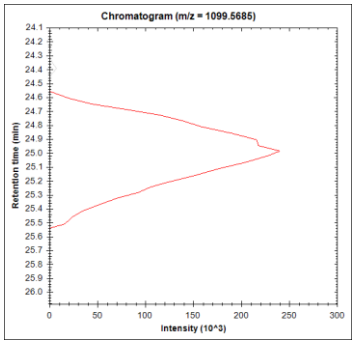

**Note**: by default all the features are included in the selection for the next section of the analysis.

To highlight a group of features drag out a selection on the table.

## The 1D Display

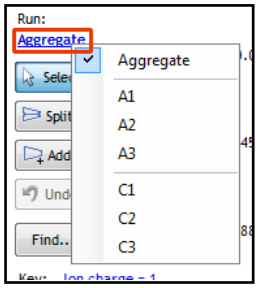



**Window B**: displays the Mass spectrum for the current feature on the selected Run (in window D).




**Window C**: displays the Chromatogram for the current feature on the selected Run (in window D).

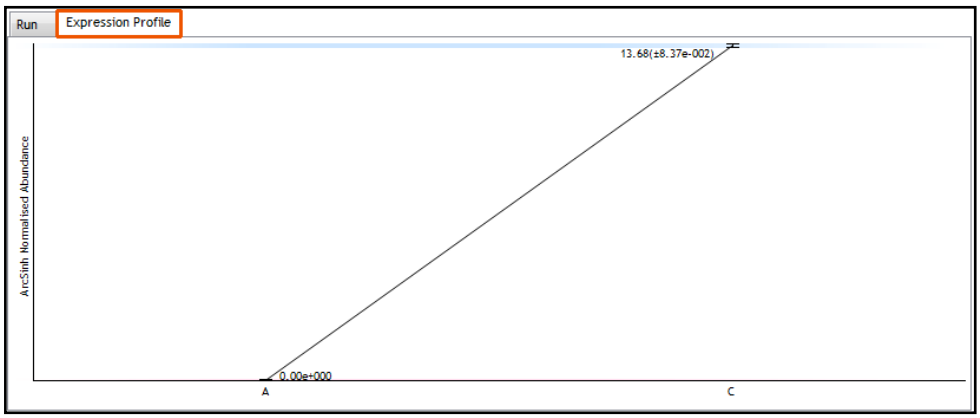
**Window D**: displays the details of the currently selected run. By default the selected run is an Aggregate of all the aligned runs.



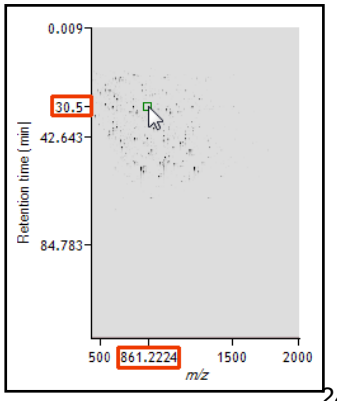

Details of individual runs can be viewed by using the 'Run' link and selecting the run you wish to view.

The feature editing tools are located in this window (see page 27 for functional explanation).

Clicking on the Expression Profile tab in Window D shows the comparative behaviour of the feature across the various biological groups based on group average normalised volume. The error bars show +/- 3 standard errors.



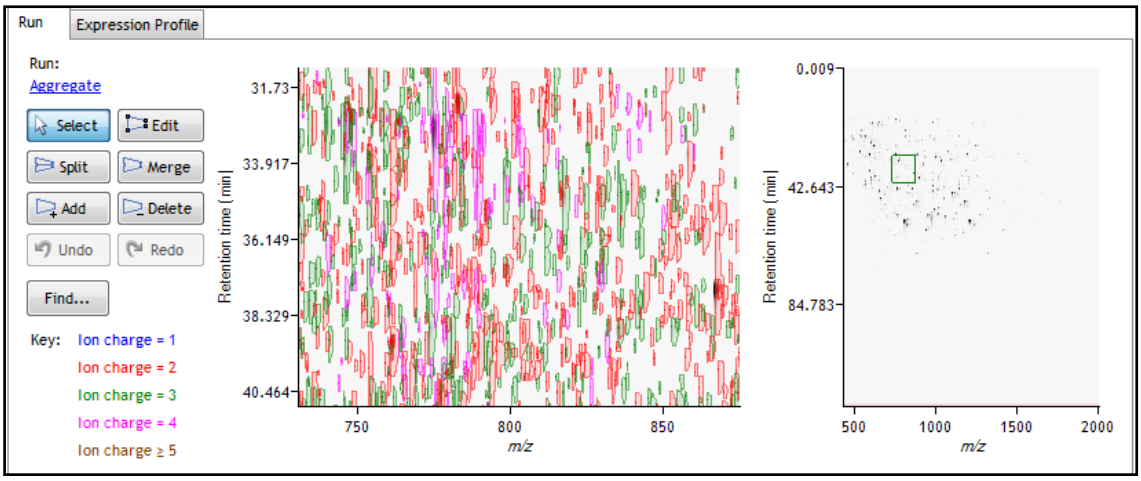

**Window E**: shows where the current feature is located on the LC-MS run by means of the 'Green' rectangle.




To change the current location, click on the image of the run (**note**: the retention time and m/z values update as you move the cursor around this view).

**Note:** doing this updates the focus of all the other windows.

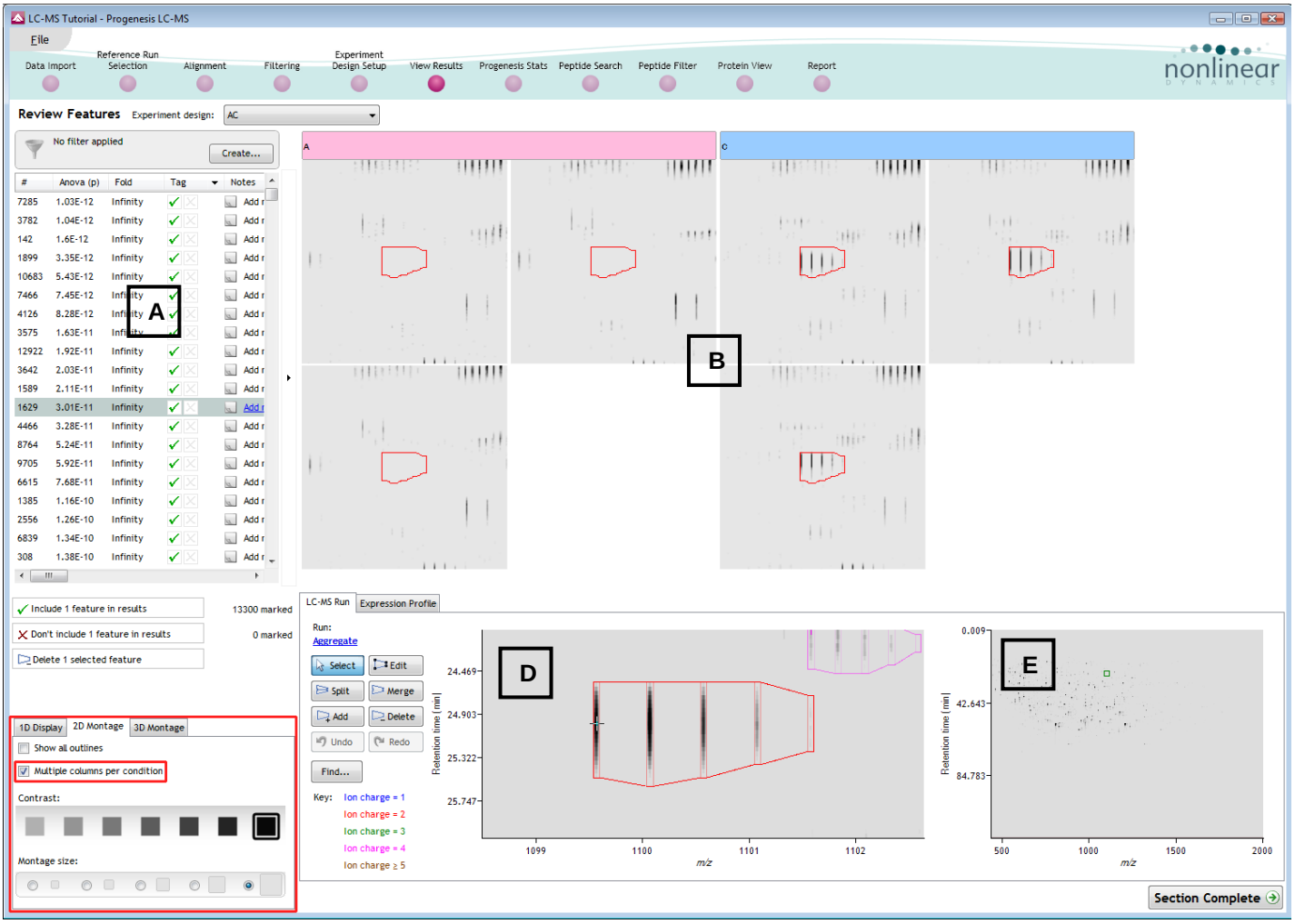
You can also drag out a larger area on this view that will refocus the other windows.




### The 2D Display

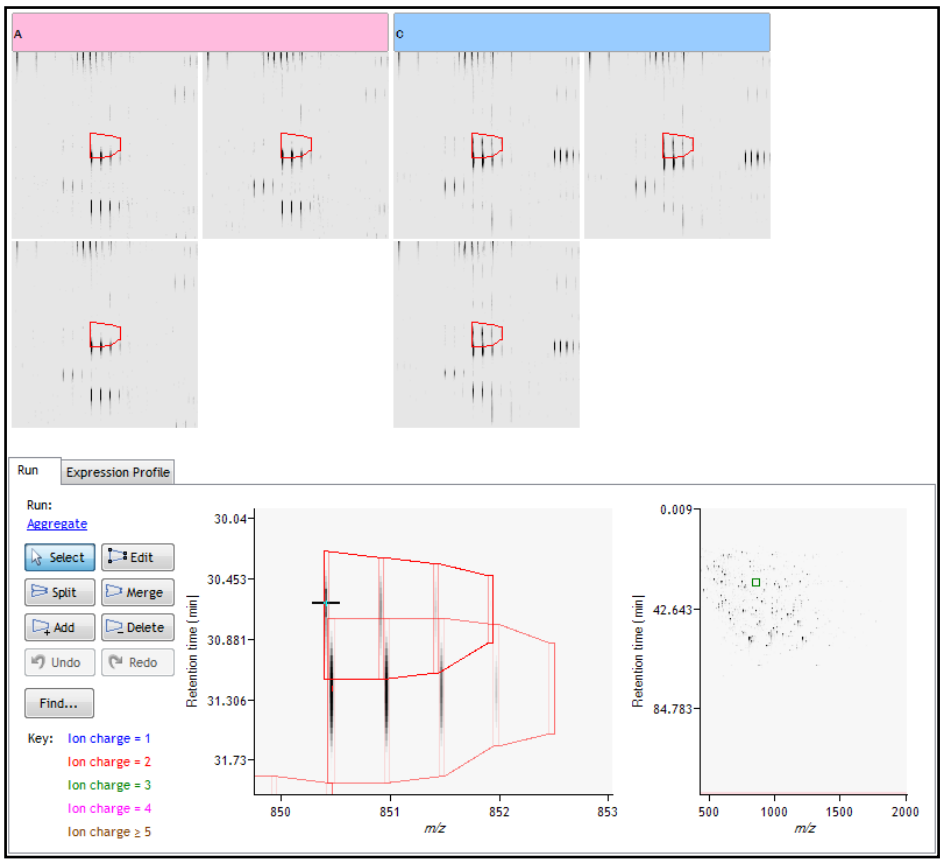
Windows **A, D and E**: perform the same functions across all 4 display modes.

In the 2D Montage mode, Window B displays a montage of the current feature across all the aligned LC-MS runs.



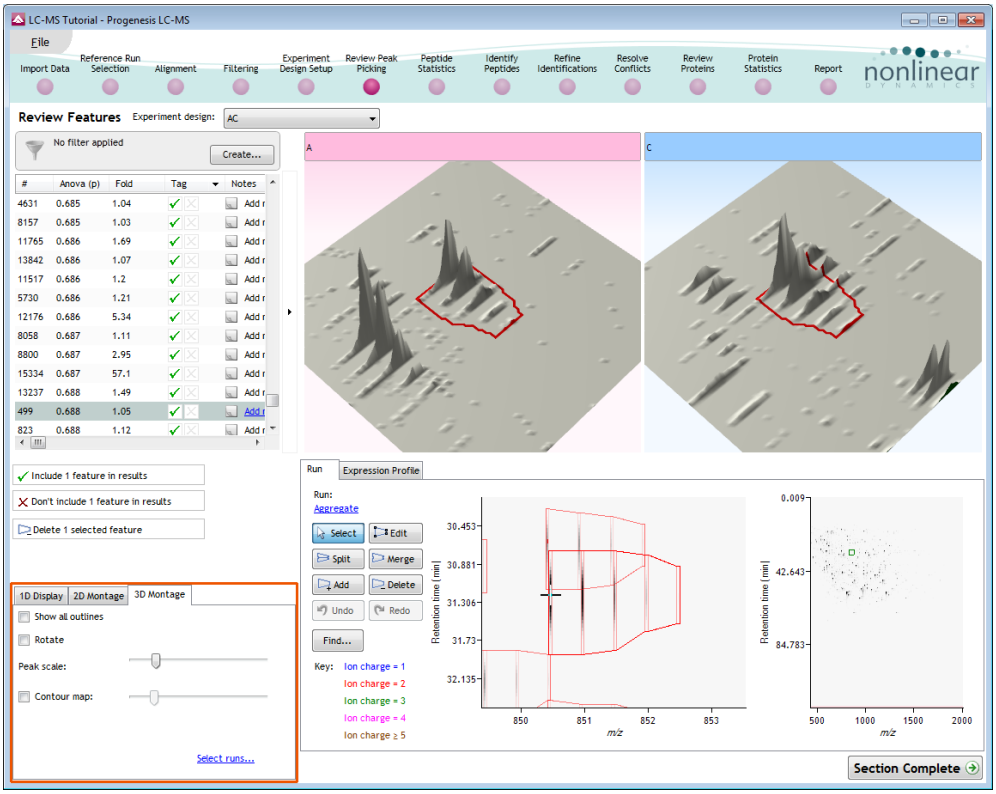

The appearance of the Montage (window B) is controlled by the panel on the bottom left of the display.

Using the the various views in the 2D display one can examine the feature detection in detail to validate the correct detection of even fully overlapping features as shown above.




### The 3D Display

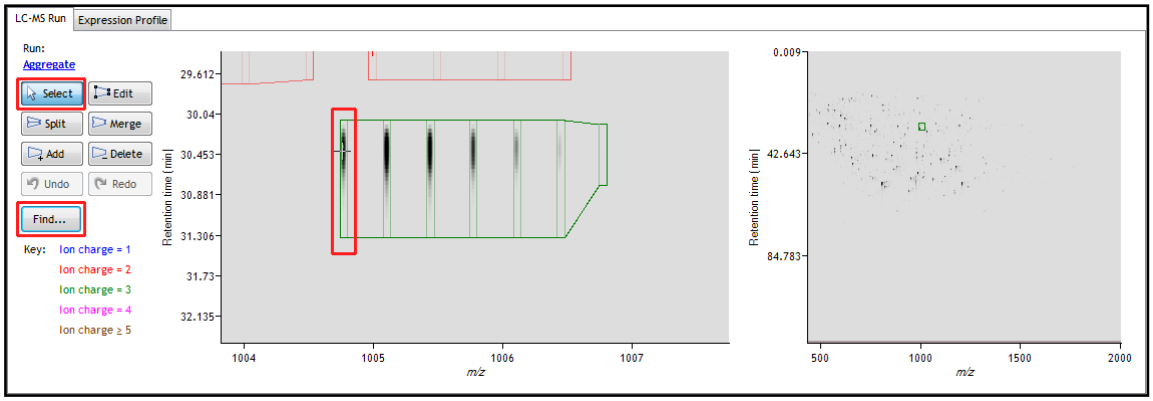
Window B changes into a 3D view by selecting the 3D Montage tab on the bottom left of the display.



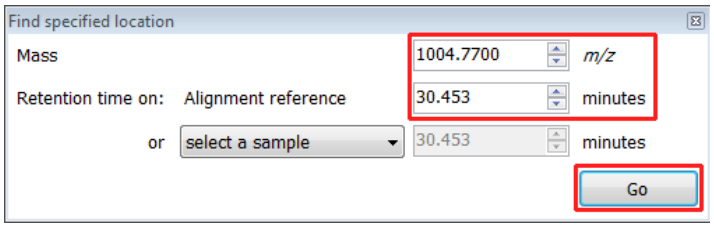

The number of 3D views displayed in the montage is controlled using the Select runs link on the 3D Montage tab. The views can be set to **Rotate** automatically or you can rotate them manually by clicking and dragging them with the mouse.

## Editing of features in the View Results stage

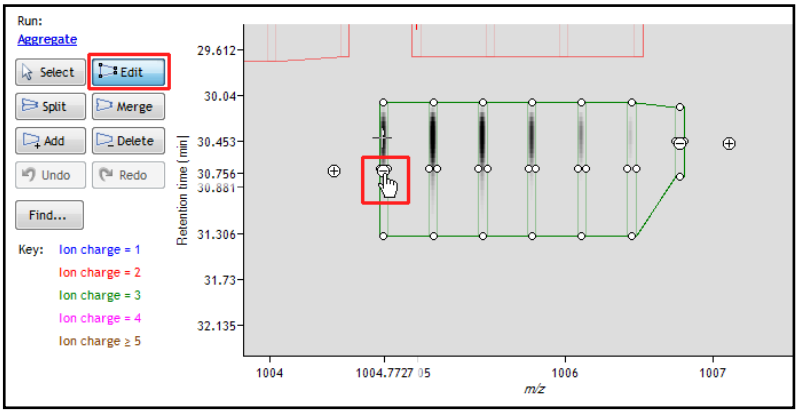
As an example of using the editing tools which are located on the left of the LC-MS Run view, we will remove and add back the 'monoisotopic peak' for the detected feature selected below. A feature can be selected from the 'Features' list or located using the various views.



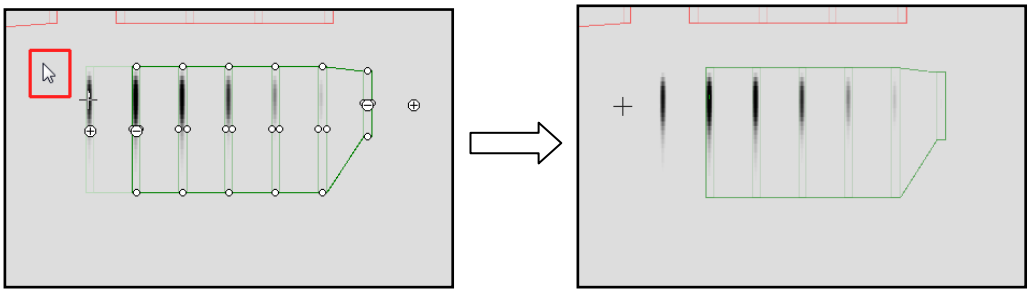

1. Locate the feature at approx 1004.77 m/z and 30.453 min using the **Find** tool.




2. Select the **Edit** tool and click on the feature to reveal the 'edit handles'

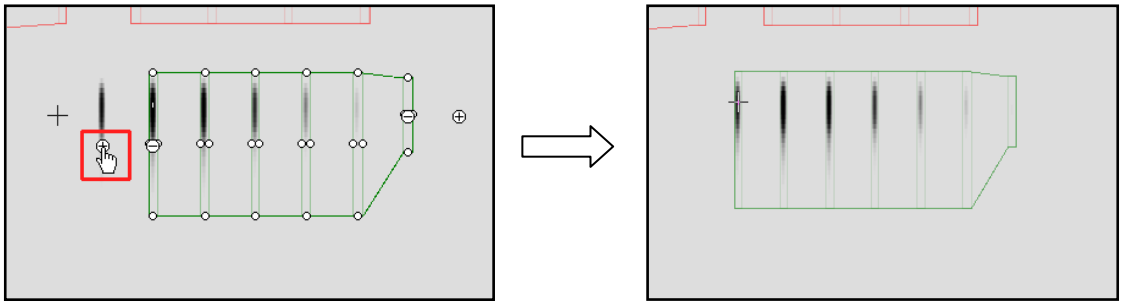



3. Click on the 'minus' handle over the monoisotopic peak to remove it.



4. Click outside the boundary of the feature to update the view.

5. To add a peak to an existing feature, ensure that **Edit** is selected then click inside the feature to reveal the handles.

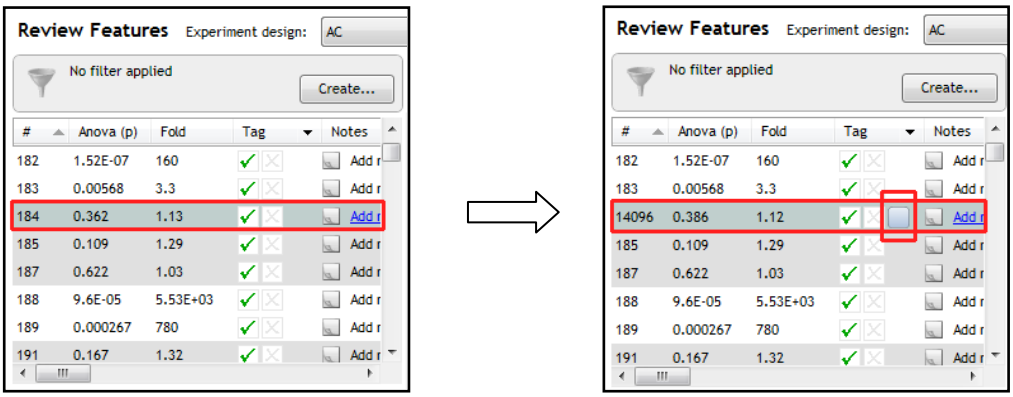


6. Click on the ‘plus’ handle on the peak to add it.

7. Then click outside the feature to update the view.

8. **Note**: If you are not satisfied with the editing use the **Undo** button and retry.

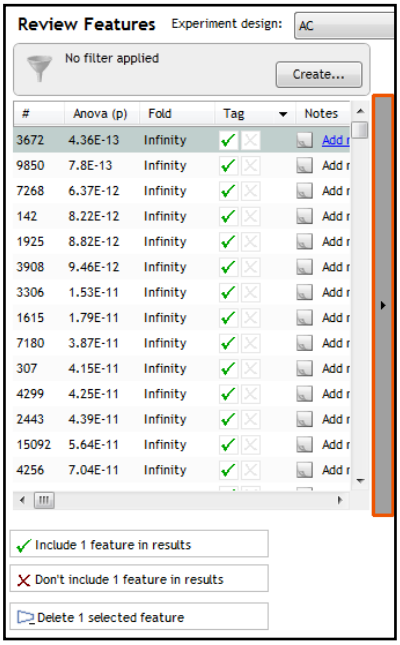
9. Finally note: that a tag is automatically added to the edited feature in the table and the features id number is changed to the next available one at the end of the list.




The other tools: split, merge, add and delete behave in a similar fashion and their use can be combined to achieve the desired results.

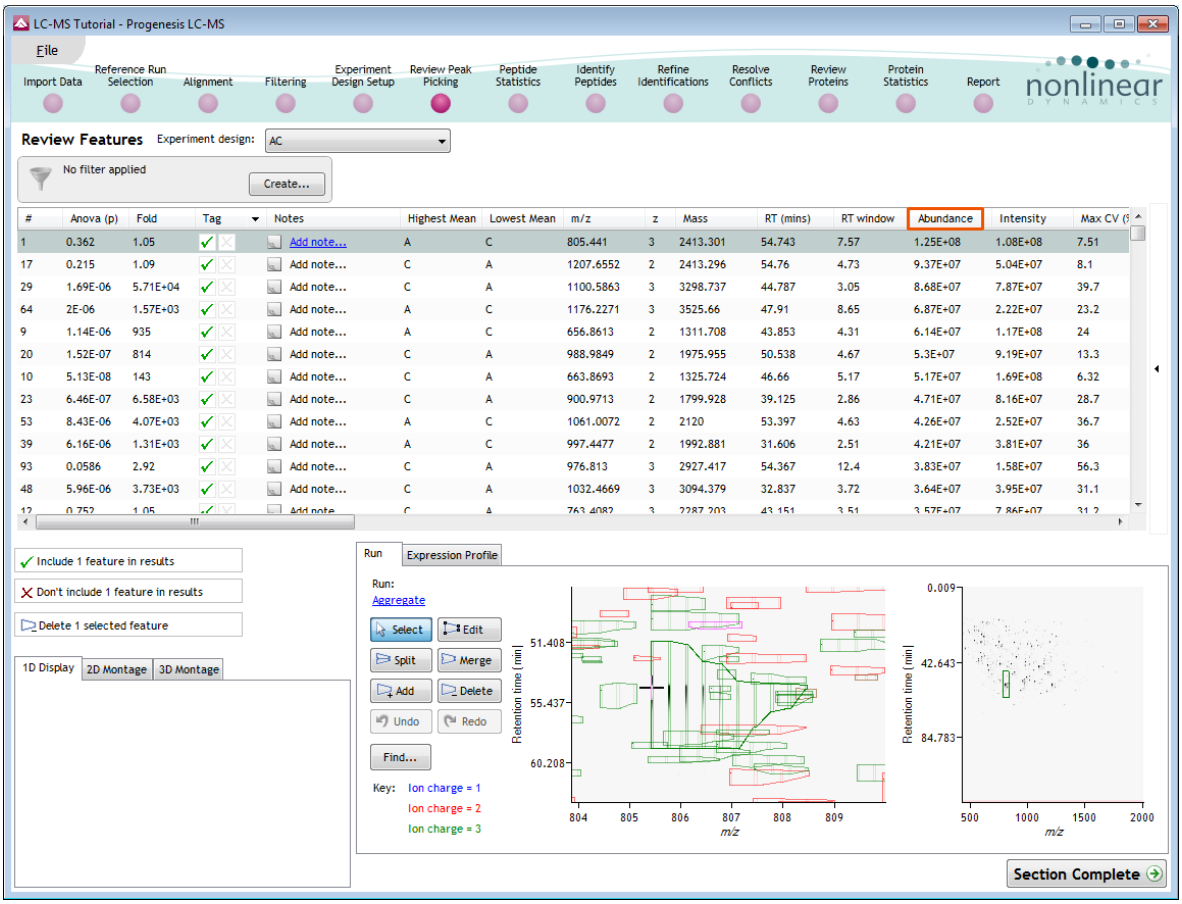
## Selecting and tagging features for Peptide Statistics

There are a number of ways to ‘refine’ your ‘Ranked List’ of analysed features before examining them with the Statistical tools in **Peptide Statistics**. These make use of simple ‘Selection’ and ‘Tagging’ tools that can be applied to the various groupings created in Stage 6 (page 20). An example is described below.

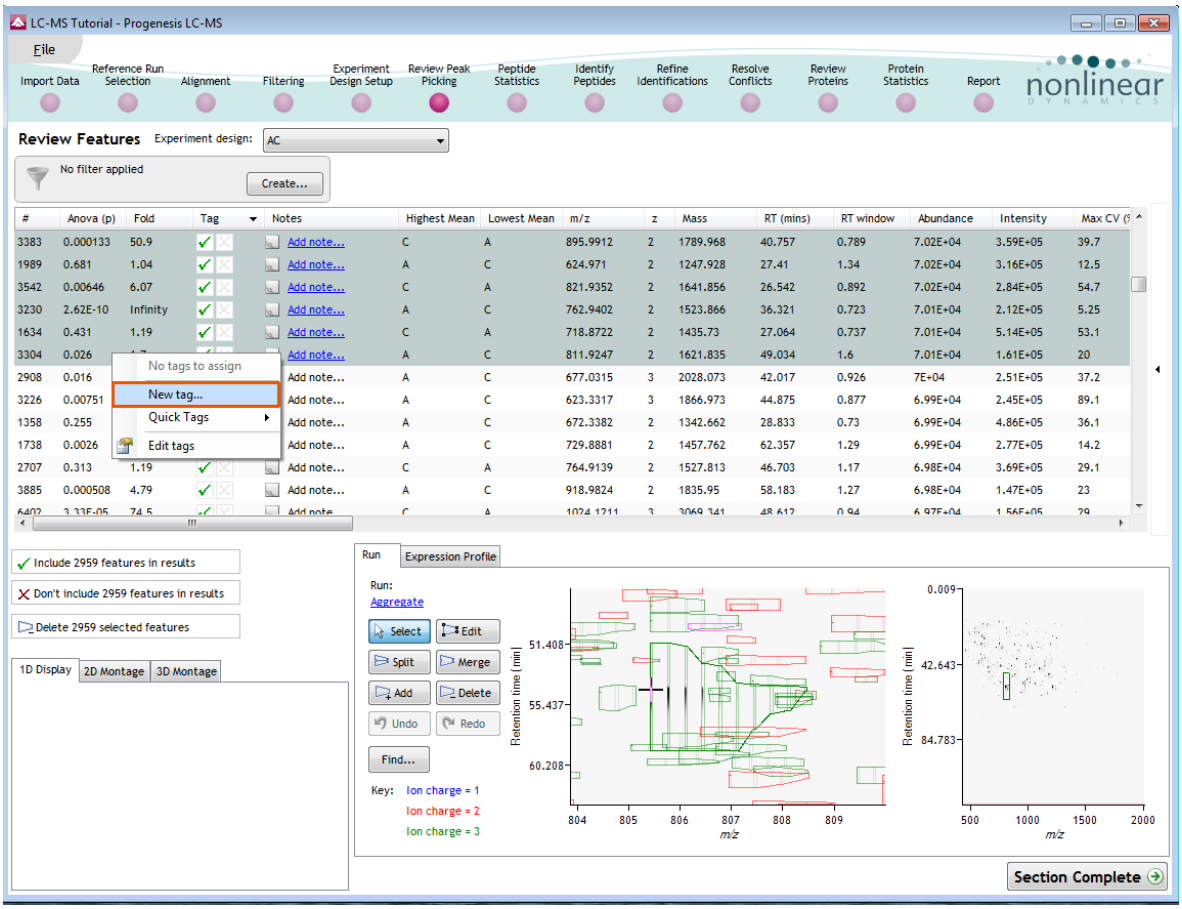



First expand the 'Features' table to show all the details by clicking on the ‘Expander bar’ to the right of the Review Features table.

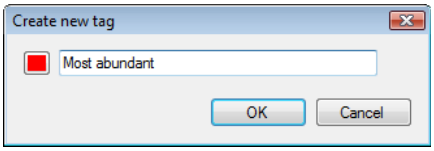
Then order on Abundance and select 2000 to 4000 of the highest abundance features, the exact number is not important.



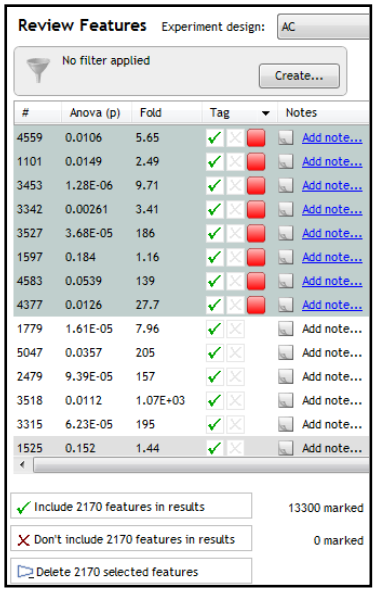

With the 2000-4000 features still highlighted right click on them and select 'New Tag'.



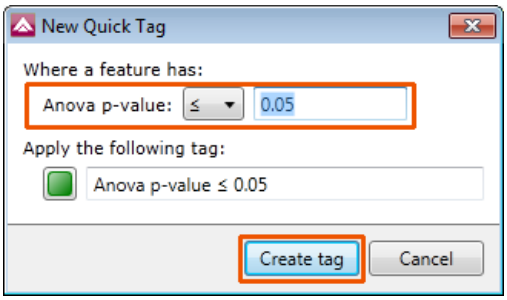
Give the Tag a name. i.e. **'Most abundant'**.

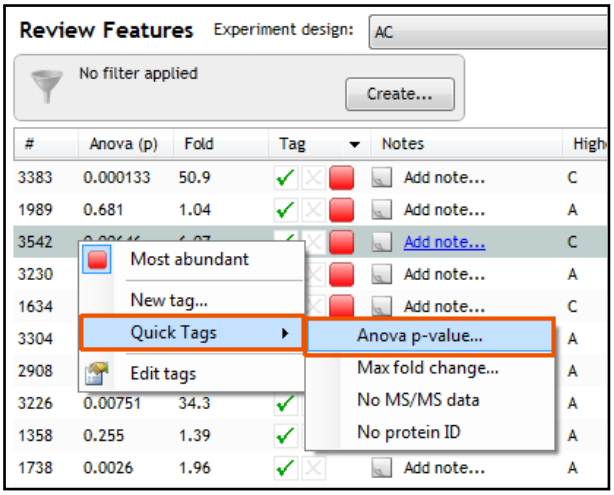



On clicking **OK** the Tag is added to the features highlighted in the table (signified by a coloured square).

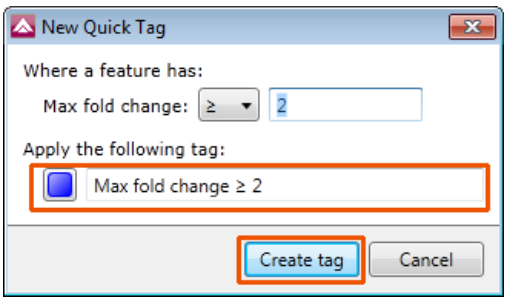


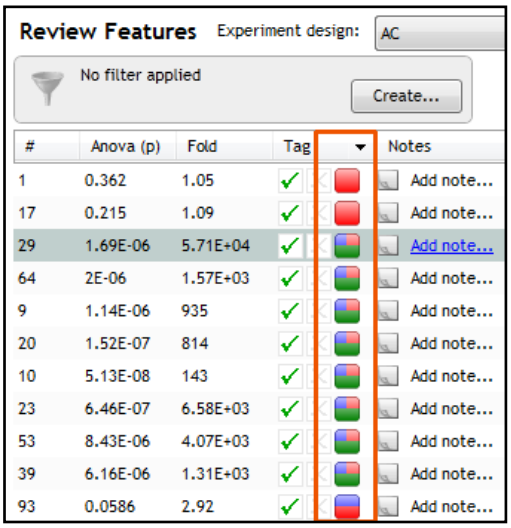

Now right click on any feature in the table and select **Quick Tags** this will offer you a number of standard tag options. Select **Anova p-value....** Then set the threshold as required and adjust default name as required and click **Create Tag**.




Once this tag appears against features in the table right click on the table again and create another Quick Tag, this time for features with a **Max fold change ≥ 2**



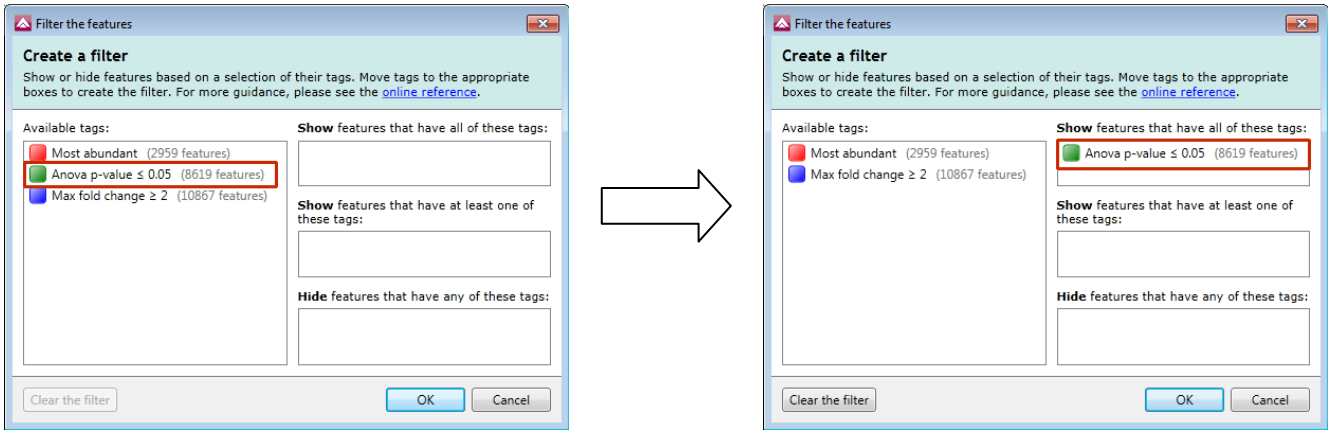




The table now displays features with multiple tags.

The tags can be used to quickly focus the table on those features that display similar properties.

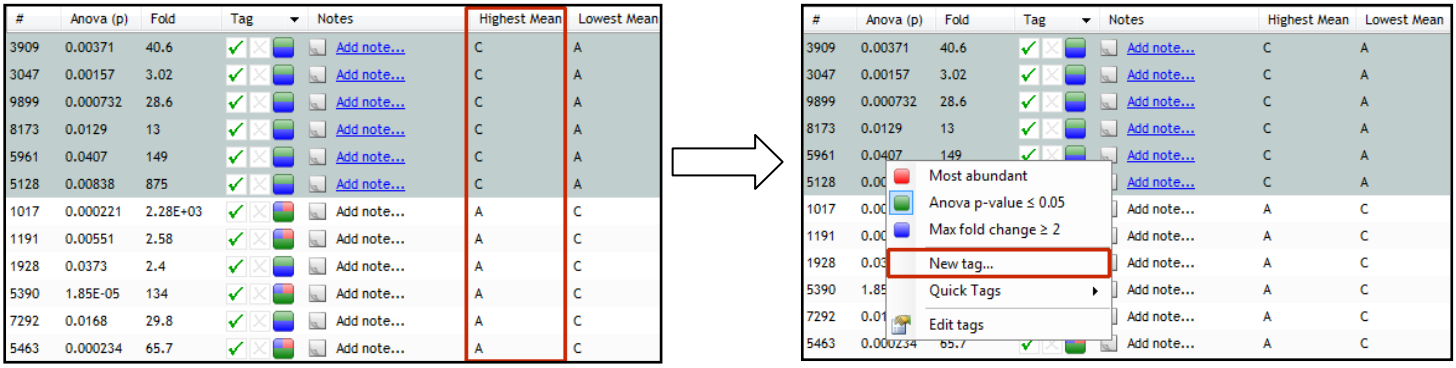
For example: to focus the table on displaying those features that have an **Anova p-value ≤ 0.05** click on **Create** on the filter panel above the table.


Drag the tag on to the panel **Show features that have all of these tags** and press OK.



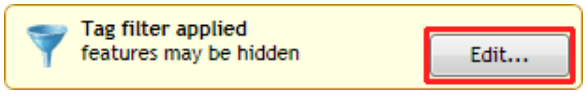

Now order the current features in the table by the **Highest mean** so that all the features showing the highest mean for **condition C** are at the top of the list.

Then highlight all the features with the highest mean for **condition C** and create a new Tag for them.



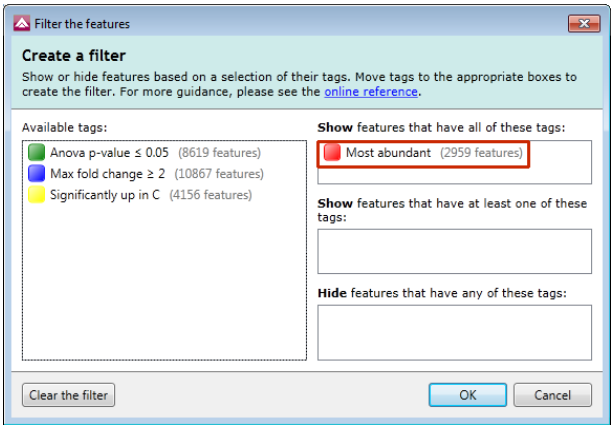

Create a tag for them called **Significantly up in C,** tagging 4156 features

Finally view the tags you have just created by clicking on Edit in the Tag filter panel, above the table.




Make sure that only the tag for the M**ost abundant** features is shown and press **OK.**

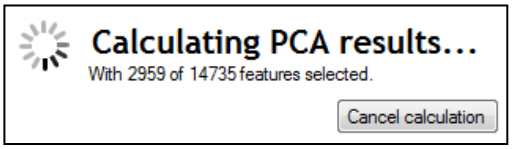
To move to the next stage in the workflow, Peptide Statistics, click **Section Complete.**



# Stage 8: Peptide Statistics on Selected Features

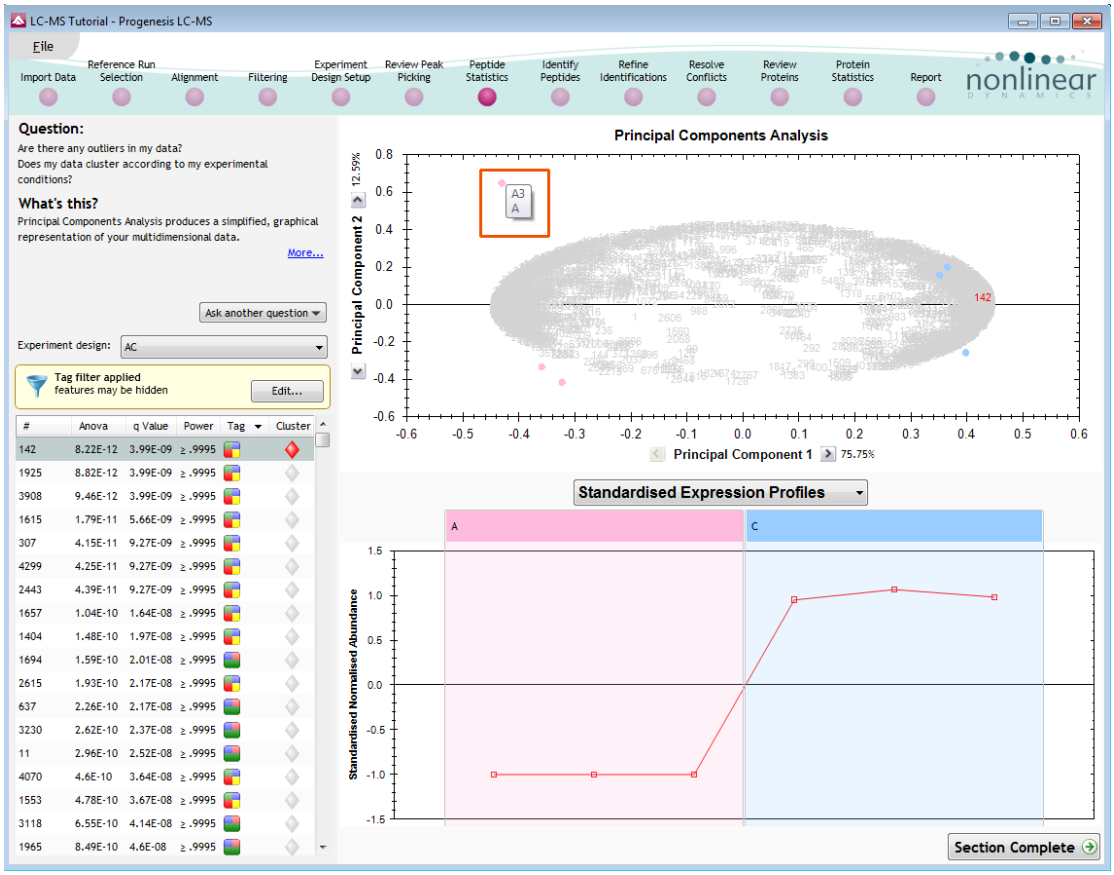
The user guide now describes the functionality of the Multivariate Statistics. This section is only available if Progenesis Stats is licensed.

Peptide Statistics opens calculating the Principal Components Analysis (PCA) for the active 'tag' in this case the **Most abundant** features.




As an example we will start by examining the behaviour of the **Most abundant** features from the previous stage, **Review Peak Picking**.

The statistical analysis of the selected data is presented to you in the form of interactive graphical representation of answers to questions asked of the analysed data.




**Note**: the LC-MS runs (samples) are displayed as solid coloured circles on the plot. To identify the runs, a tooltip is displayed when the cursor is held over each circle.

### Principal Component Analysis (PCA)

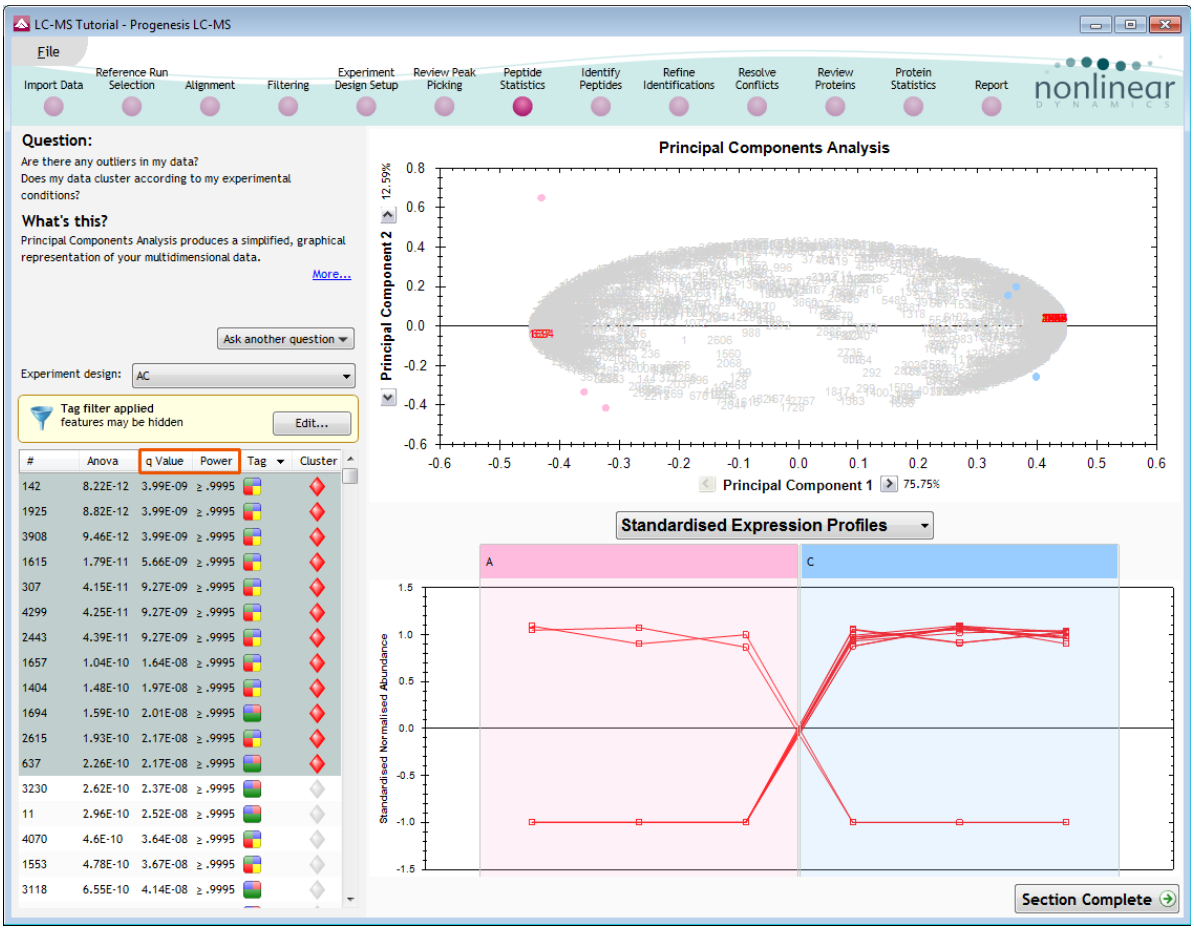
In **Progenesis Stats** the first statistically based question asked of the data takes the form of a Quality Control assessment:

*Are there any outliers in my data? And does my data cluster according to my experimental conditions?*

It answers this question by:

*'Using Principal Components Analysis (PCA) to produce a simplified graphical representation of your multidimensional data'.*

PCA can be used to determine whether there are any outliers in the data and also look at how well the samples group. The groupings that can be observed on the 2D PCA plot can be compared to your experimental conditions and conclusions can be drawn regarding possible outliers in your data. Selecting features in the table will highlight the features on the 'Biplot' and their expression profiles will appear in the lower panel.




**Note**: the Table in the Stats view contains additional columns:

**q value**: tells us the expected proportion of false positives if that feature's p-value is chosen as the significance threshold.

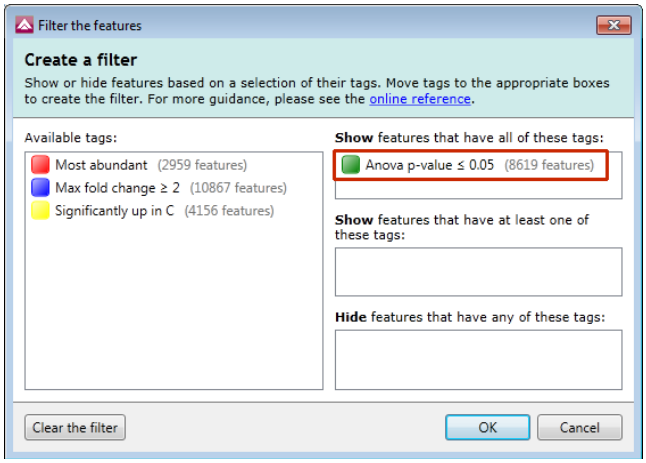
**Power**: can be defined as the probability of finding a real difference if it exists. 80% or 0.8 is considered an acceptable value for power. The Power Analysis is performed independently for each feature, using the expression variance, sample size and difference between the means.

Also, for a given power of 80% we can determine how many samples are required to ensure we find a difference if it actually exists.

**Note**: Power analysis is discussed in **Appendix 6** (page 66)

## Correlation Analysis

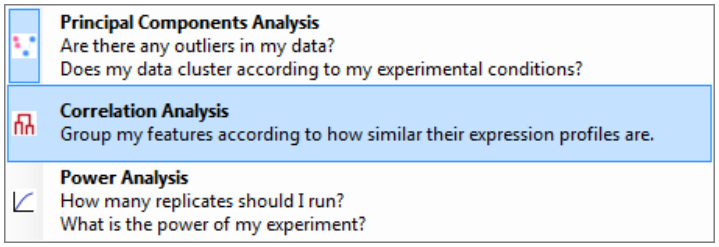
Use the tags, created in Review Peak Picking, to filter the features in the table.




We are going to explore the Correlation Analysis for all the features that were tagged at the view results stage for having an **Anova p-value≤0.05**.

On pressing OK the PCA will recalculate using these 8618 features, you can (to save time) stop this calculation by pressing **Cancel calculation** and then set up Correlation Analysis for the 8618 features.

To set up the **Correlation Analysis** using this filtered data set click on **Ask another question** (above the table)




A selection of 3 tools will appear in the form of questions.

Select the second option to explore 'feature correlation based on similarity of expression profiles'

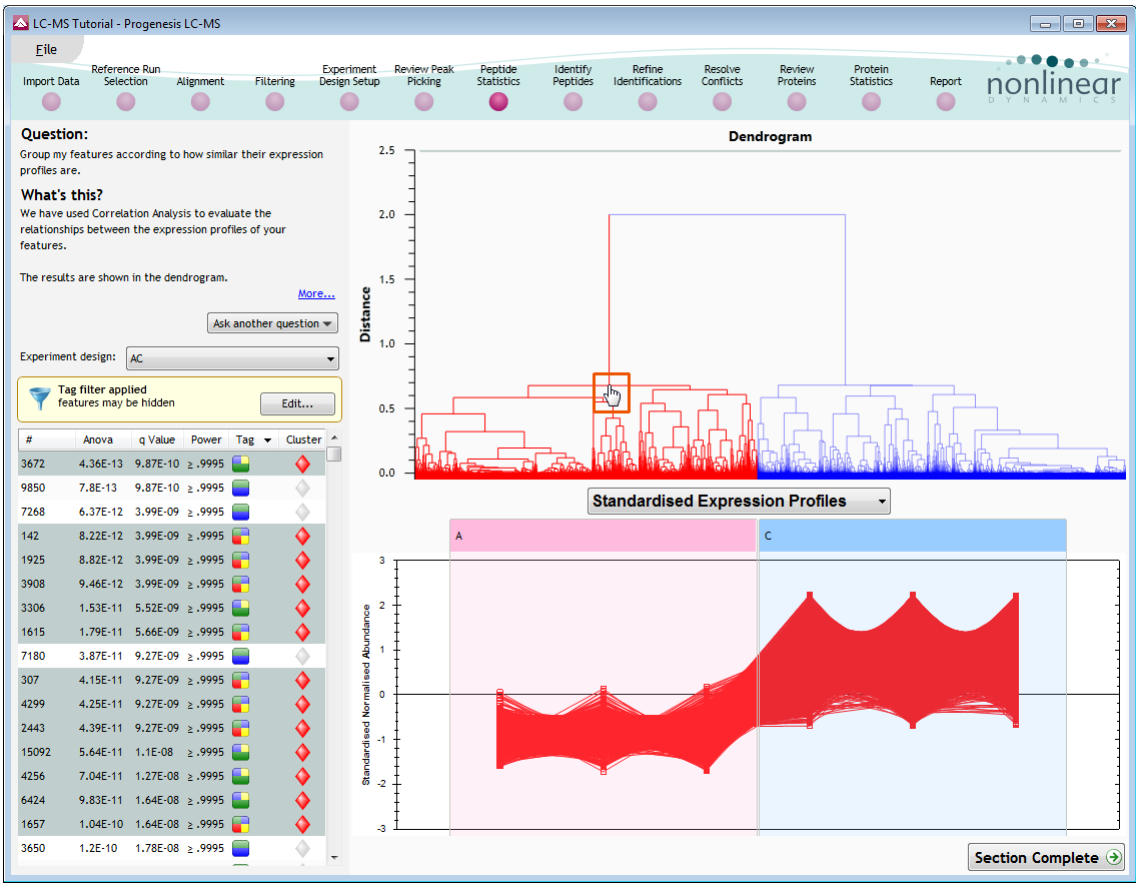

This time the statistically based question(s) being asked is:

*'Group my (selected) features according to how similar their expression profiles are'*

The question is answered by:

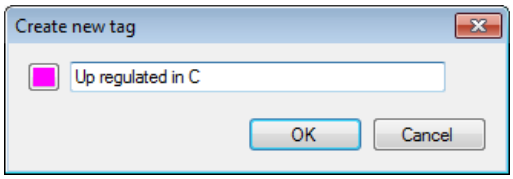
*'Using Correlation analysis to evaluate the relationships between the (selected) features' expression profiles'.*

The answer is displayed graphically in the form of an interactive dendrogram where the vertical distance, between each feature can be taken as indicative of how similar the expression profiles of each cluster of features are to each other.

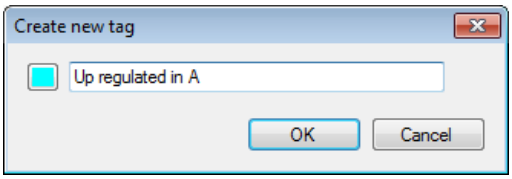



**Correlation Analysis** enables the grouping of features together according to how similar their expression profiles are.

To highlight all the features demonstrating **Increased expression in the C** group click on a 'node' for a branch of the Dendrogram (as shown above). As before right click on the highlighted features in the table and create a Tag for these features (Up regulated in C).

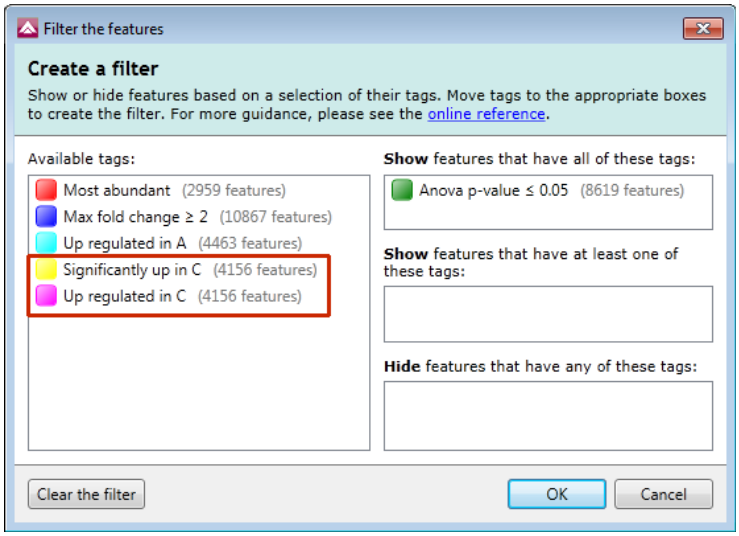



Also create a tag for those features showing **Increased expression in A** by first clicking on the other 'main' node then right click on the highlighted features in the table and creating the New tag (Up regulated in A).




**Comment**: When you review the tags using **Edit** you can see that the Magenta and Yellow tags have been assigned to the same number of features. This shows how tabulated information about features can be used alongside interactive graphical plots of multivariate statistical analysis to explore your data.

**Note**: two groups is a special case, for more groups this will not be the case additional tagging will be required.

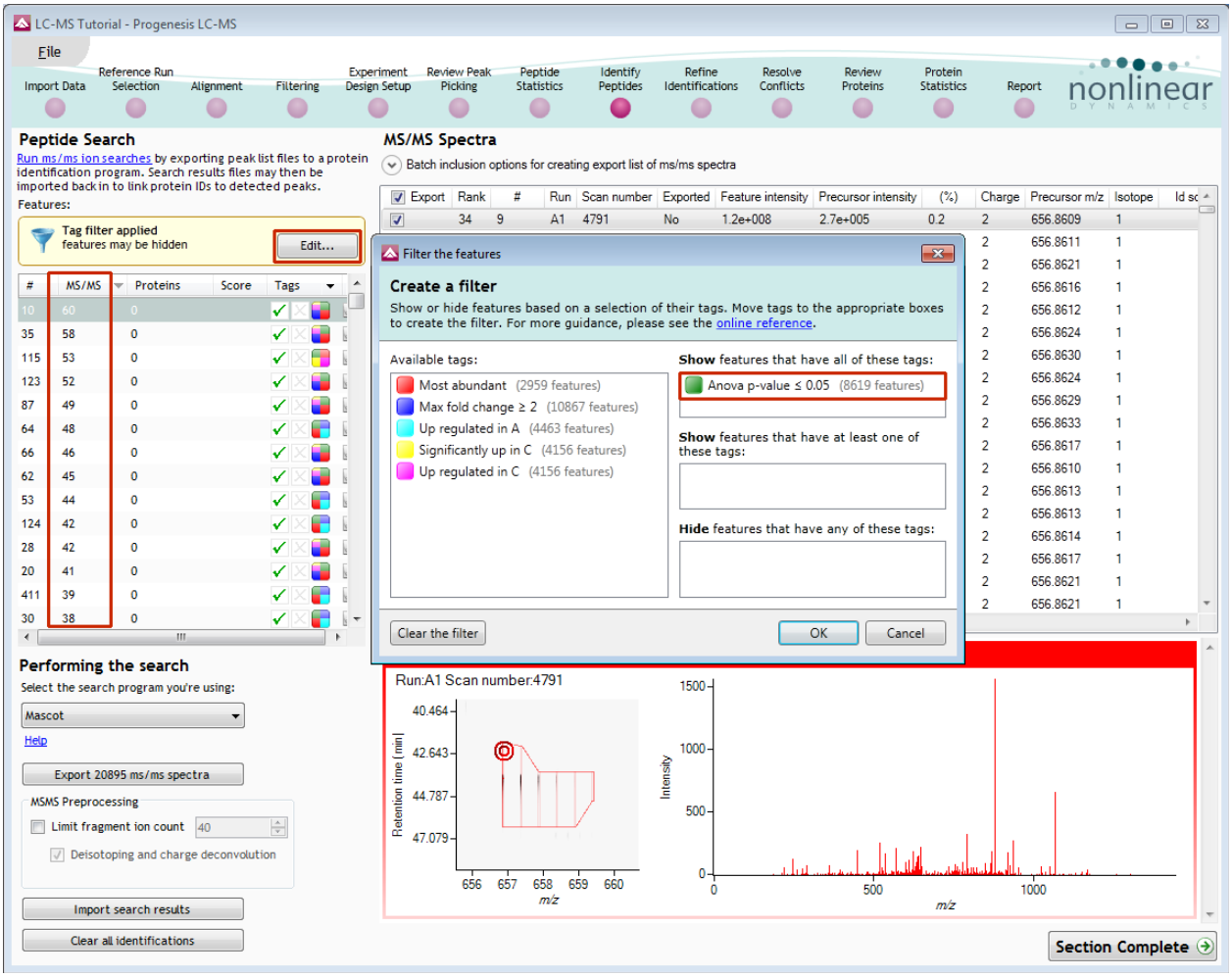



***Tip***: when reviewing the tags (see above) if you are not applying a new filter then use the **Cancel** button to return to the main view, this prevents unnecessary recalculation of your data.

To move to the next stage in the workflow, Peptide Search, click **Section Complete.**

## Stage 9: Identify peptides

Progenesis LC-MS does not perform peptide identifications itself. Instead it supports identifications by allowing you to export MS/MS peak lists in formats which can be used to perform peptide searches by various search engines. The resulting identifications can then be imported back into Progenesis LC-MS, using a number of different file types, and matched to your detected features.

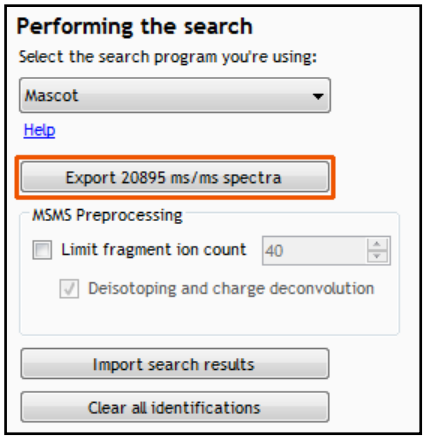


Determining protein identification is dependent on the availability of MS/MS data for the LC-MS runs. This data may be available but limited if the LC-MS was performed in a data dependant MS/MS detection mode due to under sampling. Under these conditions MS/MS data acquisition is dependent on thresholds and parameters set prior to the acquisition of the LC-MS run.

For this example we are using LC-MS runs containing MS/MS data where the data was acquired in a data dependant mode (DDA).

The Peptide Search page shows the number MS/MS that have been matched to each feature in the Feature list (see above). MS/MS scans are matched to a feature if their precursor *m/z* and aligned retention time fall within the area of one of the isotopes of the feature. The MS/MS scans which are matched to the displayed features are shown in the MS/MS spectra list on the right.

The first step is to decide which MS/MS scans you wish to export to be identified. By default this is all the available spectra for the features displayed in the Features list (in this case all the features that have an **Anova p-value≤ 0.05**. This number is visible on the Export button.




The set can be targeted using the tags and also refined with respect to quantity and quality of the spectra being sent to the search engine.

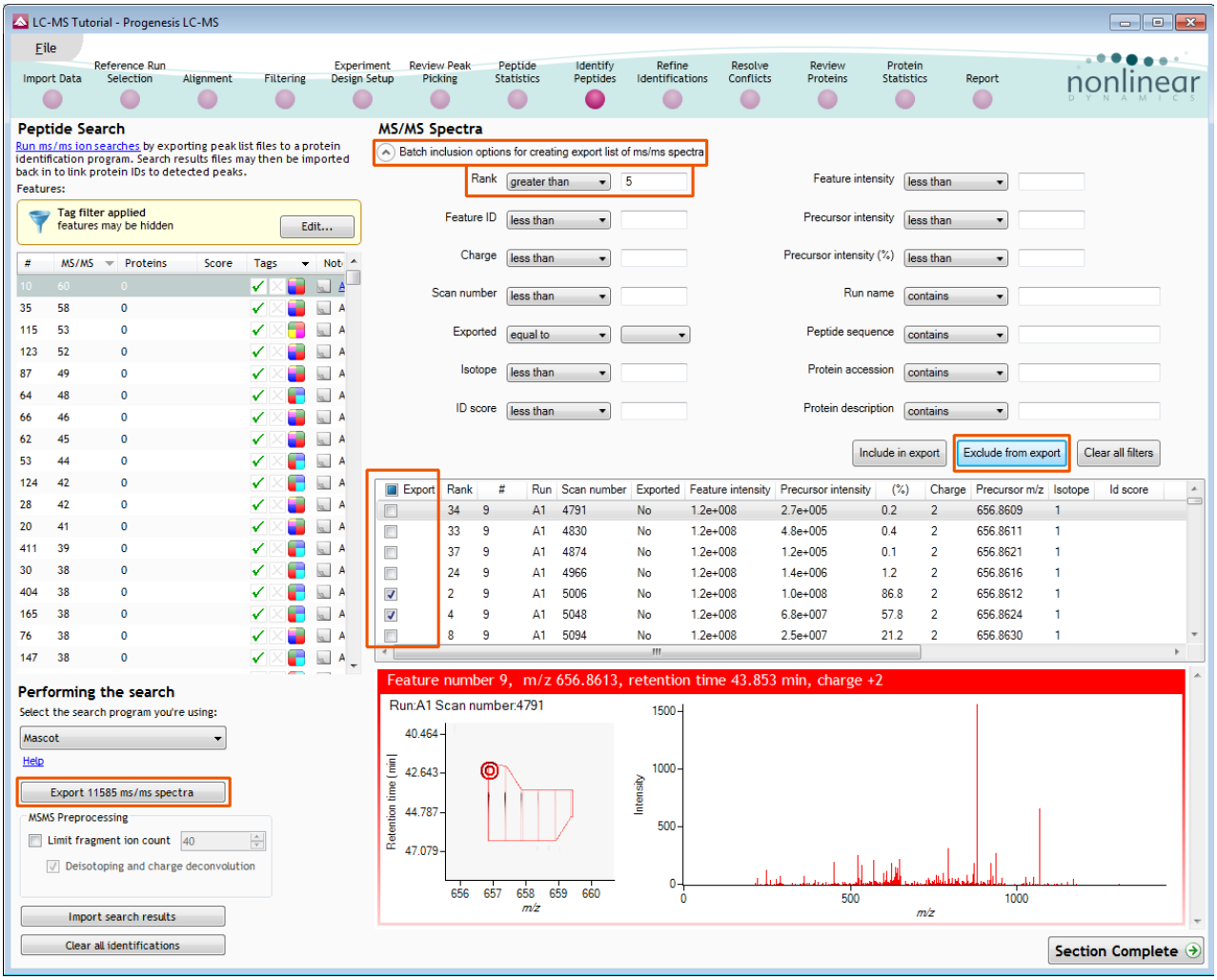
Filter the table to show only the features tagged **Significant p<0.05** as shown.

**Note**: by default the table is ordered on the number of MS/MS spectra available for each feature.

The total number of spectra included in this set is **20895** as shown on the Export button.

Before exporting the spectra, the set can be further refined.

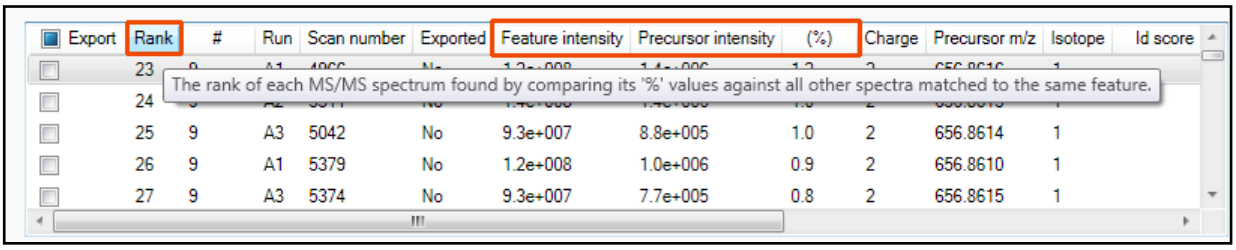
**Note**: many of the abundant features have a large number of spectra associated with them.



To control the number of spectra for each feature, expand the **Batch inclusion options**.

**For example**: We will make use of the 'Rank' value to reduce the number of Spectra being used for each feature to a maximum of 5.

The 'Rank' of each MS/MS spectra is determined by comparing its **% value against all other spectra matched to the same feature.**

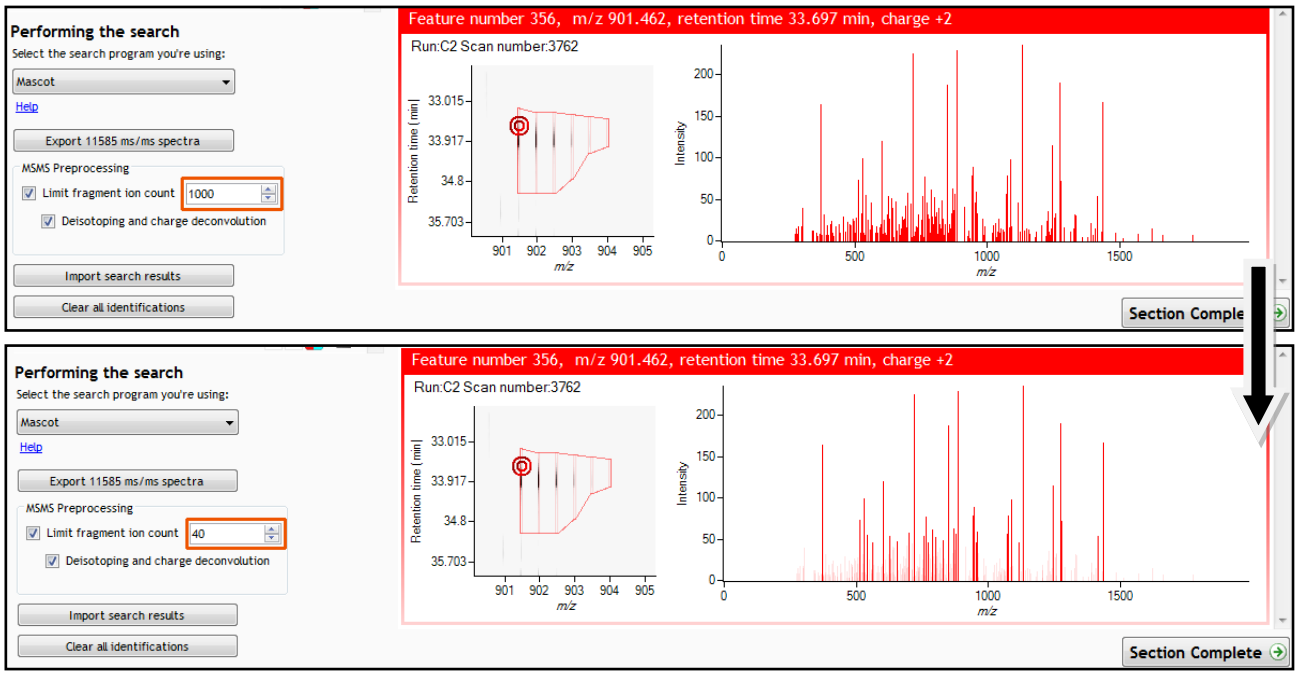



**Note**: the **% value** for each spectra is the **Precursor intensity as a percentage of the Feature intensity**

Set the Rank filter to 'greater than' 5 and click **Exclude from export** this reduces the number to spectra to export to **11585**.

Limiting the 'fragment ion count' (FIC) for the spectra being exported can improve the quality of the spectral data being used in the search by removing noisy peaks.

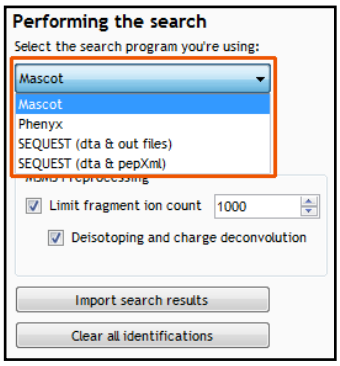
For example for the current spectra, reduce the FIC from 1000 to 40.




**Note:** the effect this has on the number of peaks in the spectra. This 'limitation' is applied to all the spectra being exported; hence the export file size will be reduced.

For this example we will **not** limit the fragment count, so leave it un-ticked (the default setting).

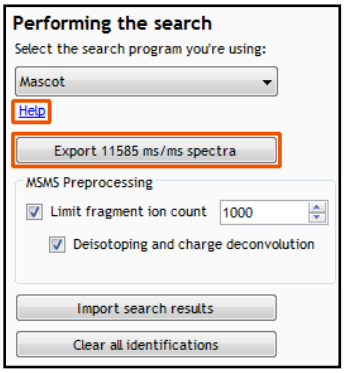
## Performing an MS/MS Ion Search




Having chosen 11585 spectra to export, as described above:

1. Select appropriate search engine i.e. Mascot
2. Click 'Export current query set' to save search as file
3. Perform search on appropriate search engine and save results file
4. Click 'Import search results', locate results file and open

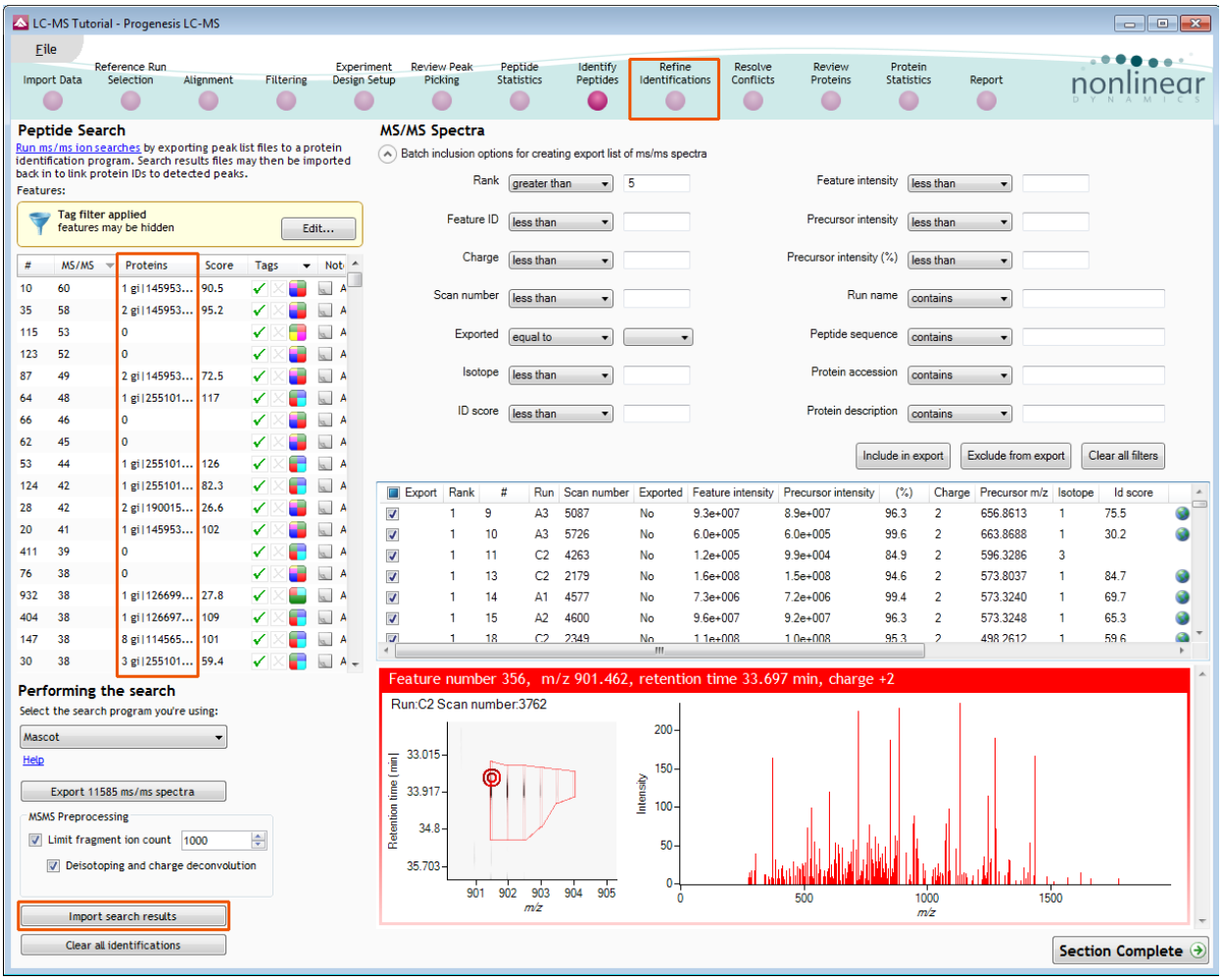
**Please refer to Appendix 7 (a and b) (pages 67 and 68) for details of the 'Search Engine' parameters**




**Note**: the blue link provides you with details on the appropriate formats for exporting search results and access to additional formats

**Note**: an example Search Results file, from a MS/MS Ion search, is available in the folder you restored the Archive to (Protein Search Results.xml). Select the 'Mascot' method and import this file to see results like those below.

On importing the Search results the Features table updates to reflect the identified proteins and the relevant score for each searched feature.



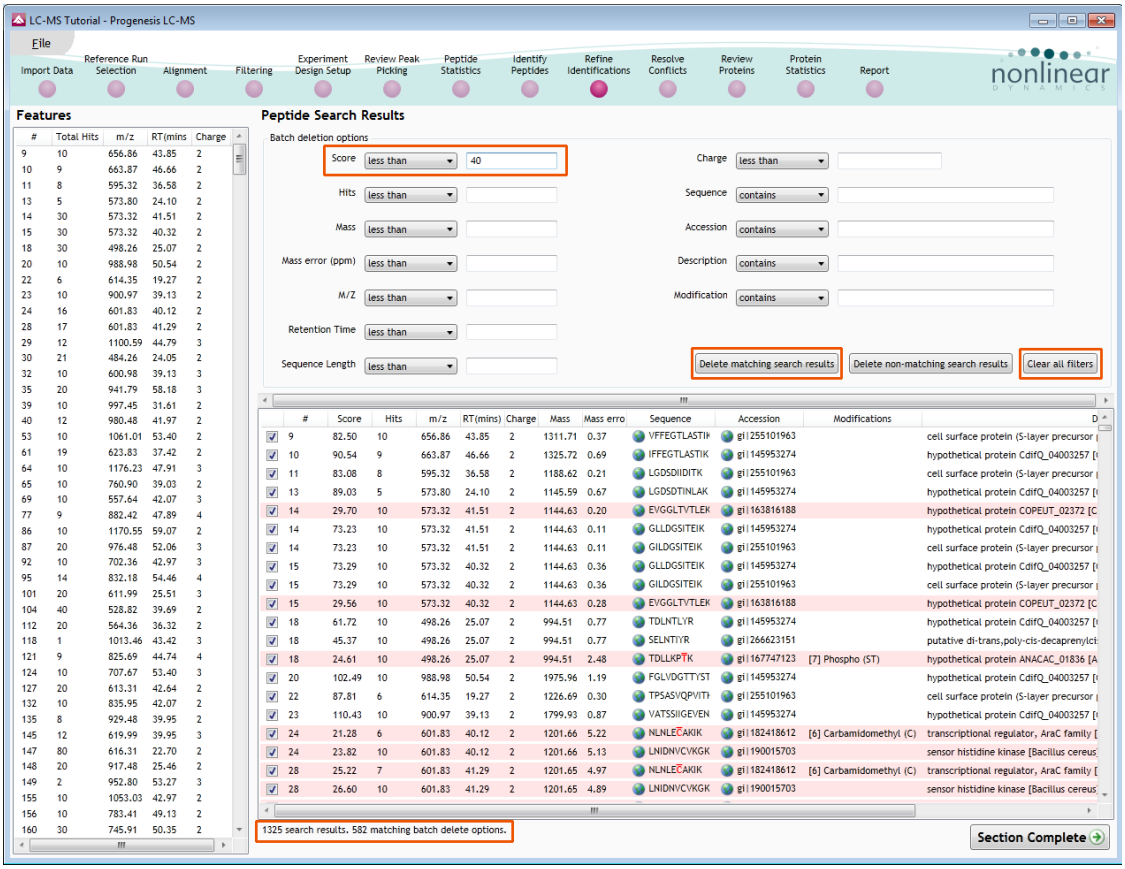

In order to review, and refine the quality of the **Peptide Search** results click on the next stage in the workflow, **Refine Identifications**.

# Stage 10: Refine Identifications

In this example the organism under study is *Clostridium difficile*

As an **example** 'Acceptance Criteria' on which to base the sequential filtering of the Peptide results, the following thresholds will be applied:

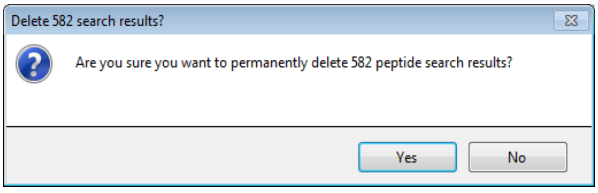
- Remove identifications with a Score less than 40
- Remove identifications where less than 2 hits were returned
- Remove all identifications where the Protein Description **Contains** 'hypothetical'
- Remove all identifications where the Protein Description **Doesn't contain** 'Clostridium difficile'




On the Batch detection options panel, set the Score to less than 40, then **Delete matching search results**.

**Note**: the search results matching the filter criteria turn pink and the total is displayed at the bottom of the table (582 matching out of 1325)

**Note**: a dialog warns you of what you are about to delete




Now **Clear all filters** and then apply the next filter (Hits: less than 2) followed by the remaining two filters (page 40)

Having applied all 4 filters the **Peptide Search Results** should be reduced to **475**.

To validate the Peptide search results at the protein level select the next stage in the workflow by clicking on **Protein View**.

# Stage 11: Resolve Conflicts

This stage allows you examine the behaviour of the identified peptides and resolve any conflicts for the various peptide assignments at the protein level.

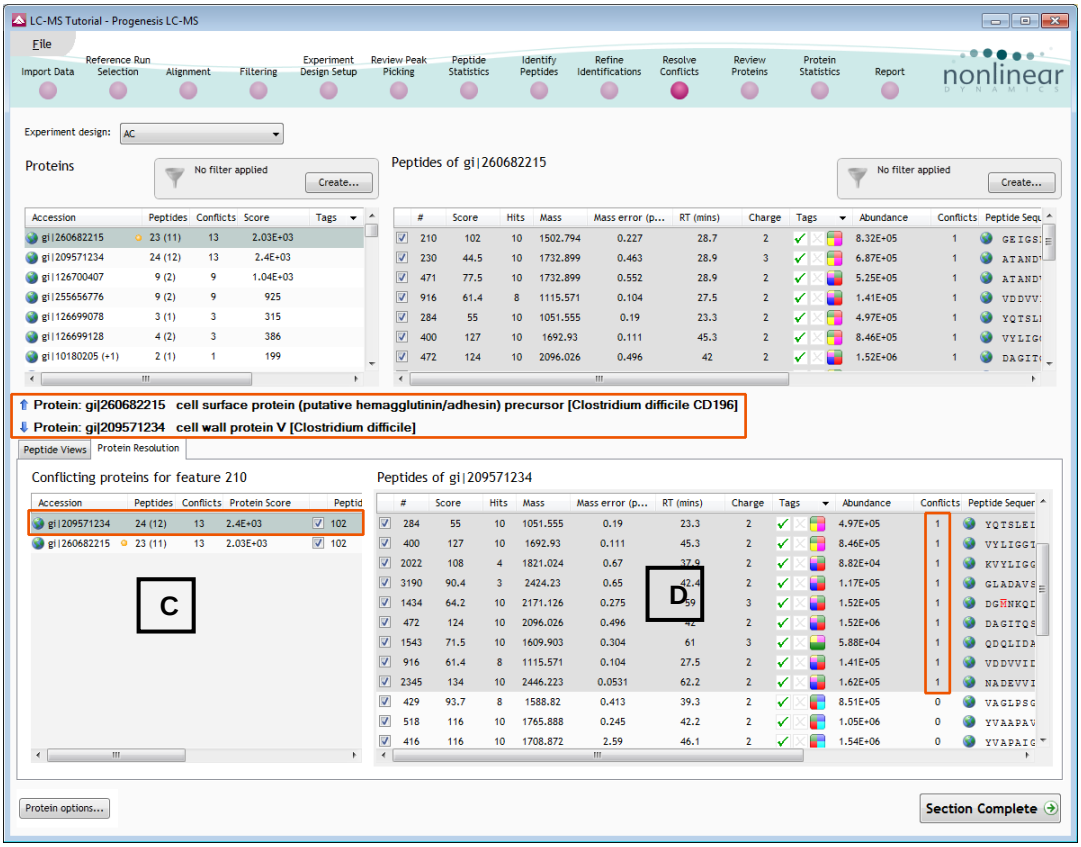
The **Resolve Conflicts** stage provides a number of interrelated graphical and tabular views to assist you in the validation of the peptides that have been assigned to proteins and also to review the relevance of the data returned from the search.

Open **Resolve Conflicts** and order the data in the Proteins table (A) on the basis of **Conflicts**.

**Note: the look of the tables (with regards to ordering) in the following section may vary slightly.**

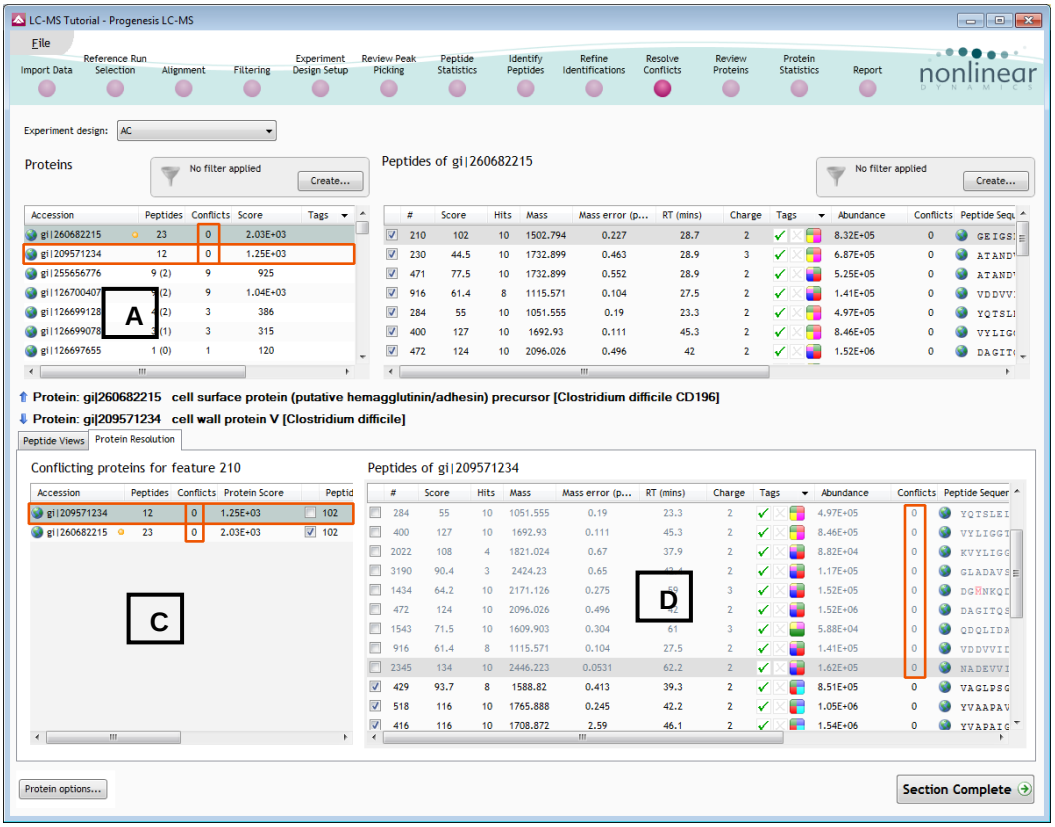

Depending on the ordering, make '**cell surface protein precursor etc '** the current protein by clicking on it in Window A (a circular orange symbol indicates current protein).

This protein has 23 peptides assigned (window B) which have a total of 13 conflicts. To view the conflicting assignments click on the **Protein Resolution** tab (window C).



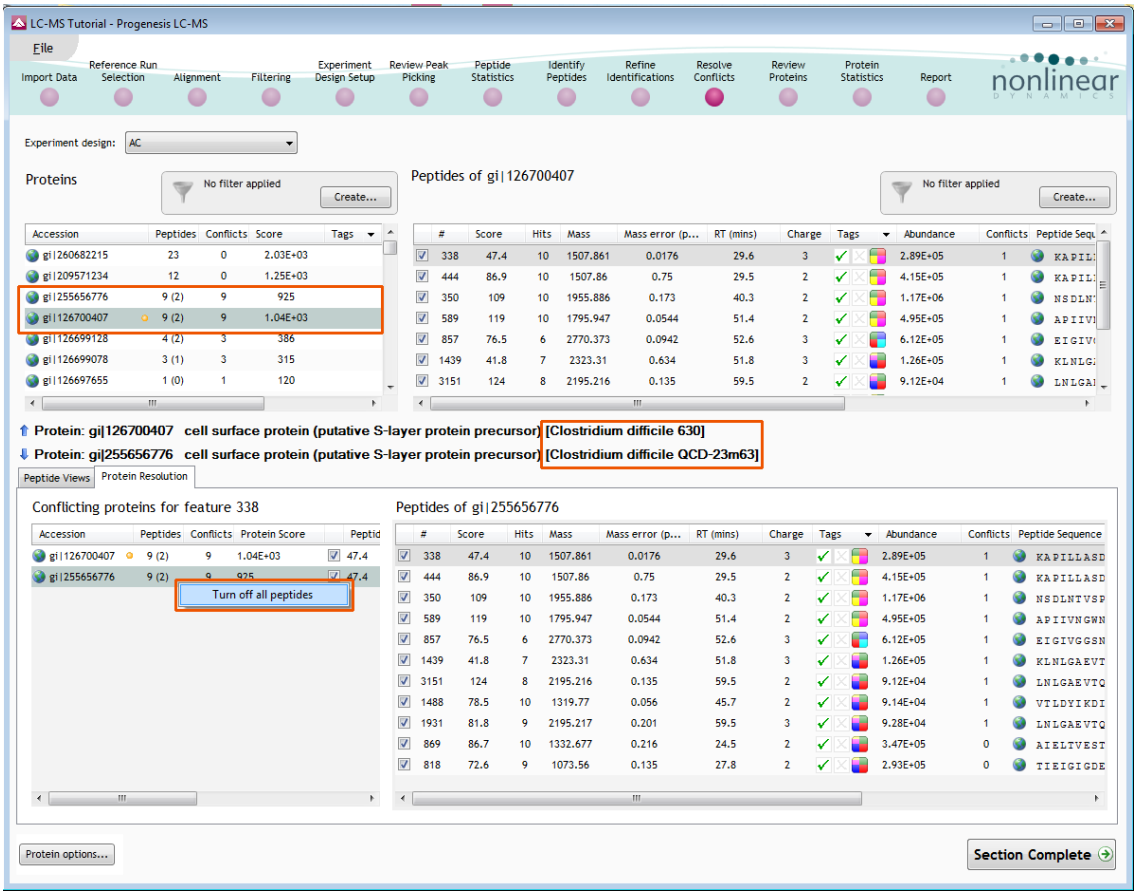

In this case the conflicting peptide assignments are with **Cell wall protein V** (from a different strain) which contains 13 conflicts. To resolve this conflict un-assign all the conflicting peptides (showing 1) assigned to the **Cell wall protein V**, by selecting and then unticking all the peptides in window D.

**Note**: as you un-assign the peptides the number of conflicts update 'on the fly' in all the windows.

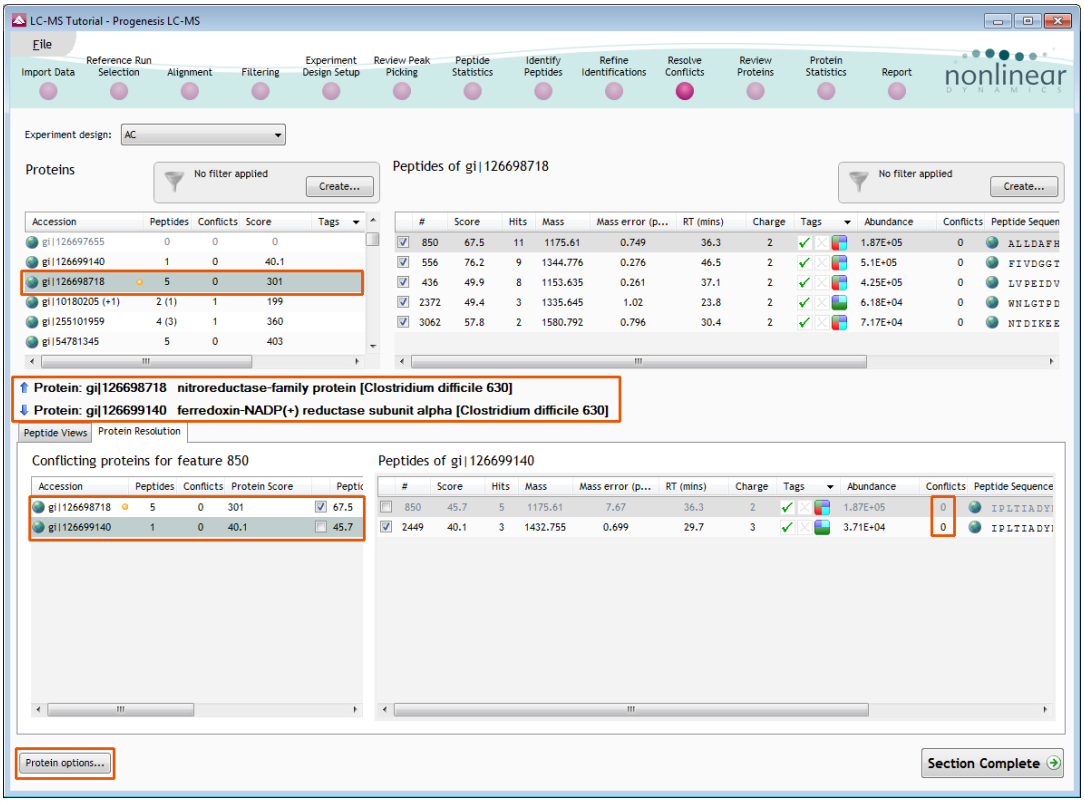


In this case the conflicting peptides are unassigned from the 'precursor' protein.

In the next example the conflict(s) are a result of entries in the data base being made for different strains of the same organism. In this case all the peptides are common to both strains therefore it would make sense to turn all the peptides off for one of the stains (i.e. retain peptides for the strain of interest). So right click on the conflicting protein and **'Turn off all peptides'.**



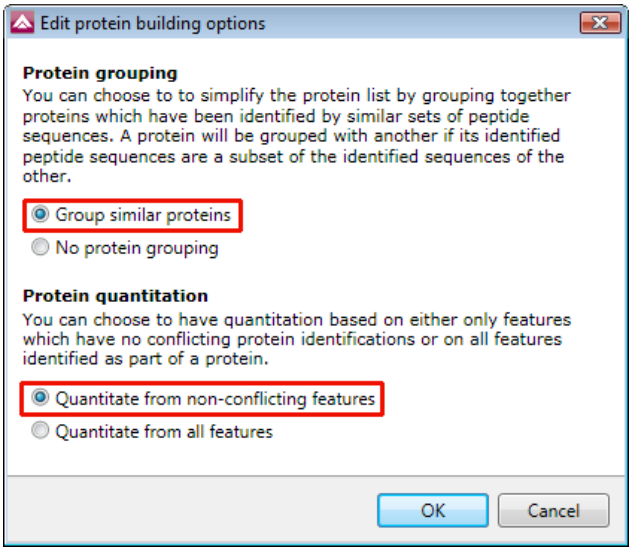

In many cases the rationale for resolving a peptide assignment conflict is based on the number of peptides assigned to each protein, often the conflict(s) being resolved are in favour of the protein with the greater number of assigned peptides.



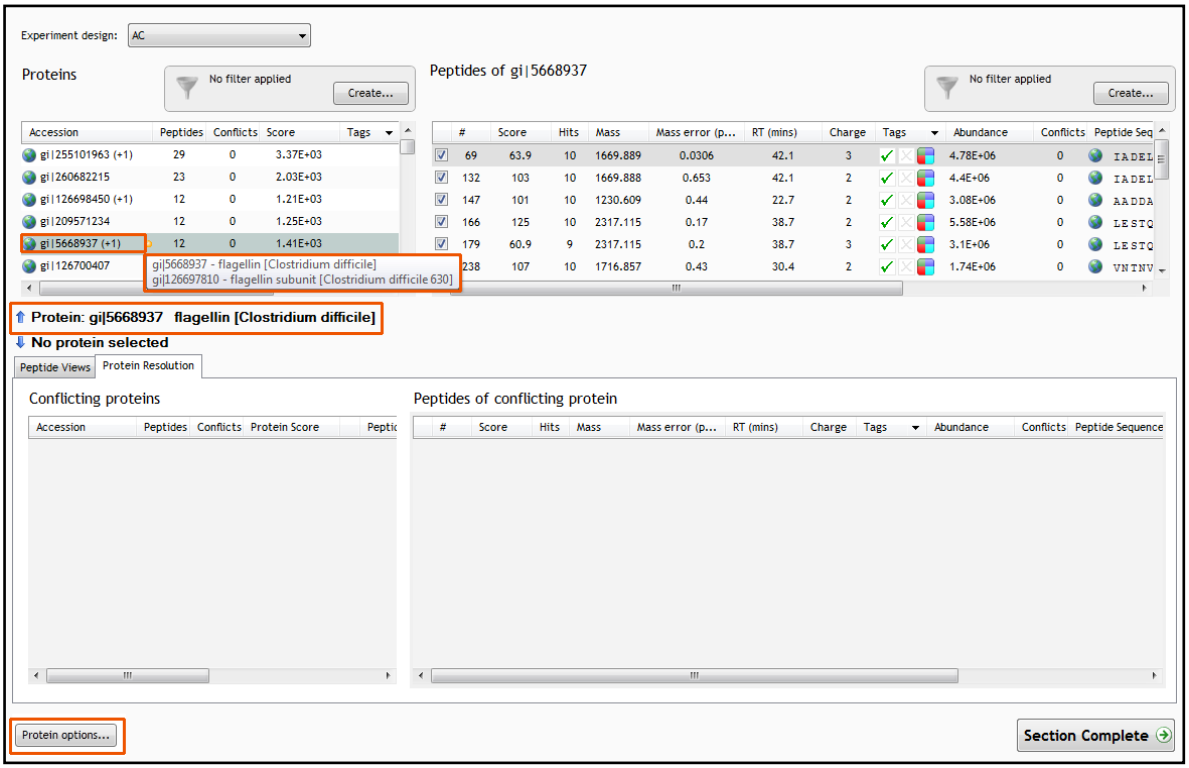
In the above example the conflict would be resolved in favour of the protein with 5 peptides.

**Note**: the number of conflicts you have to resolve will depend on the scope and stringency of the filters you apply at the **Refine Identifications** stage.

**Note**: the default **Protein options** for protein grouping and Protein quantitation are set as shown

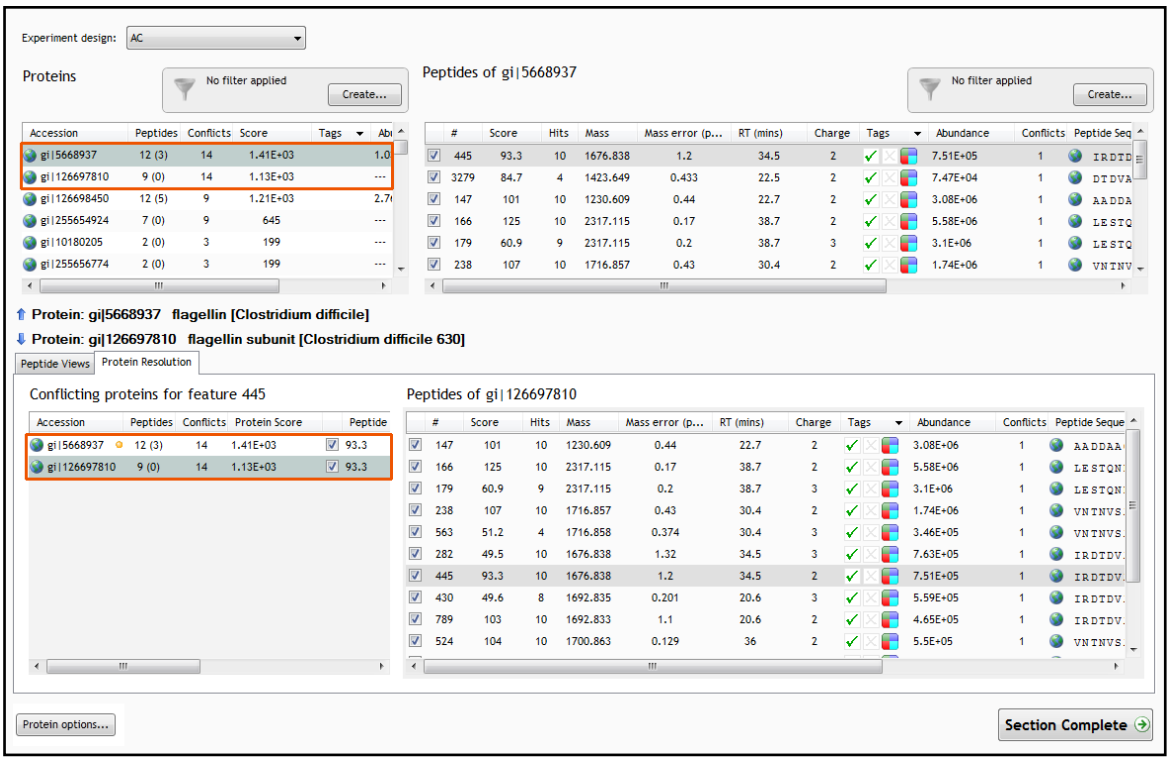



With protein grouping switched on protein groups and the additional members are indicated by a bracketed number located after the Accession number. Taking **flagellin** as an example, when the cursor is held over the accession number the group members appear in a tool tip.




Having performed the conflict resolution with **Group similar proteins** and **Quantitate from non-conflicting features** now switch off the protein grouping.

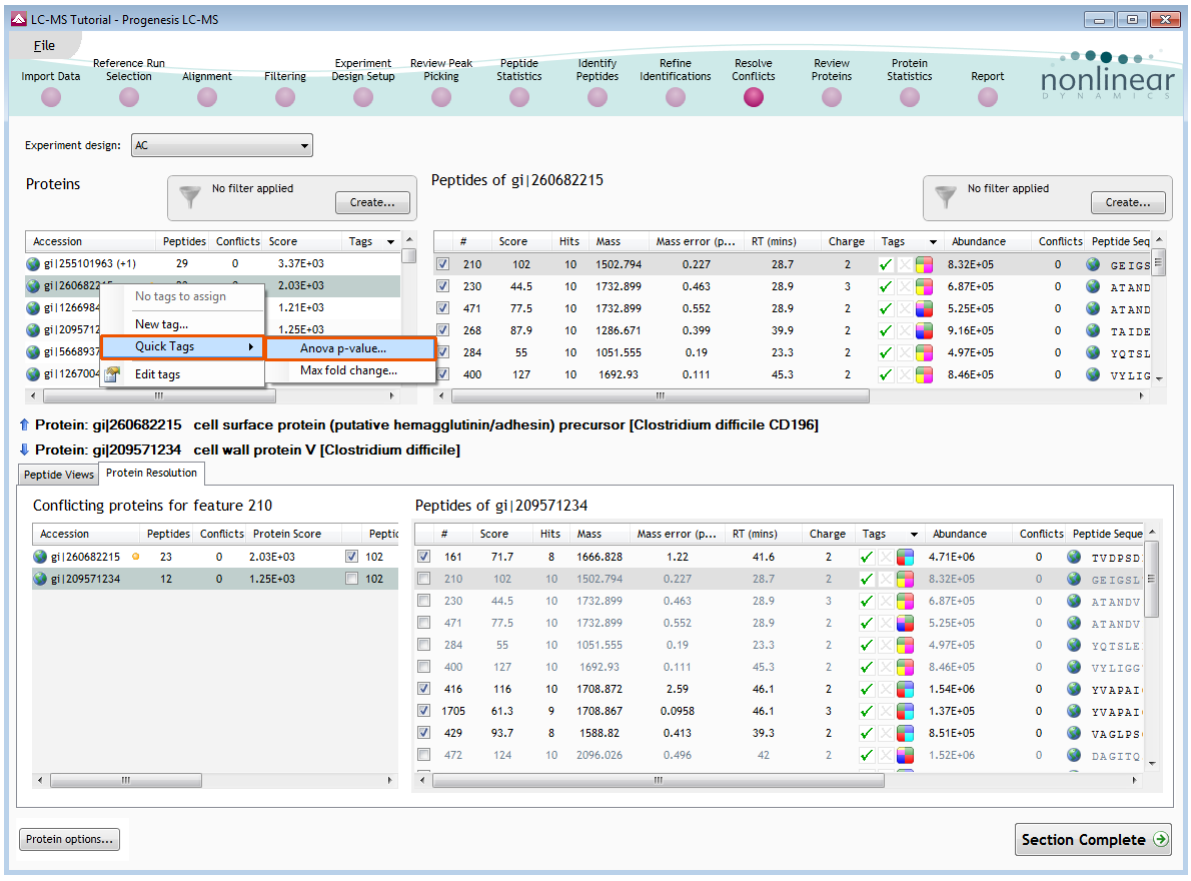
As grouping is switched off the grouped proteins appear with conflicts to the other group members and the number of unique peptides that are used for quantitation appear in brackets after the peptide number.



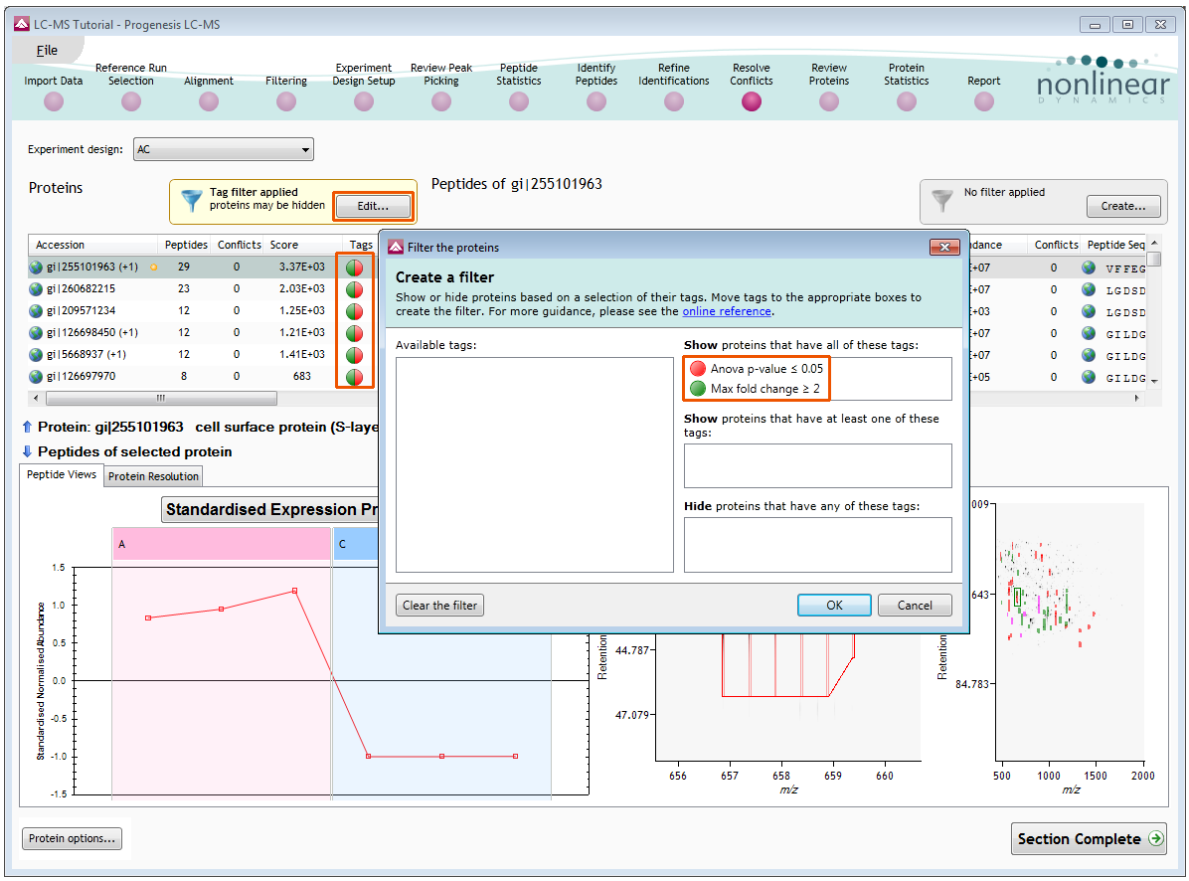

**Note**: the flagellin subunit has **no unique** peptides (brackets after the peptides field) as they are all present in flagellin protein hence the reason for grouping. As a result all the conflicts are internal to the group.

Now set the Protein Options back to **Group similar proteins**

Finally order the Protein table (A) on **Peptides**, and then right click on the table and create Quick Tags for proteins with **Anova p-value≤0.05** and a **Max fold change≥2** .

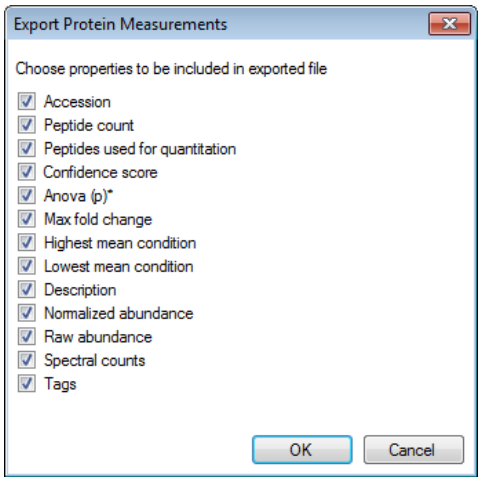


Having created these protein tags you can use them to filter the protein table so that it only displays those proteins that demonstrate a 2 fold or greater significant change.




This will filter the Protein list so that it now displays only the proteins that show significant increases in expression with a Maximum change of 2 fold or more.

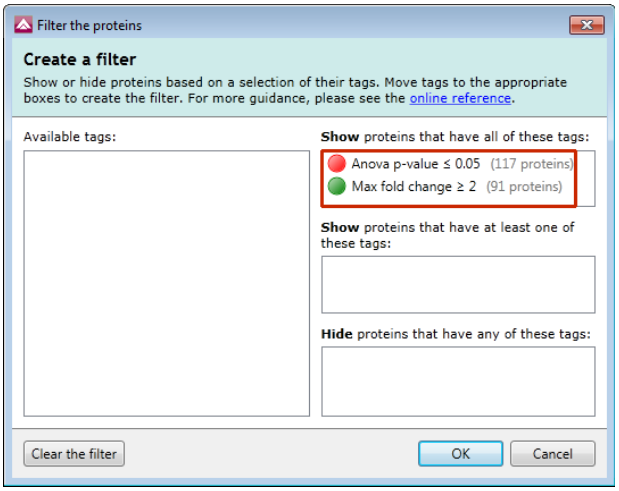
You can export this filtered Protein list (csv format) by selecting **Export Protein List** from the **File** menu. You can control the data output required, using the dialog provided.



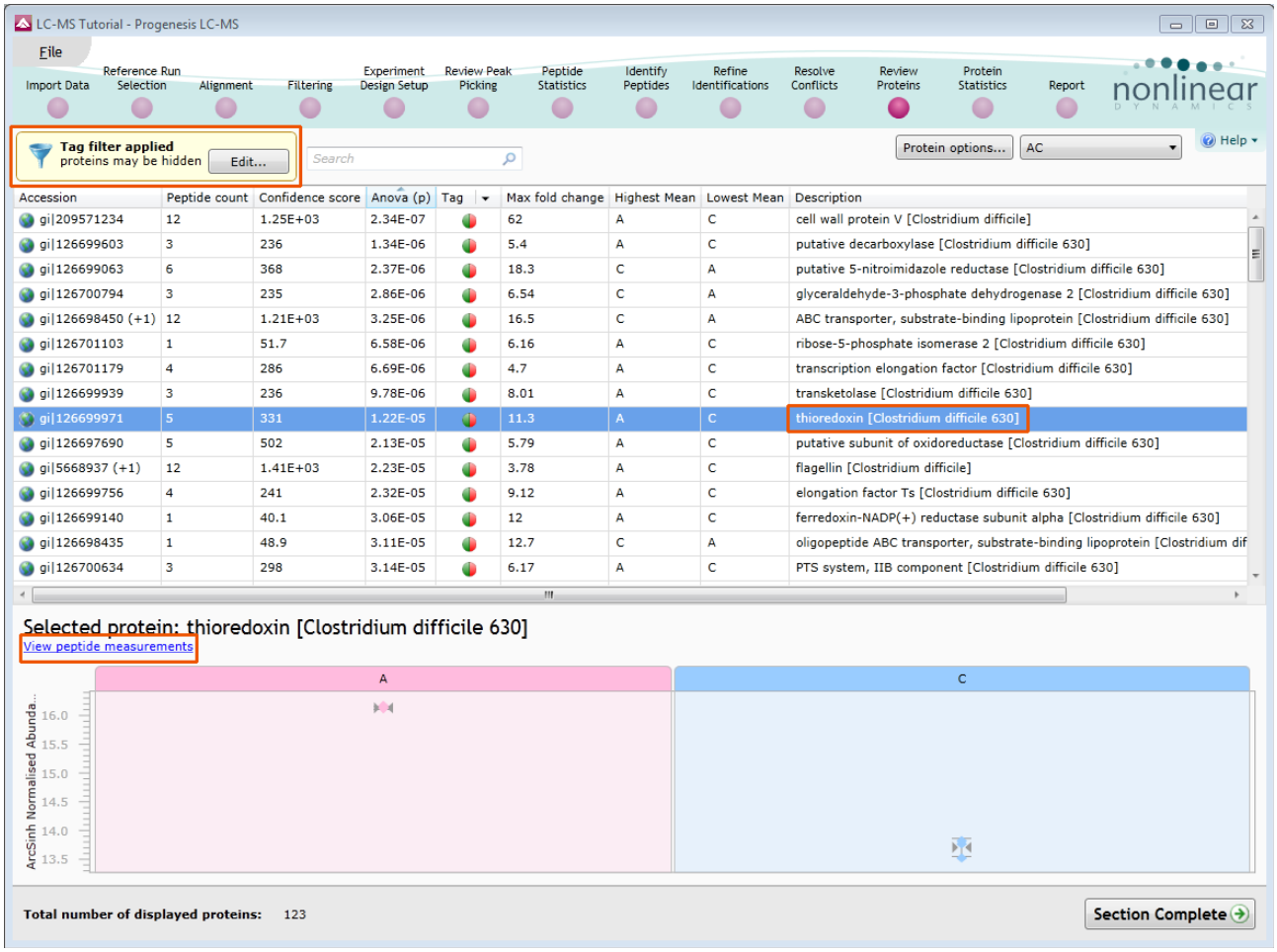

Now move to the Review Proteins section by clicking on **Review Proteins** icon on the workflow at the top of the screen.

# Stage 12: Review Proteins

The **Review Proteins** stage opens displaying those proteins that meet the conditions for the active protein filter, applied in the previous section, showing all the proteins with a significant fold change of 2 fold or greater.

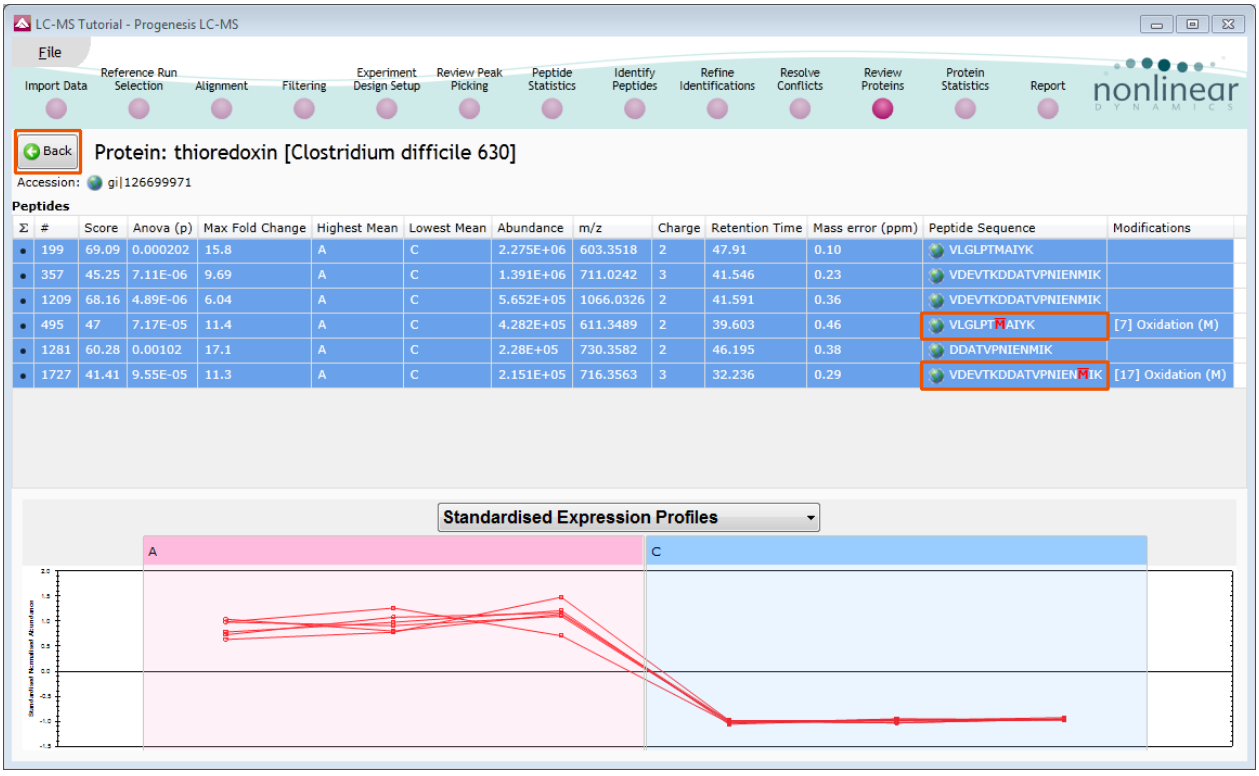



As an example let us explore Thirodoxin. The table indicates that this protein is most highly expressed in Condition A by 11.3 fold over the lowest condition (C) .

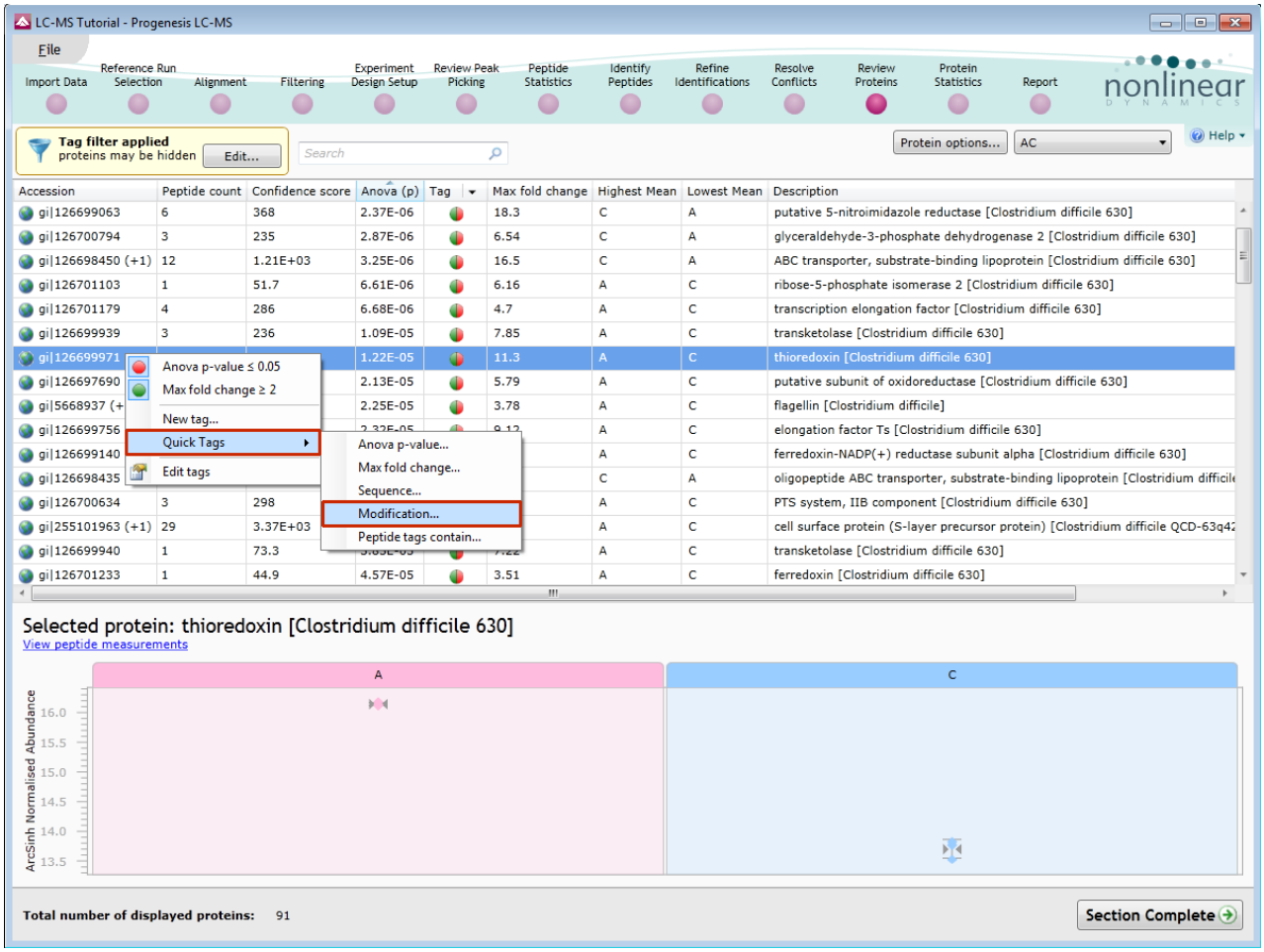



To view the corresponding peptide measurements for the current protein either double click on the protein in the table or use the **View peptide measurements** below the table.

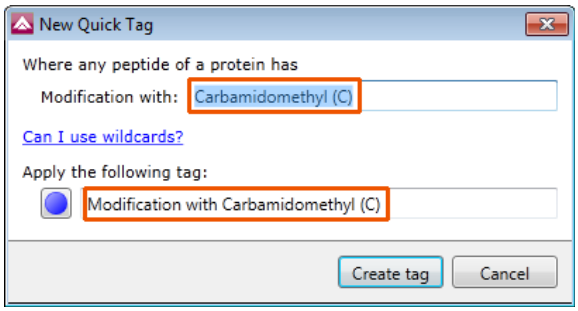
**Note**: by selecting all the peptides you can compare the pattern of expression across all the samples allowing you to identify 'atypical' behaviour of peptides assigned to the current protein.



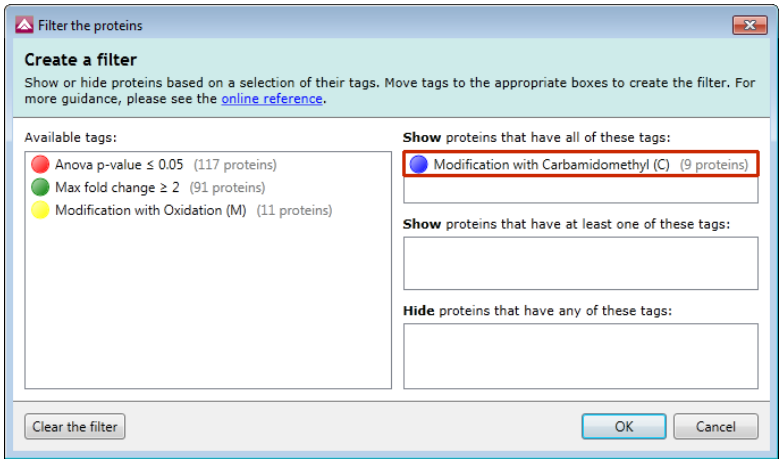

Modified proteins can be located by specifically searching for proteins containing modified peptides. Use the **Back** button to return to the Proteins List and right click on it and select **Modification** from the list of **Quick Tags.**



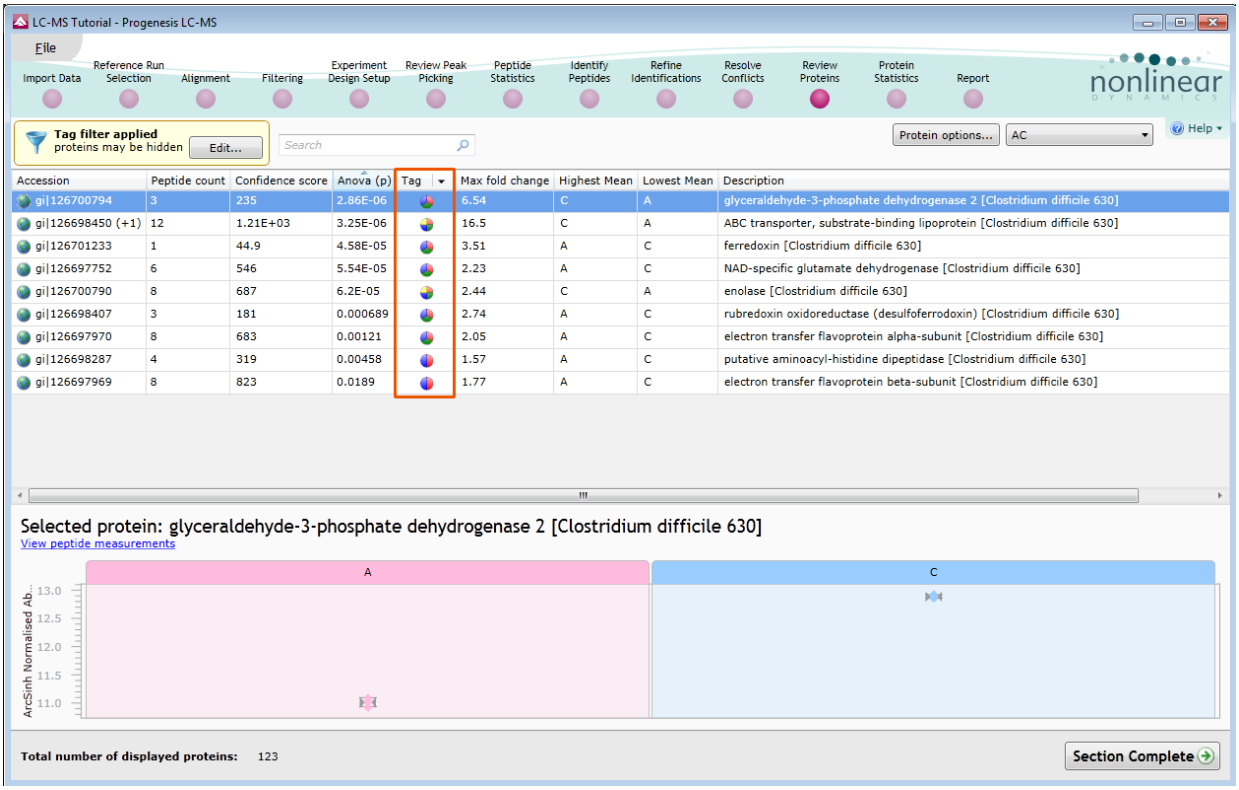
To find those proteins containing peptides with Carbamidomethyl on cysteine residues enter **Carbamidomethyl (C).** This will automatically provide a named tag when you click **Create tag**. Repeat this for Oxidation of methionine residues.




To reduce the table to displaying only these proteins with modified peptides (carbamidomethylation on cysteine) use the tag filter to focus on these proteins.




The proteins table will only display those proteins containing modified peptides.

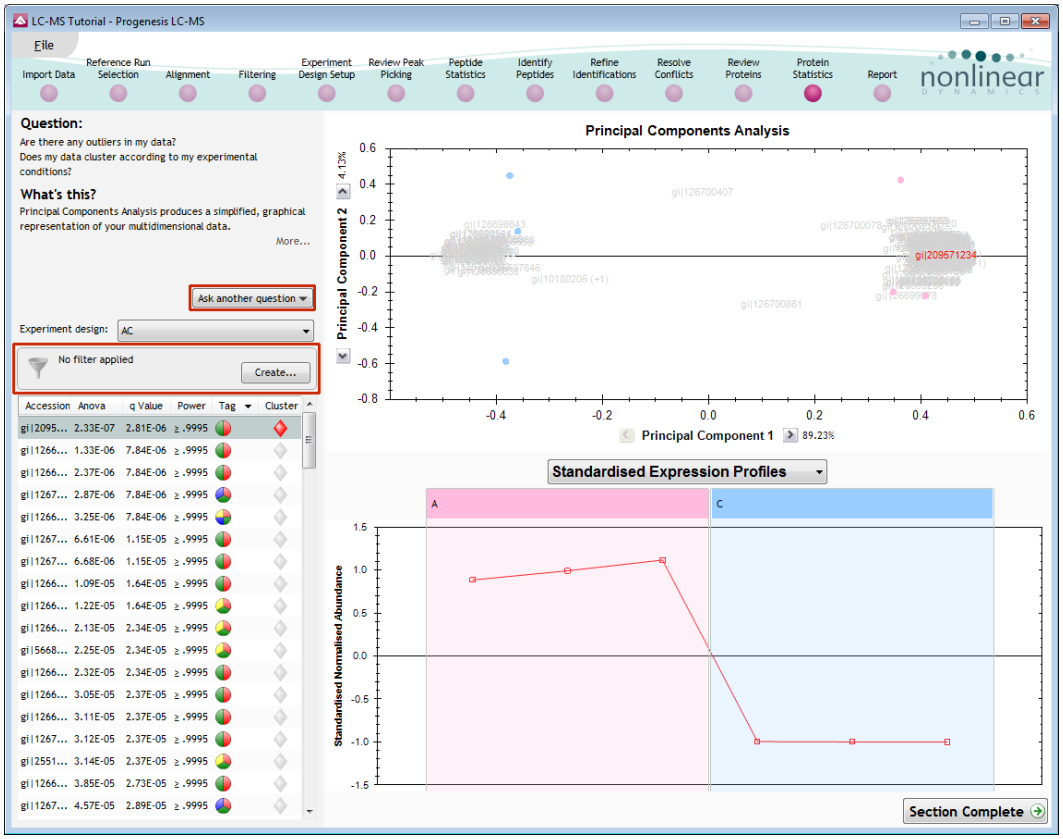



**Note**: the **Sequence** Quick tag can be used to locate Proteins containing peptides with specific motifs.

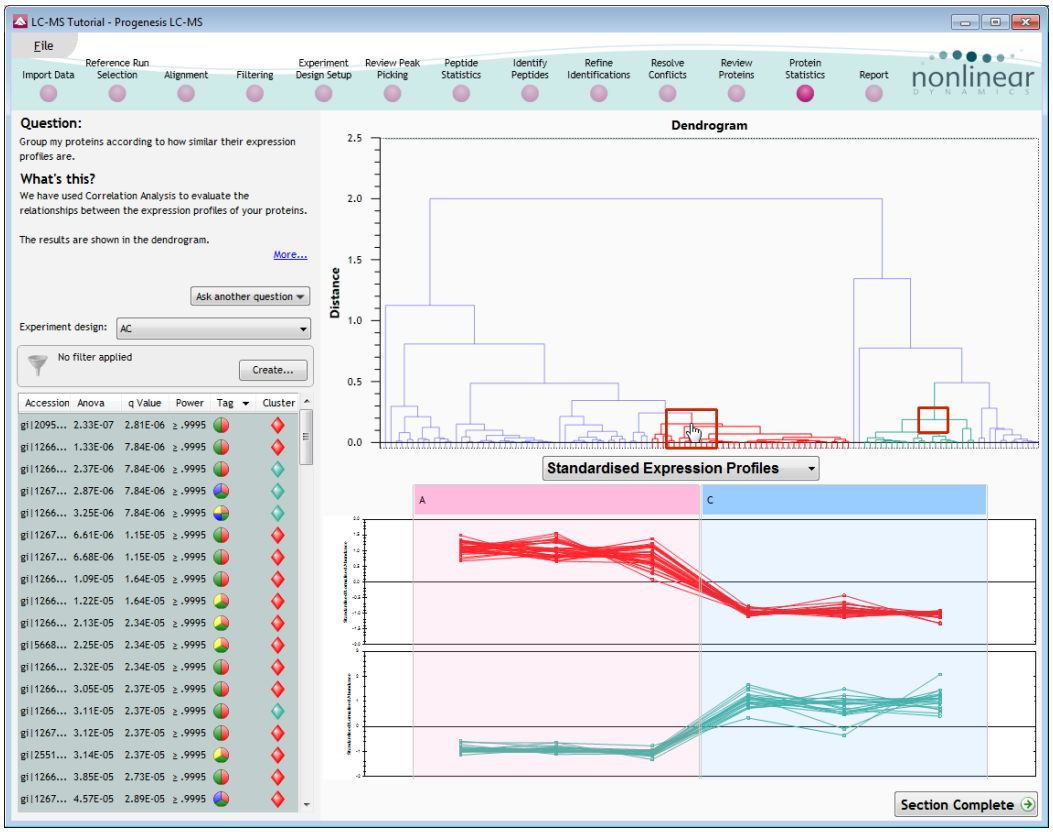
Now move to the Protein Statistics section by clicking on **Protein Statistics** icon on the workflow at the top of the screen.

# Stage 13: Protein Statistics

Protein Statistics opens with a Principal Components Analysis (PCA) for all the proteins displayed.



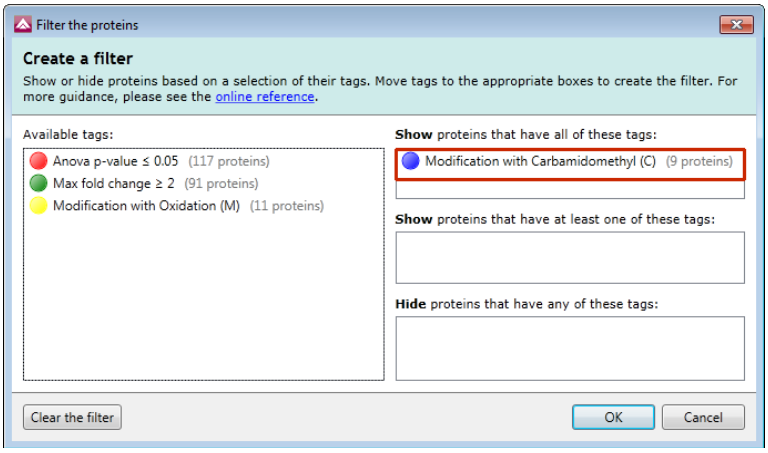

The Multivariate Stats can now be applied to all or subsets of proteins as determined by the current Tag filters. Allowing you to identify similar paterns of expression using the Correlation Analysis.




Now move to the **Report** section to report on Proteins and /or peptides.

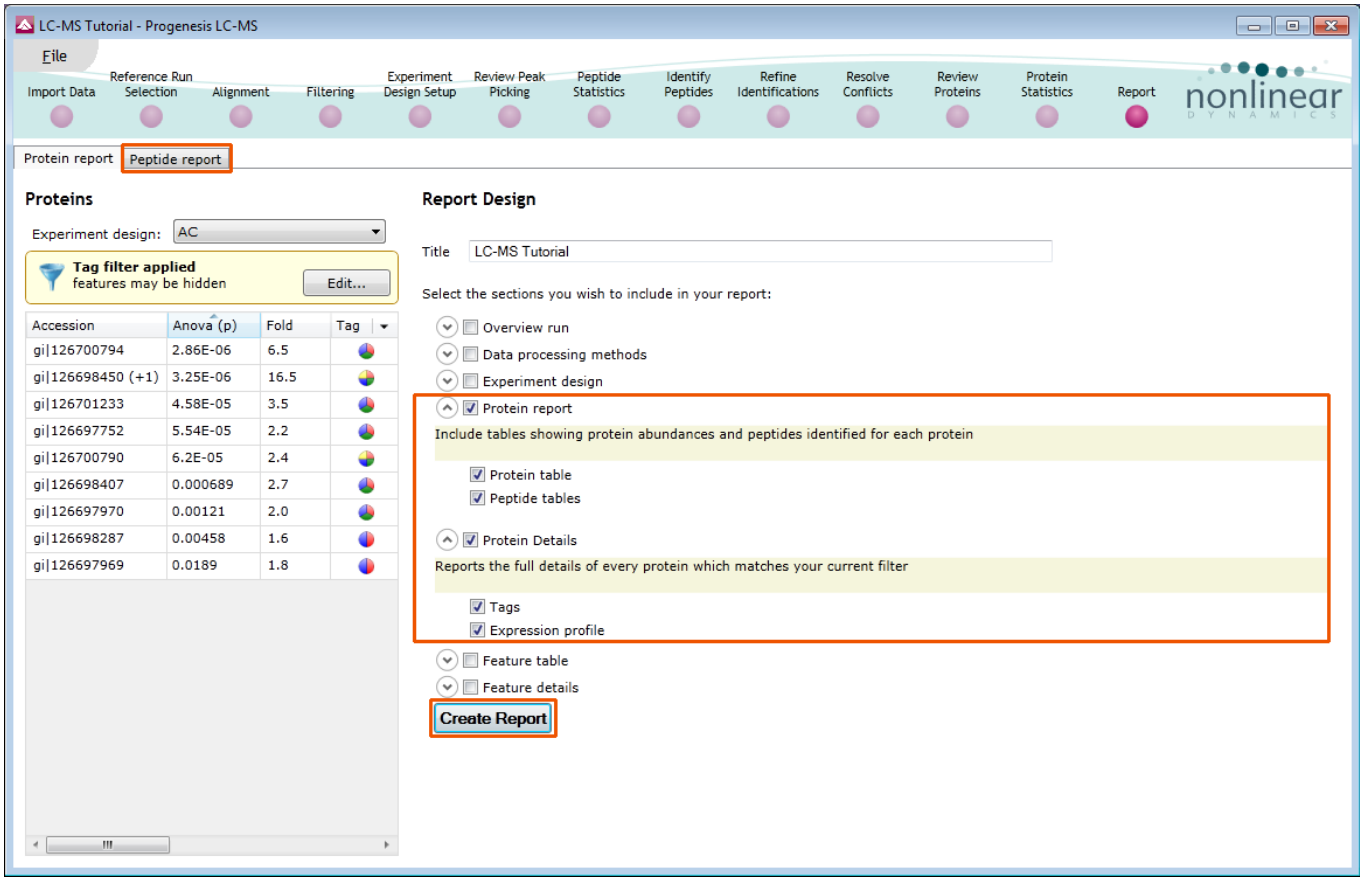
# Stage 14: Reporting

The **Report Design** stage allows you to select what views you want to include in a report based on the list of **currently selected proteins.**



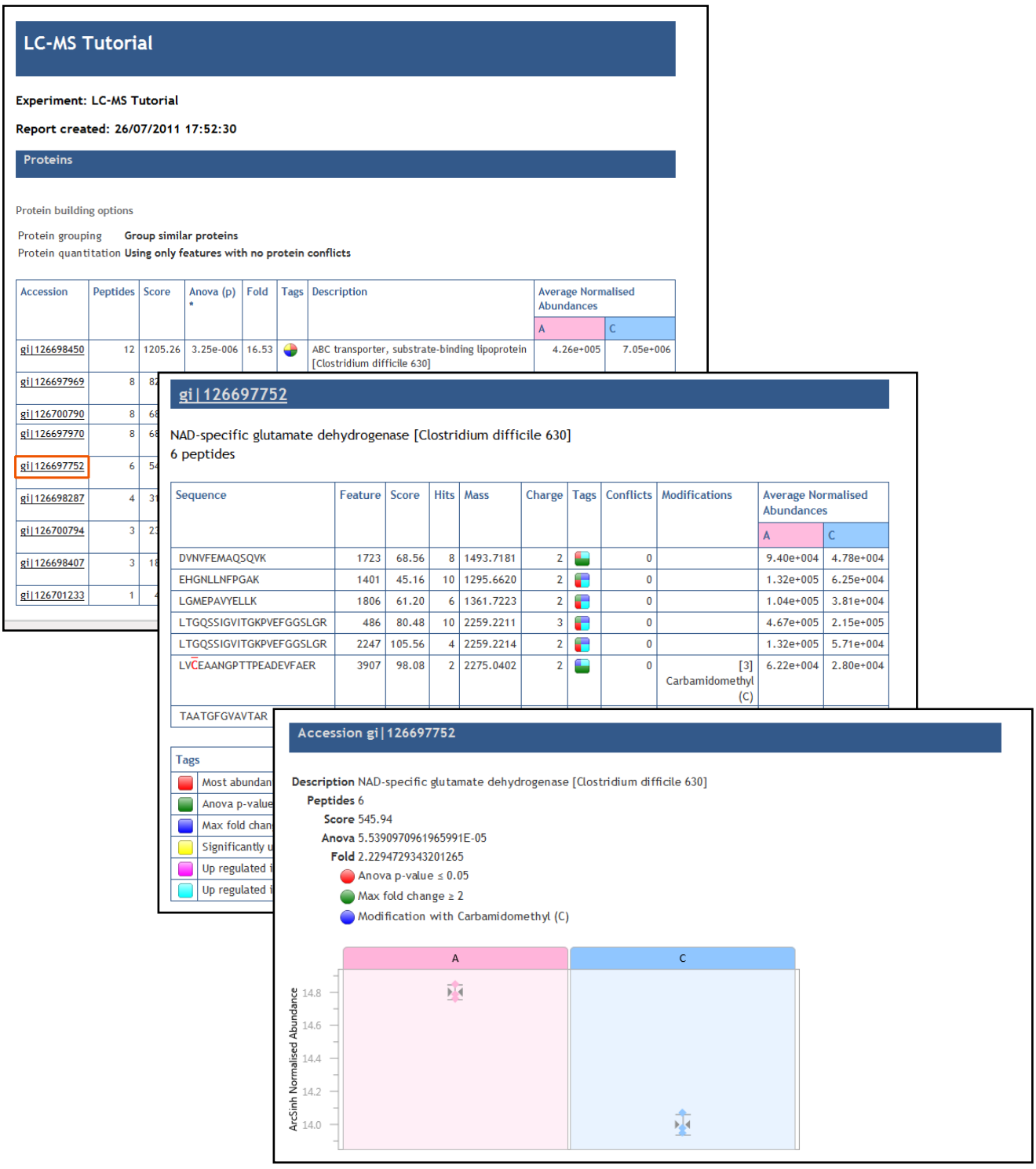

As an example we will create a report for **only** the proteins containing peptides with Carbamidomethyl on cysteine residues enter **Carbamidomethyl (C).**

1. First reduce the proteins to report on by selecting the '**Modification with Carbamidomethyl (C)**' tag. In this example it reduces the number of proteins in the table to 9.
2. Expand the various Report Design options (by default they are all selected)
3. Un-tick as shown below
4. Click **Create Report**




This opens a dialog to allow you to save the report, after which it will be opened in the form of a web page.

Click on the **Accession No**. in the proteins section of the Report and this will take you to the Assigned peptides for this protein




Having closed the report it can be reopened by double clicking on the saved html file.

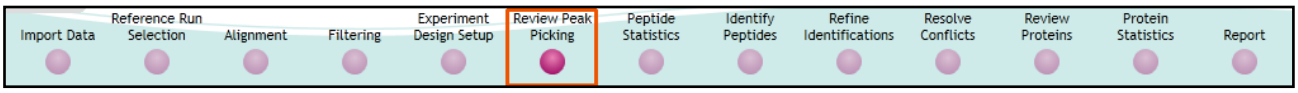
**Note:** you can also copy and paste all or selected sections of the report to Excel and/or Word.

**Note:** there are separate panels for reporting on Proteins and Peptides.

# Creating an Inclusion list

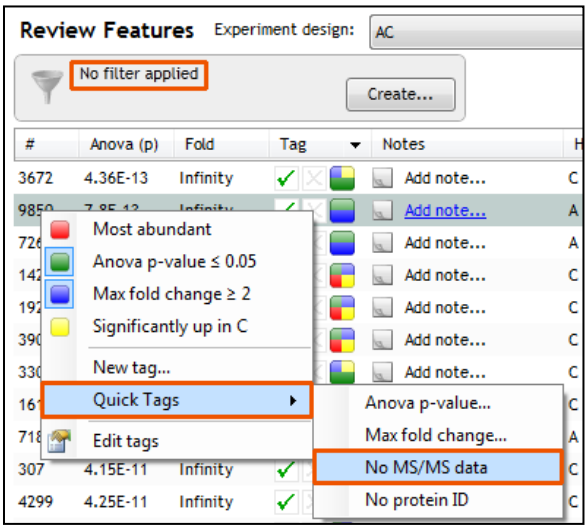
Inclusion lists can be used to try to increase the number of identified peptides you have in your experiment. They are used to control your mass spectrometer to try and concentrate the collection of MS/MS scans from specified m/z – retention time positions. Runs made using the inclusion list can then be imported into the Progenesis LC-MS and the extra MS/MS scans added to the experiment.

As an example we are going to create an inclusion list for all the features that show a Significant difference between groups A and C (Anova p<0.05) and have no MS/MS spectra.

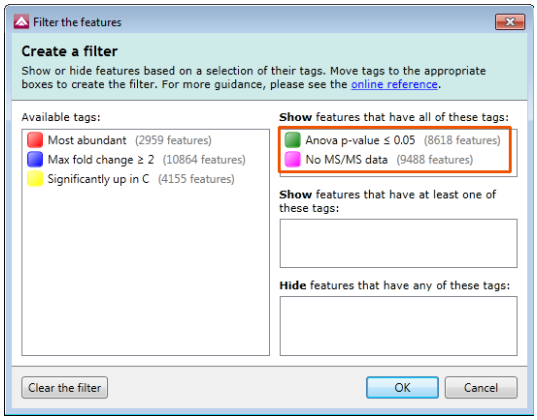



First return to **Review Peak Picking** using the Workflow icons.

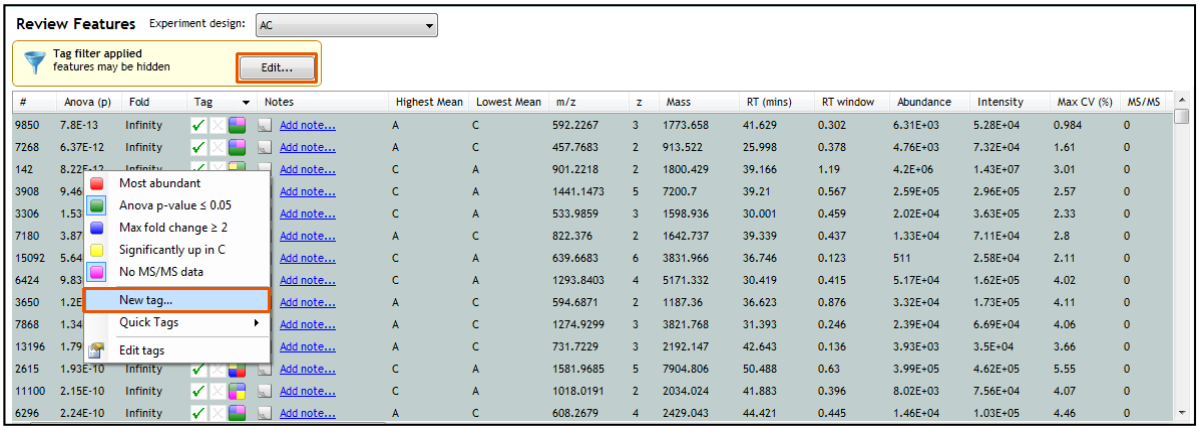
With no filters applied right click on a feature in the table, select **Quick tags** and click on **No MS/MS data**.



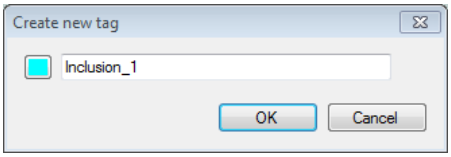

Filter the table so that it is only showing features with an **Anova p-value ≤0.05** and **No MS/MS data**. To do this click **Create** and drag the two tags on to the Show features that have all these tags and click OK.



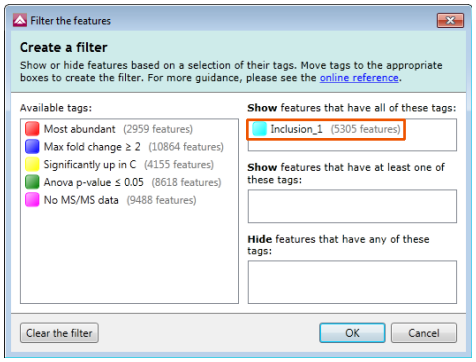

Select all of the features that are displayed, right click and create a New tag called **Inclusion 1**



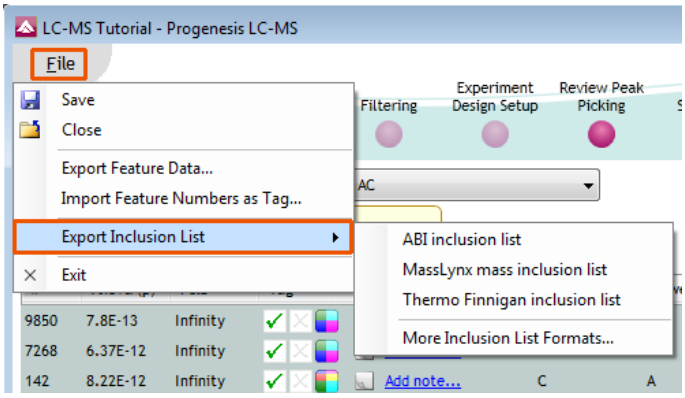

Call the new tag Inclusion_1.



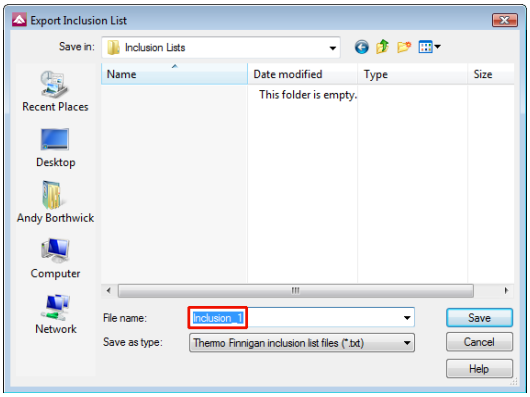
Now use the new tag to filter the table to display only those features that show a Significant Change and **DO NOT** have any MS/MS spectra.

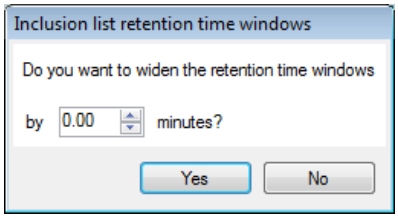



Then select **Export Inclusion List...** from the file menu and select the appropriate format.




Finally save the file to an appropriate location




**Note:** with certain MS machines it is possible to widen the retention time windows being used, this can be controlled using the following dialog.

If you require further information on the inclusion list file formats then click **Help.**

**Note**: The new LC-MS runs can then be added to the original experiment to increase the MS/MS coverage using the **Add files** facility at the Data Import Stage.

## Congratulations!

This document has taken you through a complete analysis using Progenesis LC-MS, from Alignment through Analysis to generating lists of interesting features using powerful Multivariate Statistical analysis of the data.

Hopefully our example has shown you how this unique technology can deliver significant benefits with

- Speed
- Objectivity
- Statistical Power

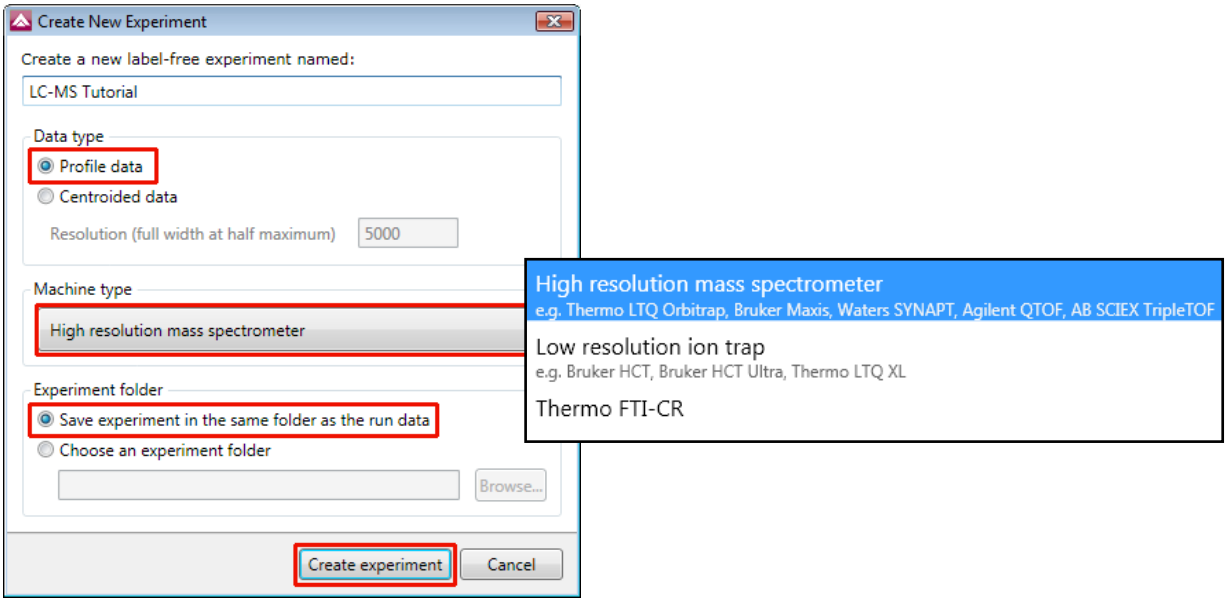
If you would like to see the benefits of running Progenesis SameSpots using your own runs and explore the Progenesis LC-MS workflow please go to Appendix 3: Licensing Runs (page 58).

# Appendix 1: Stage 1 Data Import and QC review of LC-MS data set

You can use your own data files, either by directly loading the raw files (Thermo and Waters) or, for other Vendors, convert them to mzXML or mzML format first.

To create a new experiment with your files select **New** give your experiment a name. Then select data type, the default is 'Profile data'.

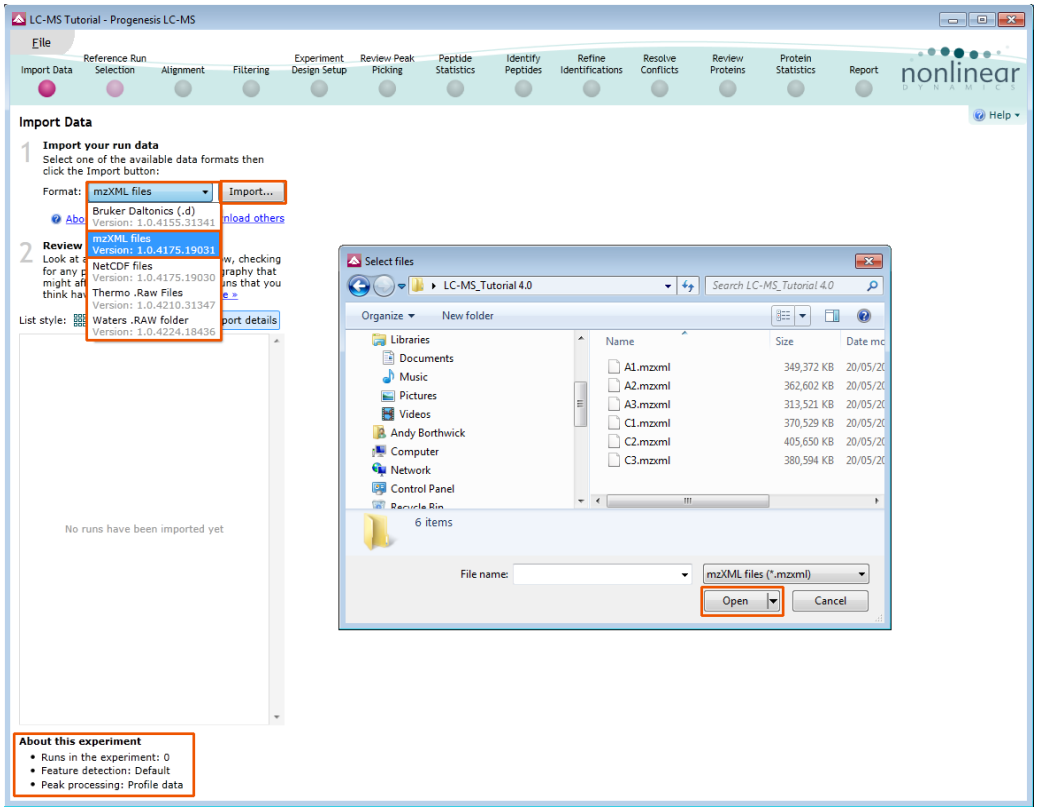
**Note**: if you have converted or captured the data as centroided then select Centroided data and enter the Resolution for the MS machine used.




Click **Create experiment** to open the LC-MS Data Import stage of the workflow.

Select the 'Import Data file format', in this example they are mzXML files

Then locate your data files using **Import...**

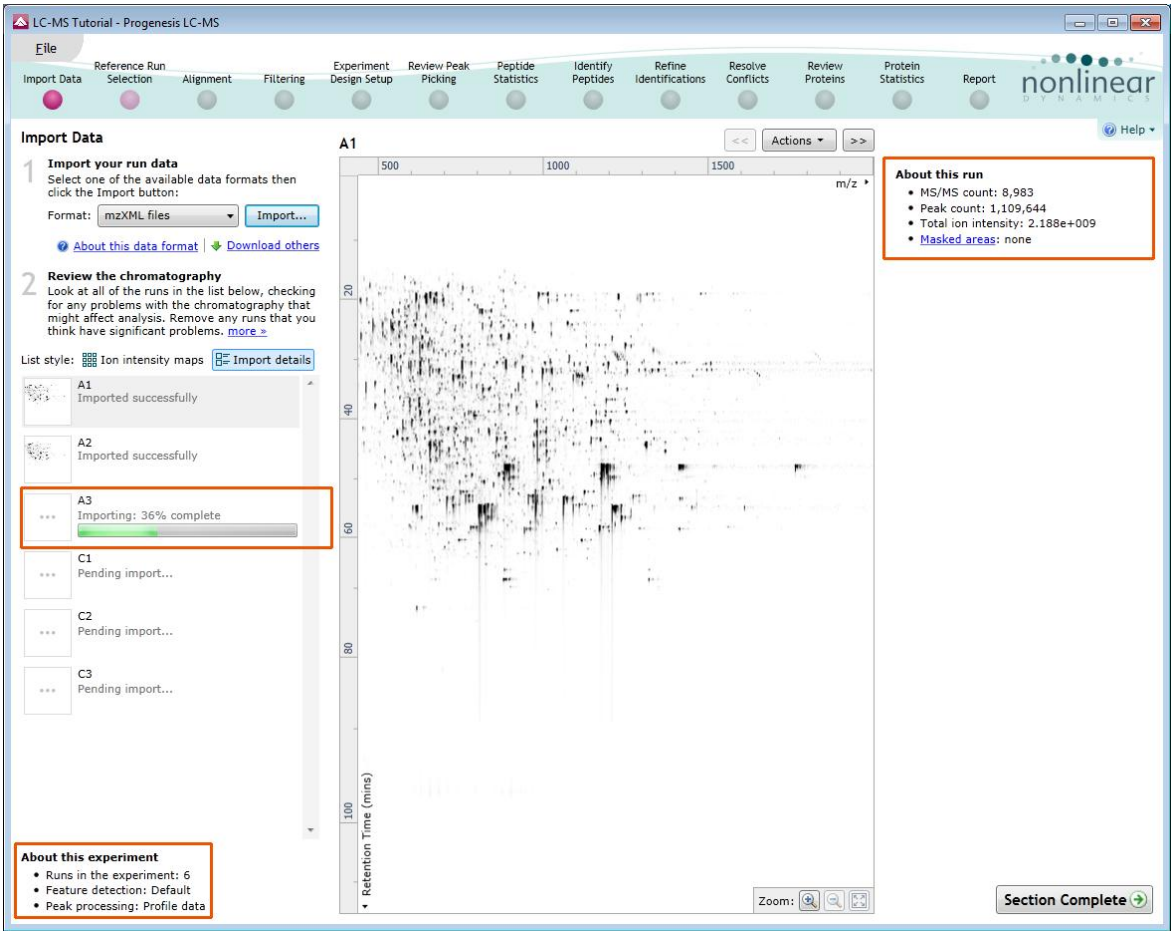



Locate and select all the Data files (A1 to C3).

On loading the selected runs your data set will be automatically examined and the size of each file will be reduced by a 'data reduction routine', which reduces the data by several orders of magnitude but still retains all the relevant quantitation and positional information.

**Note**: For a large number of files this may take some time.

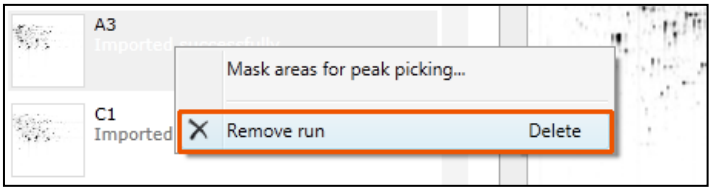
Each data file appears as a 2D representation of the run. At this stage you will be warned if any of the data files have been 'centroided' during the data acquisition and conversion process.




**Note**: as each data file is loaded the progress is reported in the **Import Data** list. The dialog below the run reports on the QC of the imported Data files. In this case 'No problems found' with the this data file

**Note**: details of the current run appear on the top right of the view.

You can delete run(s) by left clicking on the run in the list.



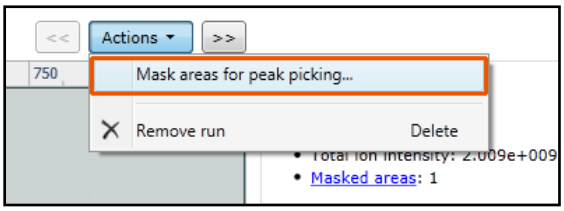

Now move to the next stage in the workflow (page 6 in this user guide) by clicking **Section Complete.**

# Appendix 2: Stage 1 Data QC review and addition of exclusion areas

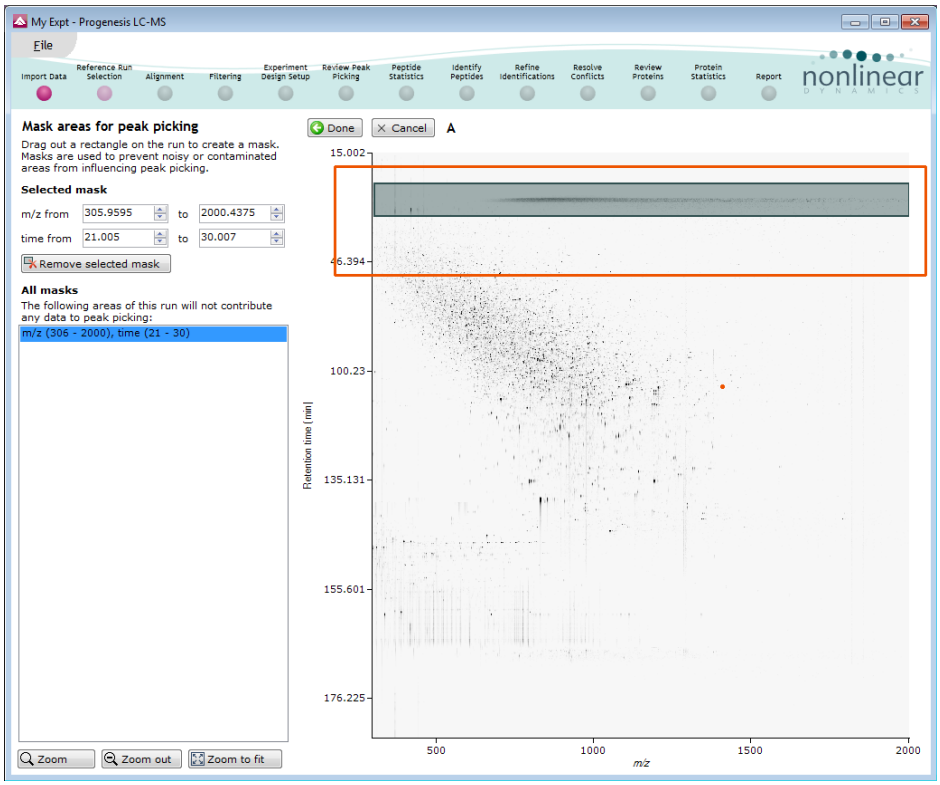
During the process of Data QC you may identify areas of the raw data for a particular run that appear 'noisy' yet still have identifyable 'isotopic patterns'.

For example if the run is part of a 'replicate set' of runs it is possible to exclude such areas on the noisy run by applying a mask to the area. By doing so this area is excluded during the initial part of the detection process in order that it does not 'interfere' with the detection of the features in the replicate group.

To do this select Masked areas from selected run on the bottom left of the screen.

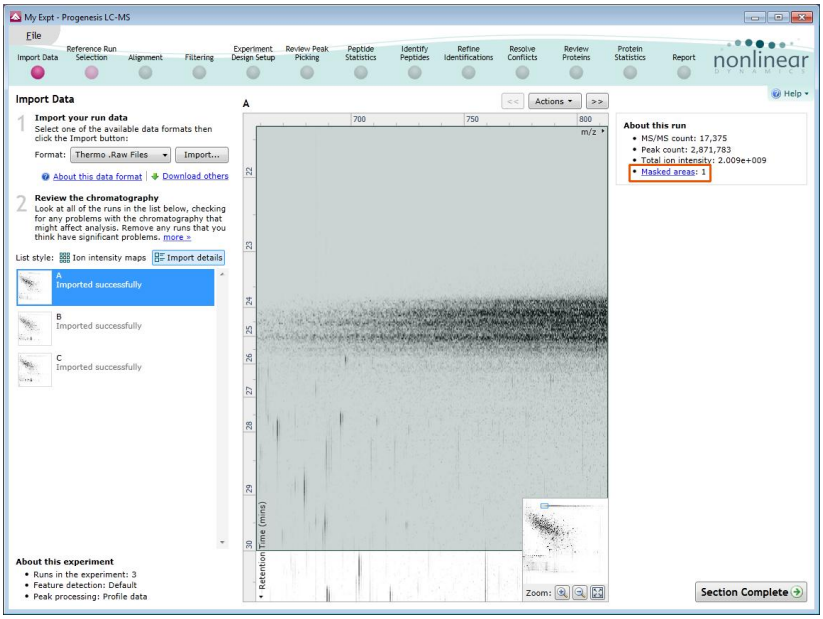



Drag out an area over the noisy part of the run to create the mask.



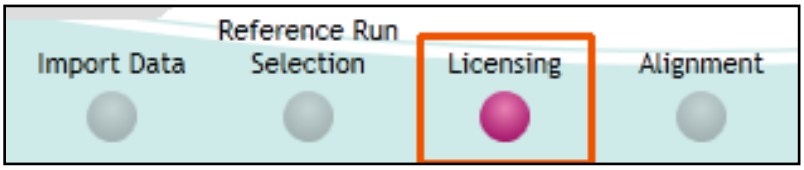

**Note**: Click **Done** to return to the **Import Data** view where you can zoom into the masked where you will see the isotopic features in the noise.

**Note**: if the level of noise is high and affecting many of your runs a preferred approach would be to re-optimise the chromatography to improve the levels of noise in your data.



# Appendix 3: Licensing runs (Stage 3)

When setting up a **New experiment** if you are evaluating Progenesis LC-MS with unlicensed runs then the licensing page will open after **Reference Run Selection**.




**If you already have a programmed dongle attached to your machine then the License Runs page will not appear.**

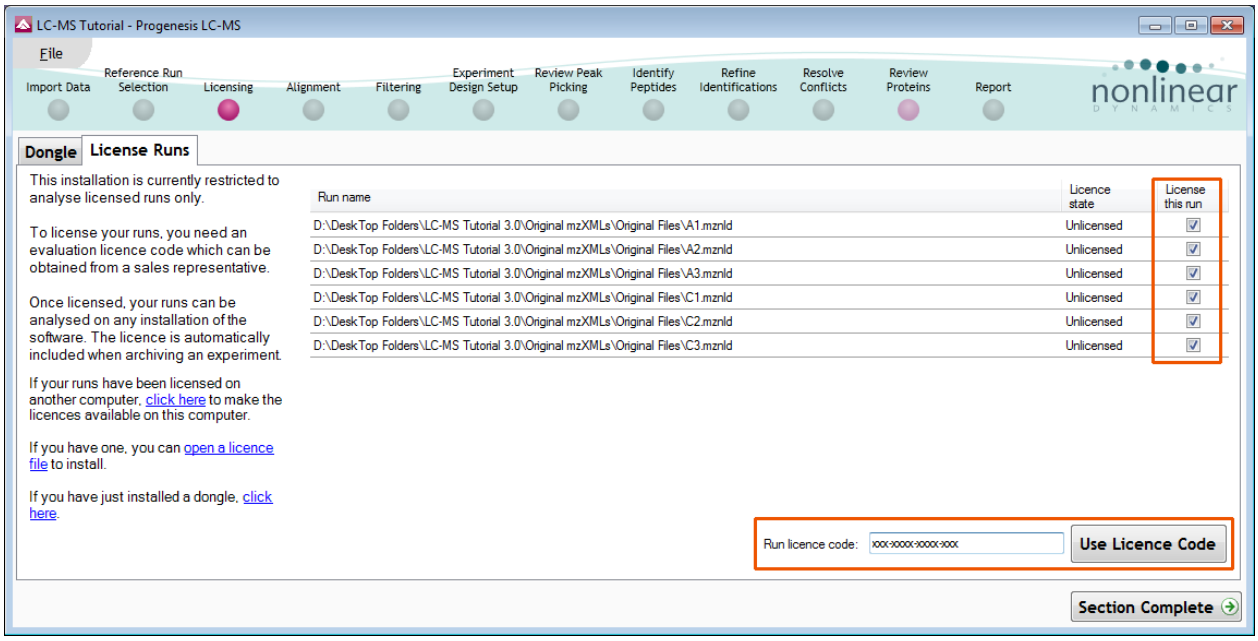
To use this page to License your Runs you must first either obtain an **'Evaluation' Licence Code from a Nonlinear Sales Person or purchase a licence code directly from Nonlinear**.

Each code will allow you to license a set number of runs.

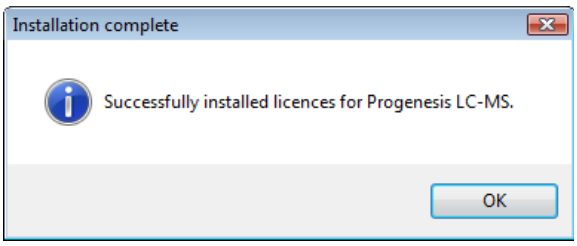
The runs in your experiment will be listed as shown below.

To activate license(s) for the selected runs enter the code in the space provided and click **Use Licence code**.

**Note**: you will need an internet connection to use this method.




A message confirming successful installation of your licences will appear.



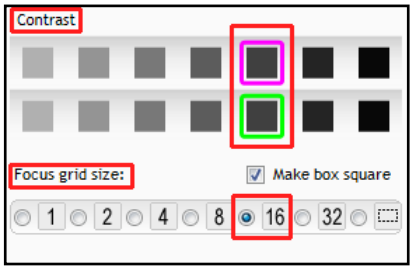

Click **OK**, the view will update and Alignment, the next stage in the workflow, will open with the licensed files.

# Appendix 4: Manual assistance of Alignment

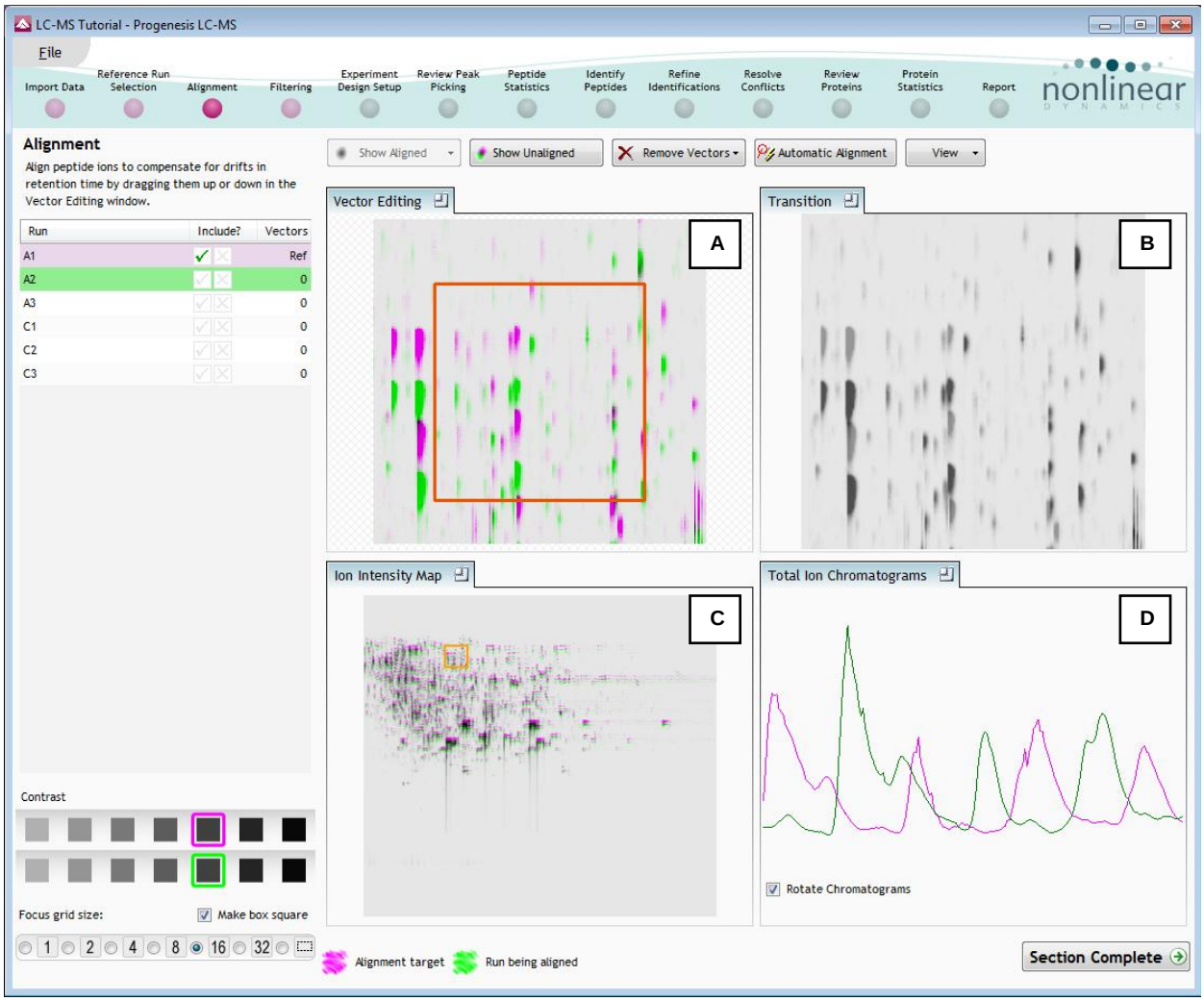
## Approach to alignment

To place manual alignment vectors on a run (A2 in this example):

1. Click on Run A2 in the **Runs** panel, this will be highlighted in green and the reference run (A1) will be highlighted in magenta.
2. You will need approximately 5 - 10 **alignment vectors** evenly distributed from top to bottom of the whole run.
3. First ensure that the size of the focus area is set to **8 or 16** in the Focus grid size on the bottom left of the screen.



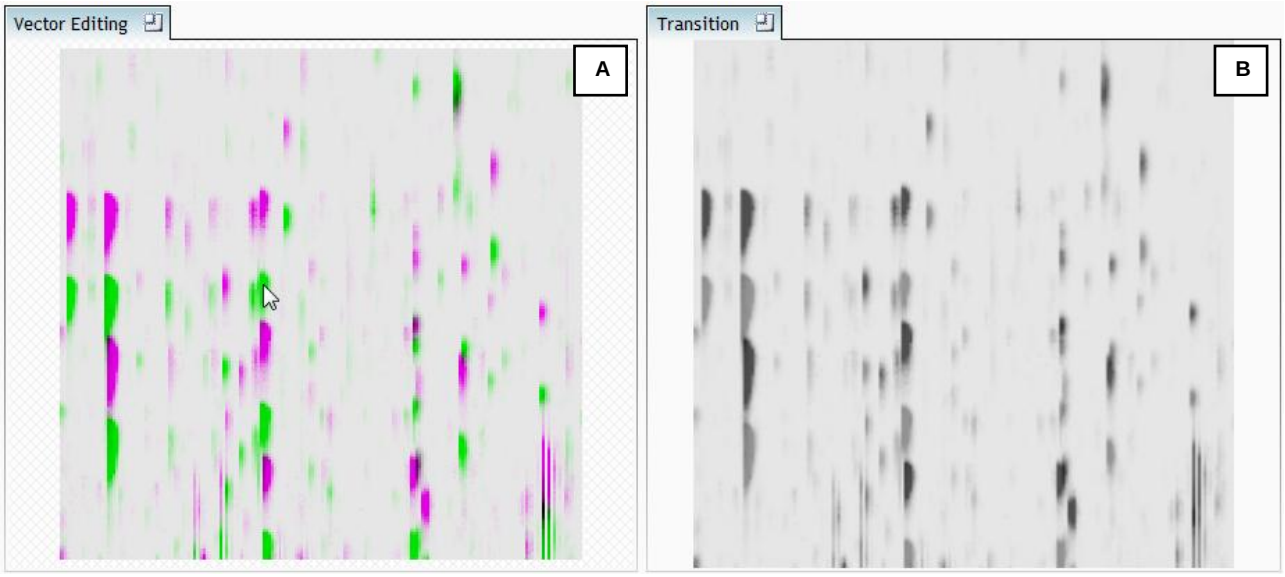

Click on an area (see below) in the **Whole Run** view (C) to refocus all the windows. Adjust Contrast as required.



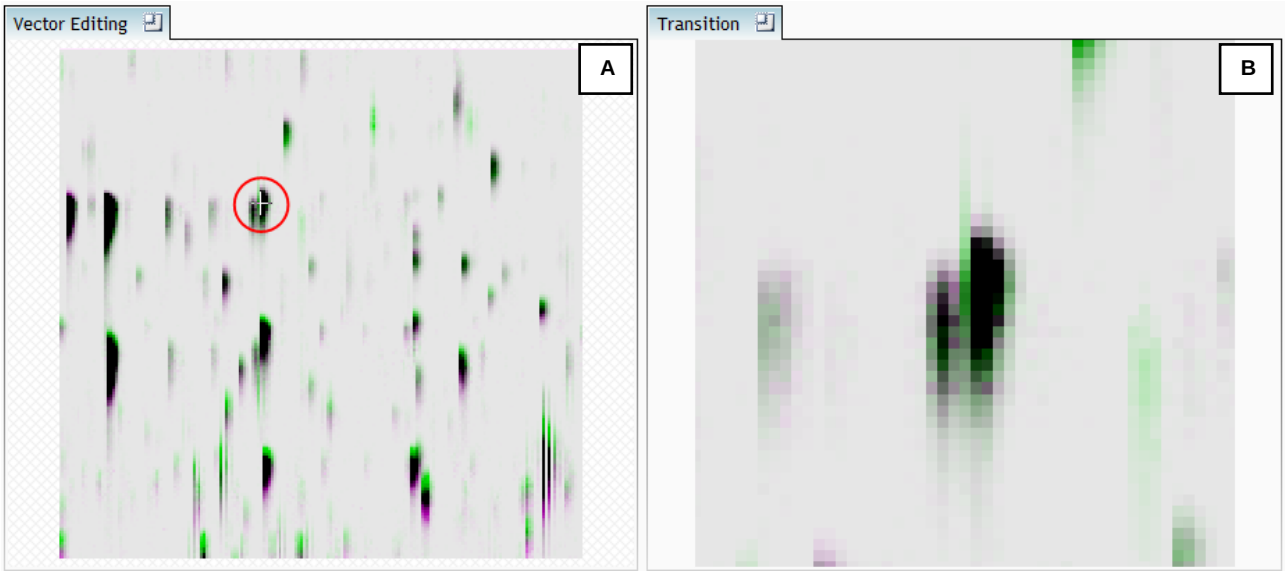

**Note**: the features moving back and forwards between the 2 runs in the **Transition** view (B) indicating the misalignment of the two LC-MS runs

**Note**: The **Total Ion Chromatogram (D)** view also reflects the misalignment of the 2 runs for the current Retention Time range (vertical dimension of the current Focus grid in the **Whole Run (C)** view.

4. Click and hold on a green feature in Window A as shown below.

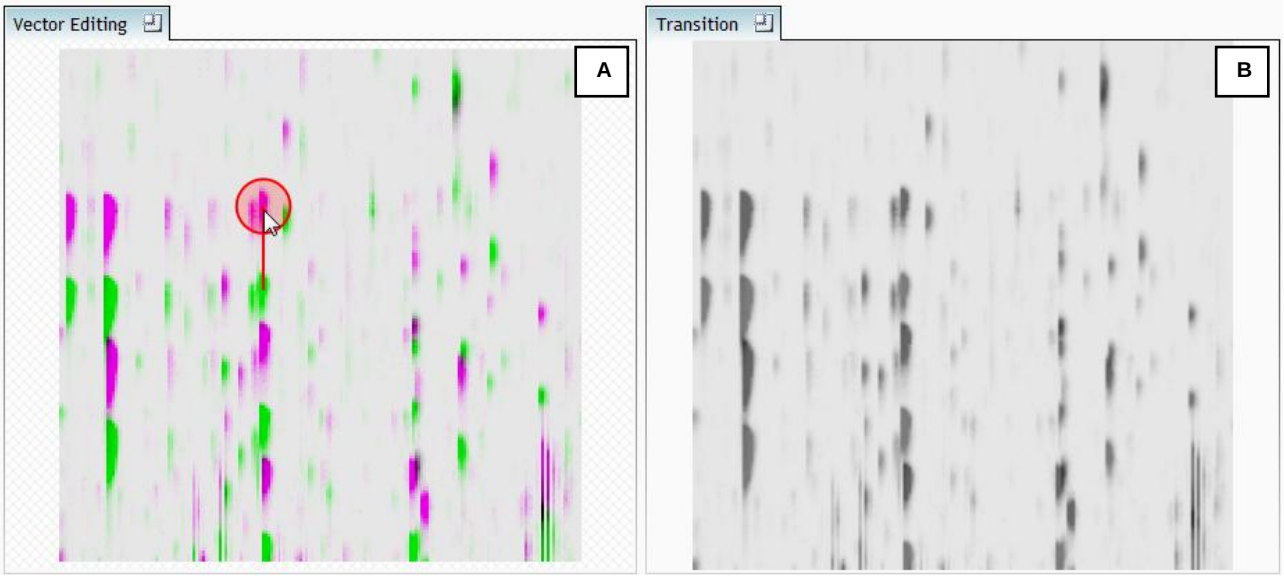



5. As you are holding down the left mouse button drag the green feature over the corresponding magenta feature of the reference run. The red circle will appear as shown below indicating that a positional lock has been found for the overlapping features.



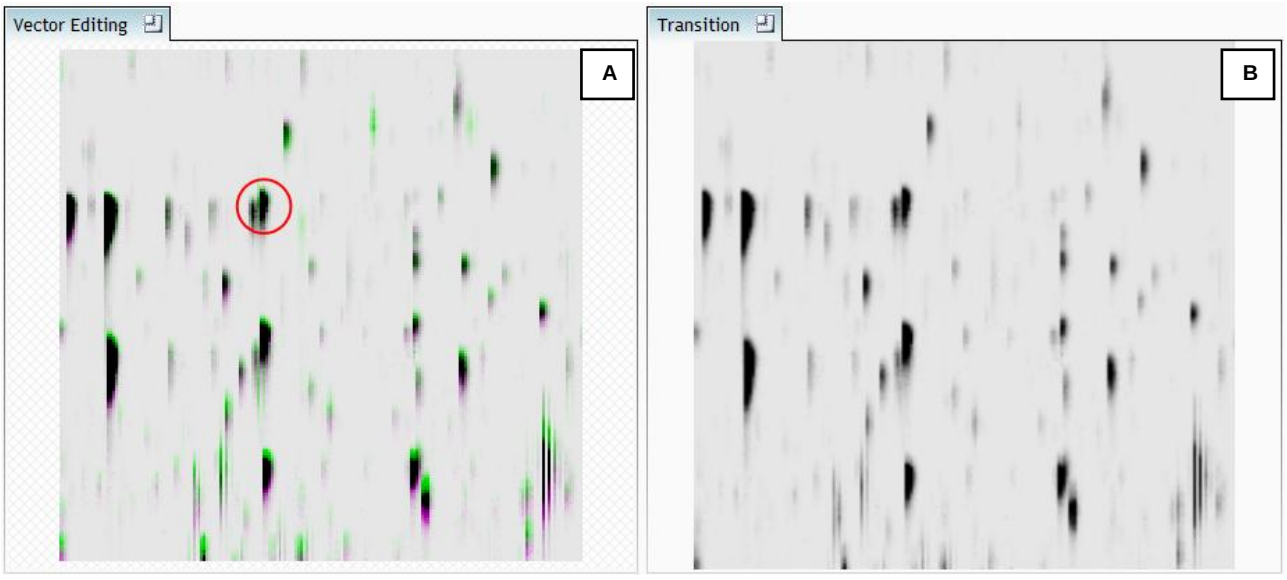

**Note**: as you hold down the mouse button, window B zooms in to help with the alignment.

6. On releasing the left mouse button the view will 'bounce' back and a red vector, starting in the green feature and finishing in the magenta feature will appear.

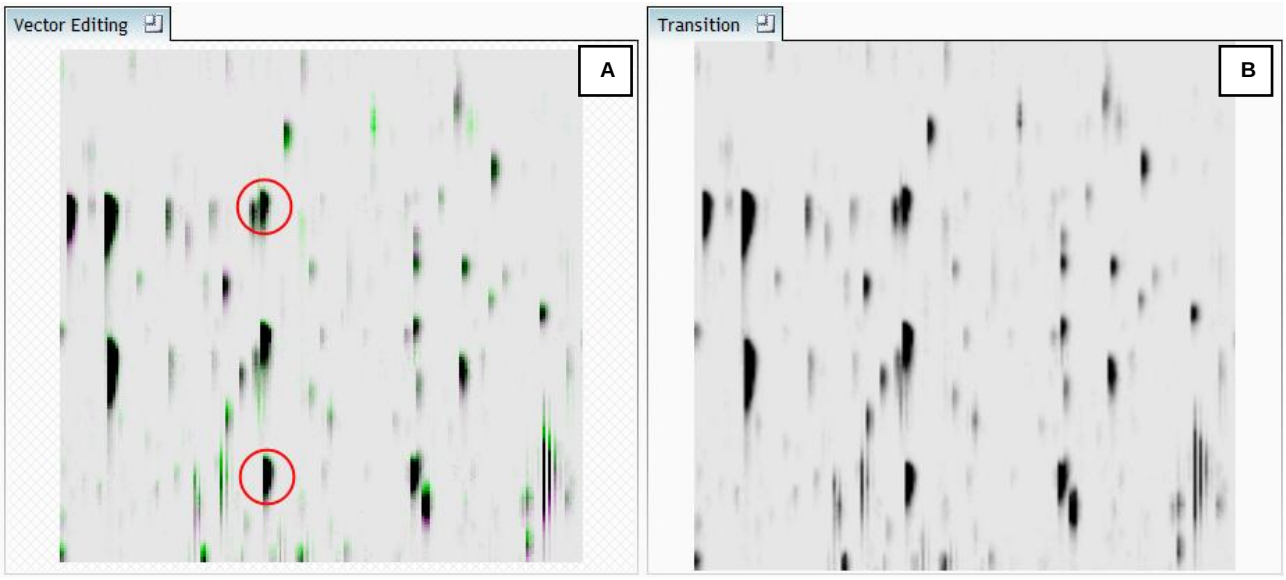



**Note**: an incorrectly placed vector is removed by right clicking on it in the **Vector Editing** window

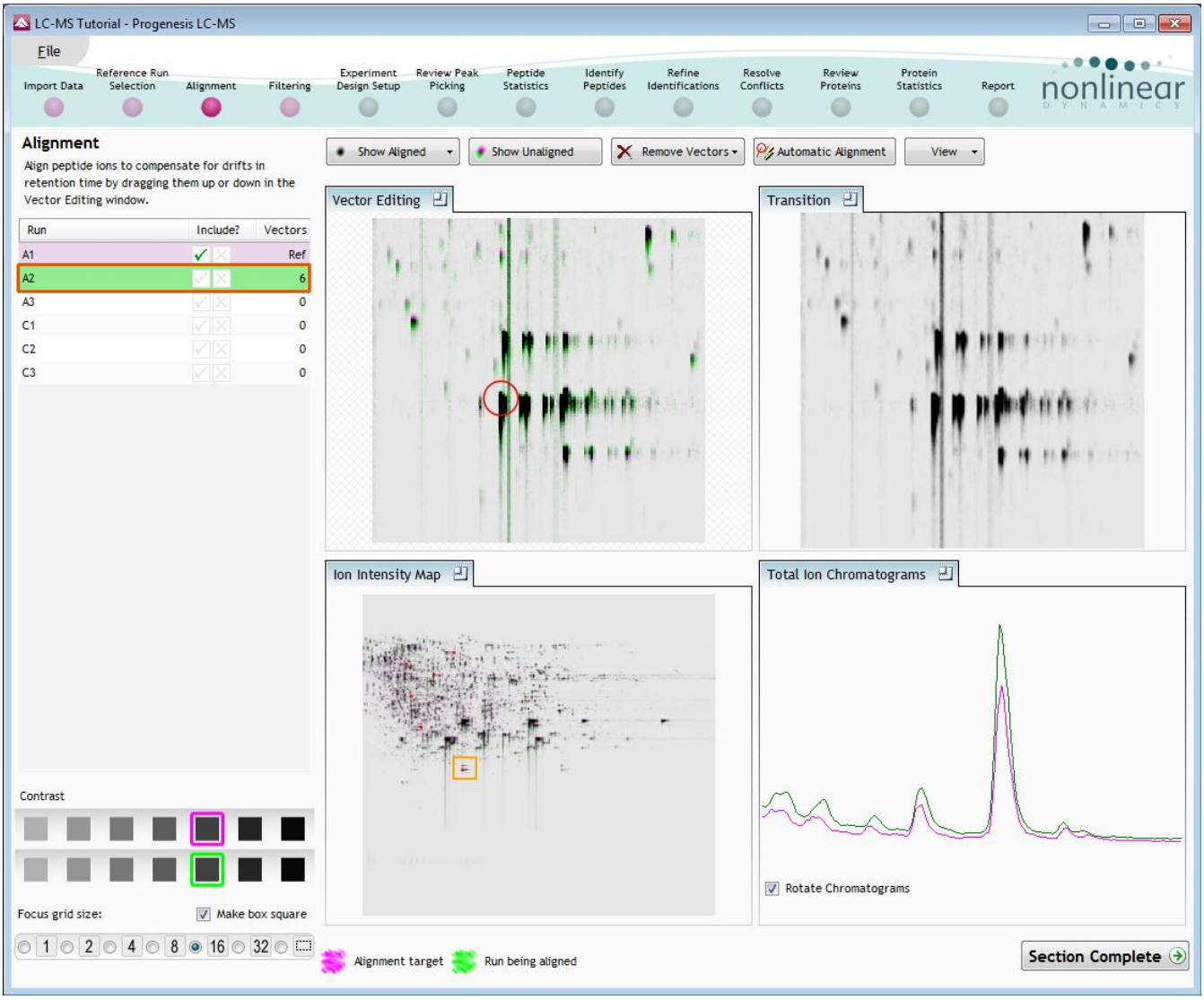
7. Now click **Show Aligned** on the top tool bar to see the effect of adding a single vector.




8. Additing an additional vector will improve the alignment further. **Note** this time as you click to add the vector it 'jumps' automatically to the correct position using the information from the existing alignment vector.



Repeat this process moving the focus from top to bottom on the **Whole Run** view

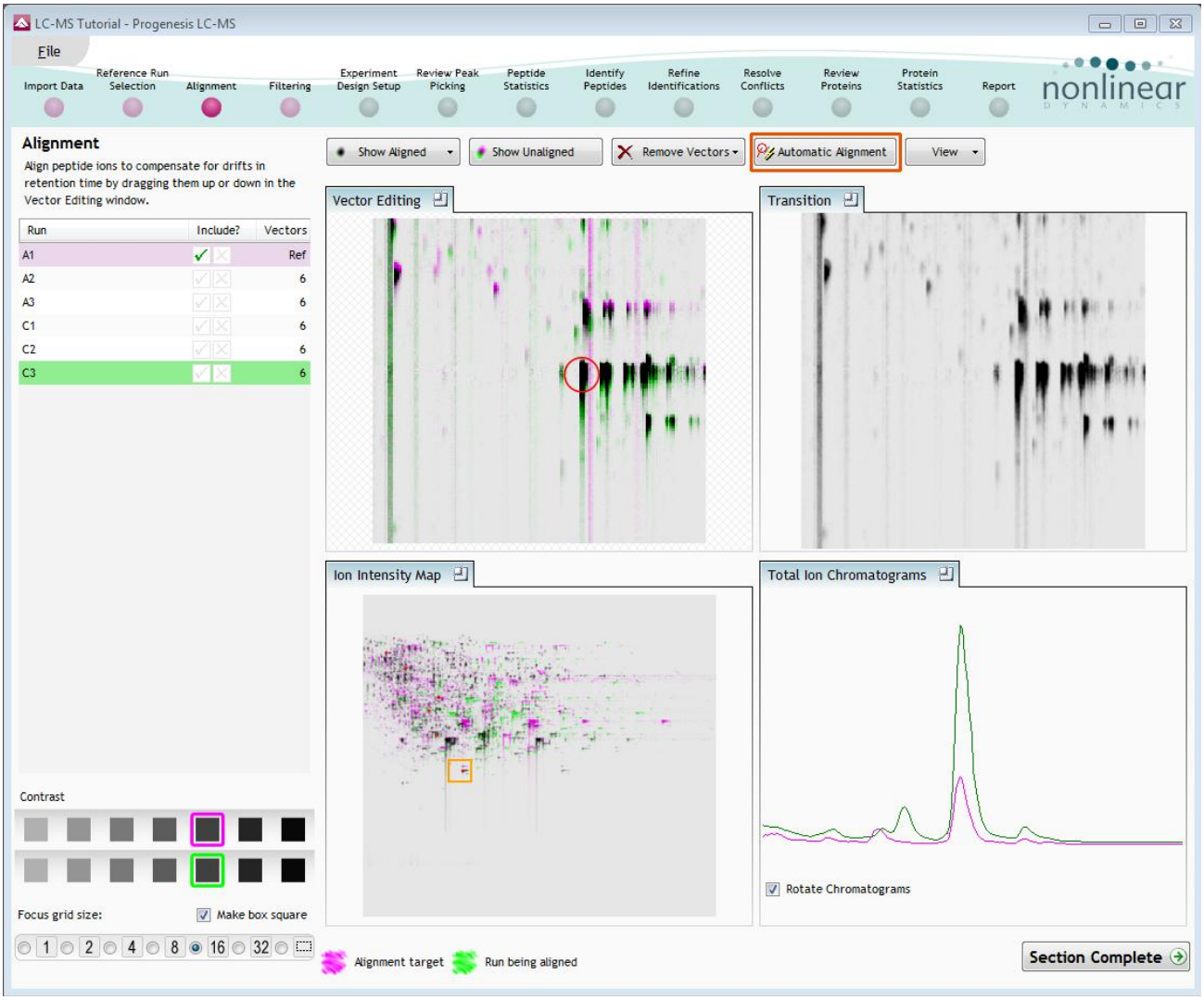



**Note**: the number of vectors you add is recorded in the **Runs** table

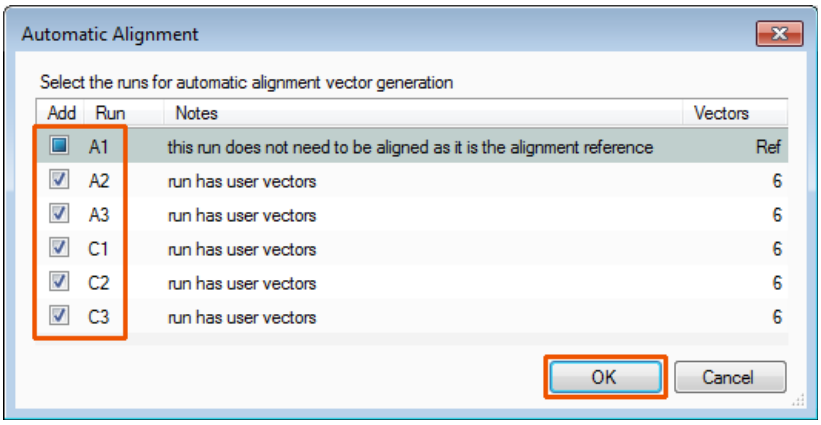
10. Repeat this process for all the runs to be aligned.

The number of manual vectors that you add at this stage is dependent on the misalignment between the current run and the Reference run. In many cases only using the Automatic vector wizard will achieve the alignment.

Also the 'ease' of addition of vectors is dependent on the actual differences between the LC-MS runs being aligned.




11. Then select Automatic Alignment to bring up the Automatic Alignment dialog and click **OK**. The automatic alignment process will begin, using the manual vectors you have added to aid in automatic vector placement.



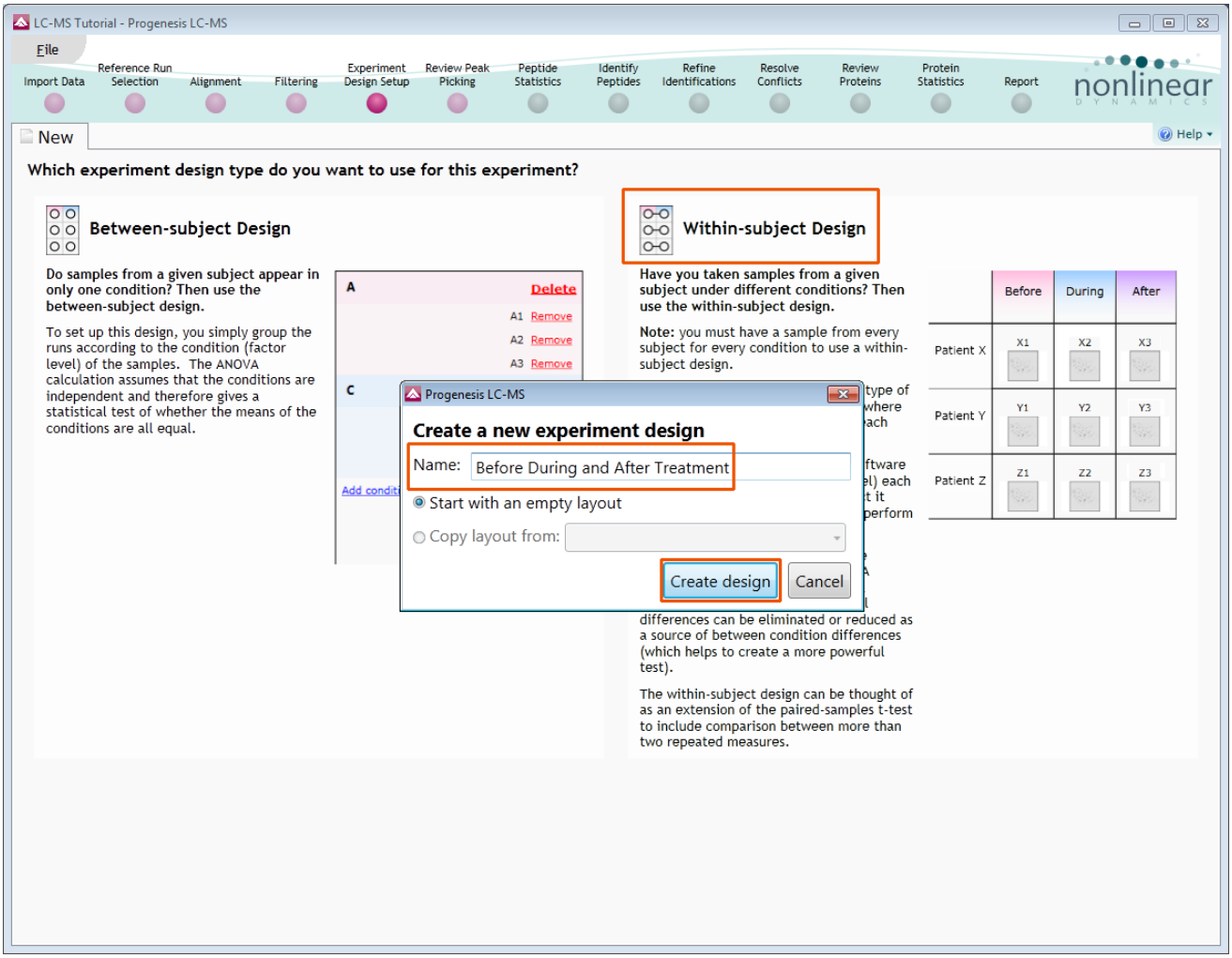

**Note**: the tick boxes next to the 'Run name control' which control whether vectors will be generated for each run.

To review the vectors, automatic and manual return to page 12

# Appendix 5: Within-subject Design

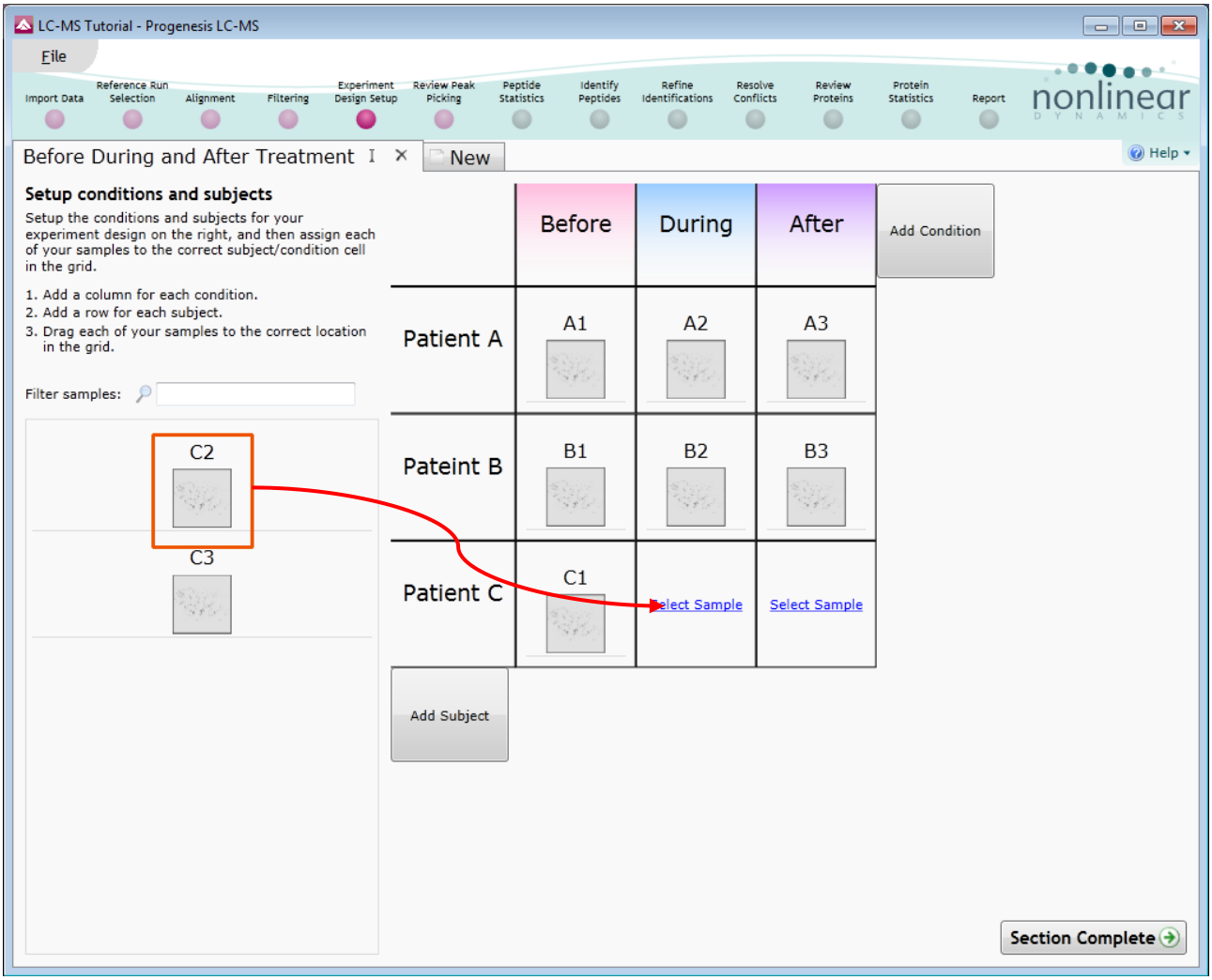
To create a **Within-subject Design** for your data set select this option on the **Experiment Design Setup** page and enter the name of the design.

In this example there are 3 Subjects (i.e. patients A, B and C) who have been individually sampled: Before(1), During (2) and After (3) treatment




When the design page opens use the **Add Subject** and **Add Condition** buttons to create the matrix that fits your experimental design, over typing the names as required.

Then Drag and drop the Samples on to the correct 'cell' of the matrix.



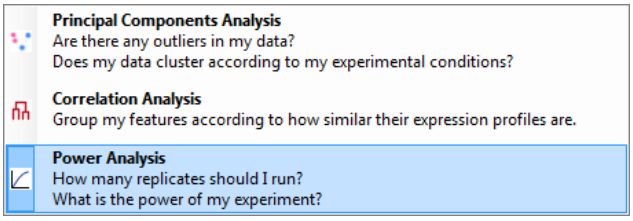

You can create additional Experimental Designs using the New tab

All of these Experimental Designs are available at all the following stages in the LC-MS workflow.

# Appendix 6: Power Analysis (Progenesis Stats)

Power analysis is a statistical technique that is used to gauge how many replicates are needed to reliably see expression differences in your data. It is available through the Progenesis Stats section of the workflow.

To perform a power analysis of the data click on **Ask another question** at the top of the table in the Progenesis Stats screen. A selection of 3 tools will appear in the form of questions.



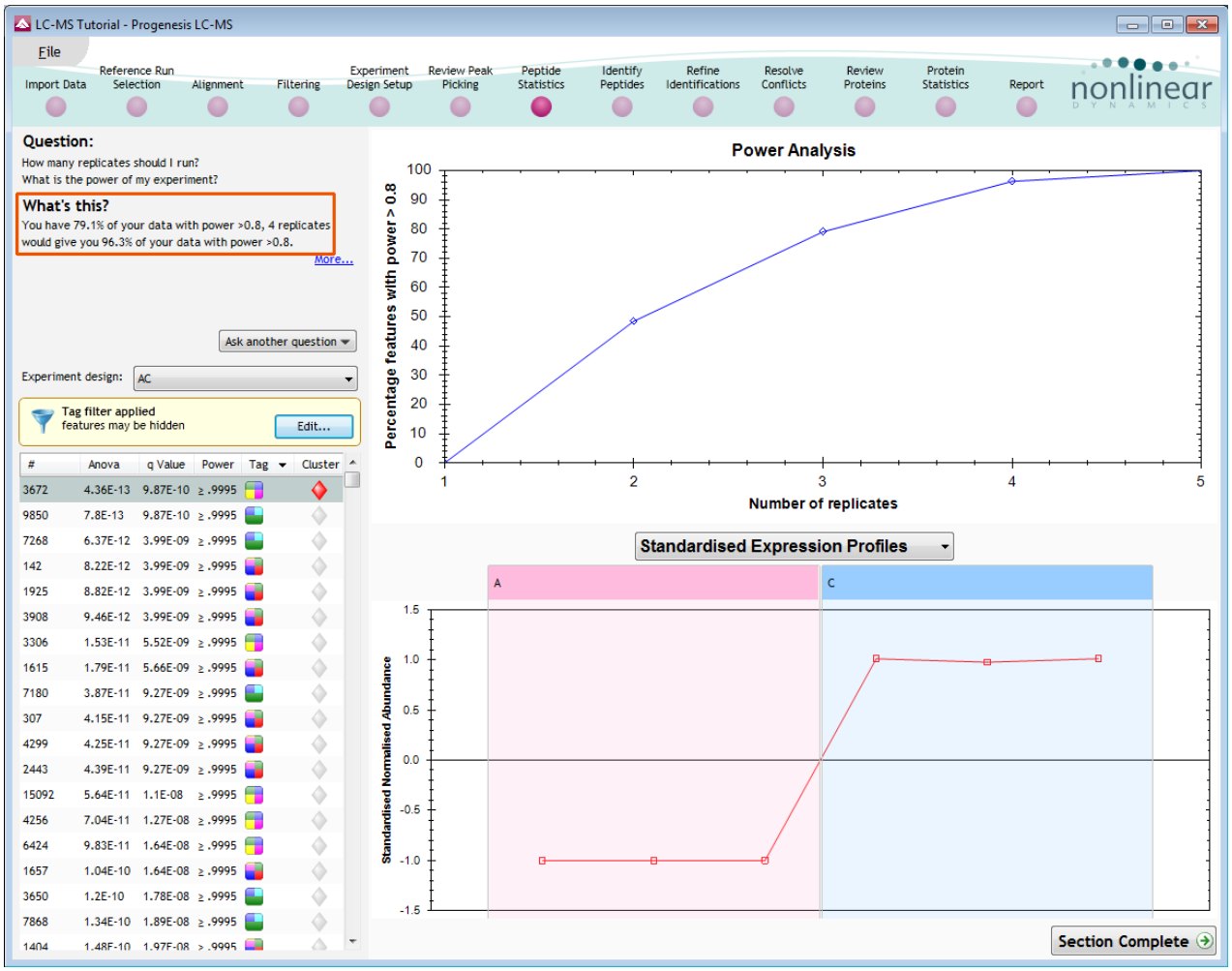

Select the option

*'How many replicates should I run and what is the power of my experiment?'*

It answers this question by informing you:

*'How many replicates you need so that at least 80% of your features with a power >0.8'*

Using the **Significant p<0.05 features (8618),** as an example, view the power analysis.

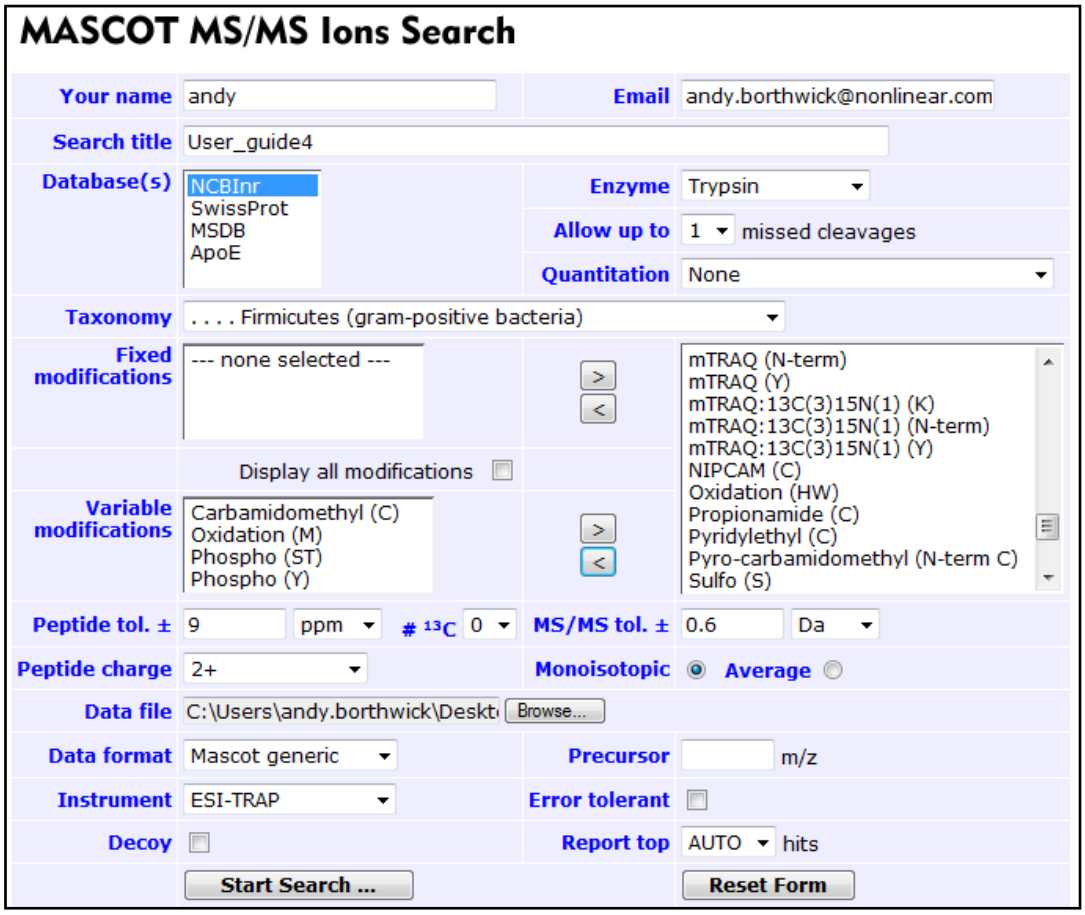



This is displayed graphically showing that 79.1% of the 8618 features have a power of 80% or that 4 replicates would give you 96.3% of your data with power > 0.8.

- **The power of a statistical test** reflects our confidence in the experimental data's ability to find the differences that do actually exist
- The power is expressed as a percentage, where 80% power is an accepted level, therefore allowing you to assess the number of sample replicates that would be required to achieve a power of 80%.

# Appendix 7 (a): Search engine parameters (Stage 9) Mascot

The parameters applied to the Mascot search that yielded the search results used in this user guide are shown below:




Database : NCBInr (circa 10/11) was used with the Taxonomy restriction set to Fermicutes

Variable modifications: Carbamylation(C), Oxidation (M), Phospho (ST) and Phospho (Y)

Peptide Tol: 9ppm

Instrument: ESI-Trap

# Appendix 7 (b): Search engine parameters (Stage 9) Phenyx

The parameters applied to the Phenyx search that yielded the search results used in this user guide are shown below:

| IDs | 60629 |
|---|---|
| Title | |
| File(s) | C:\Users\Andy.Borthwick\Desktop\LCMS Tutorial\Abundant C.mgf (mgf 108913 Kb) |
| Databank(s) | NCBInr (20080114) |
| AC | |
| Taxonomy | Firmicutes |
| Scoring Model | ESI-LTQ-Orbitrap (CID_LTQ_scan_LTQ) |
| Parent Charge | 1,2,3,4 (trust=medium) |
| Round # | 1 |
| Modifications | Oxidation_M[variable, <=4]<br>PHOS[variable, <=4]<br>Cys_CM[variable, <=4] |
| Enzyme | Trypsin_(KR_noP)<br>miss. cleav. 1<br>cleav. mode. normal |
| Parent tol. | 0.01Da |
| Pept thresholds | length>=6<br>score>=6.0<br>p-value<=1.0E-6 |
| AC Score | 6.0 |
| Conflict resolution | yes |
| Turbo scoring | tolerance=0.5Da<br>coverage >=0.2<br>series=b;b++;y;y++ |

Database : NCBInr (circa 03/09) was used with the Taxonomy restriction set to Fermicutes

Variable modifications: Carbamylation(C), OxidationM, Phospho

Peptide Tol: 0.01Da

Instrument: ESI-Trap