nonlinear dynamics

Progenesis LC-MS Tutorial

for LC-MS version 2.5

Introduction

This tutorial 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 refer to Appendix 1 then start at page 5.

How to use this document

You can print this tutorial to help you work hands-on with the software. The complete tutorial 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 Image QC 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 tutorial 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 tutorial

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 tutorial as well as invaluable discussion on the handling of the data.

Workflow approach to LC-MS run analysis

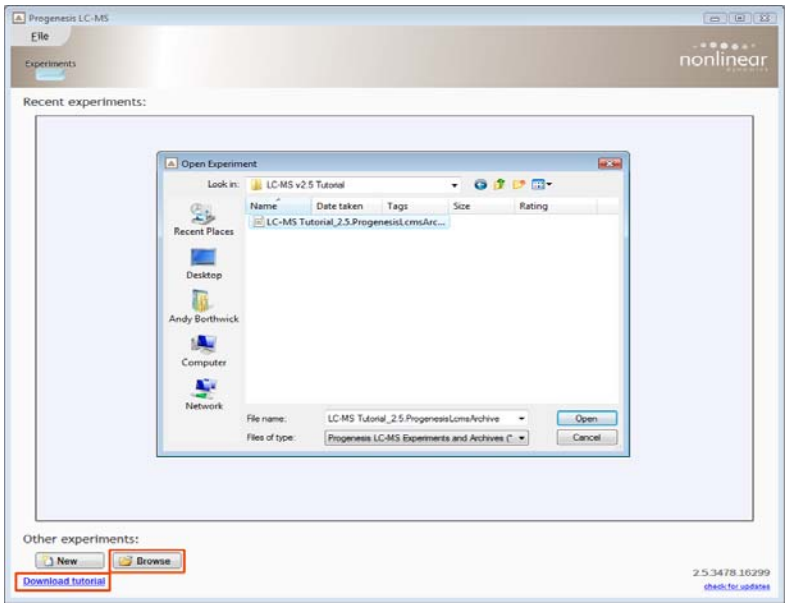
Progenesis LC-MS adopts an intuitive **Workflow** approach to performing comparative LC-MS data analysis. The following tutorial describes the various stages of this workflow (see below) focusing mainly on the stages from Alignment to Report.

Stage	Description	Page
LC-MS Data Import	**LC-MS Data Import:** Selection and review of data files for analysis.	5
Reference Run Selection	**Reference Run Selection**: Select run to align to.	6
Licensing	**Licensing**: allows licensing of individual data files when there is no dongle attached (Appendix 3)	6
Alignment	**Alignment**: automatic and manual run alignment	7
Filtering	**Filtering**: defining filters for features based on Retention Time, m/z , Charge and Number of Isotopes.	18
Group Setup	**Group Setup**: defining one or more group setups for analysed aligned runs	21
View Results	**View Results**: review and validate results, edit feature detection, tag groups of features and select features for further analysis	22
Progenesis Stats	**Progenesis Stats**: performing multivariate statistical analysis on tagged and selected groups	31
Peptide Search	**Peptide Search**: managing export of MS/MS spectra to, and import of peptide ids from Peptide Search engines	36
Peptide Filter	**Peptide Filter**: manage peptide ids and filters	39
Protein View	**Protein View**: validation and resolution of peptide id conflicts for data entered from Database Search engines	41
Report	**Report**: generate a report for peptides and proteins	46

Restoring the LC-MS Tutorial

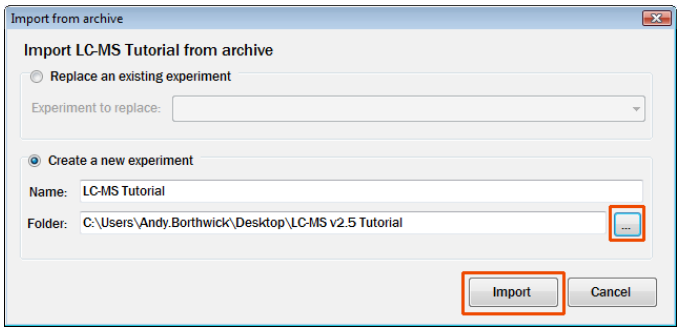
Open Progenesis LC-MS and downloaded the Compressed (.zip) Tutorial Archive file from the 'Download tutorial' link shown below, placing it in a **new folder** on your desktop. Before restoring the tutorial in the software **you must** first right click on the (.zip) file and extract it to the same folder.

Now you can restore the uncompressed LC-MS tutorial archive file. To do this, first locate the LC-MS Tutorial Archive file using the **Browse** 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) .

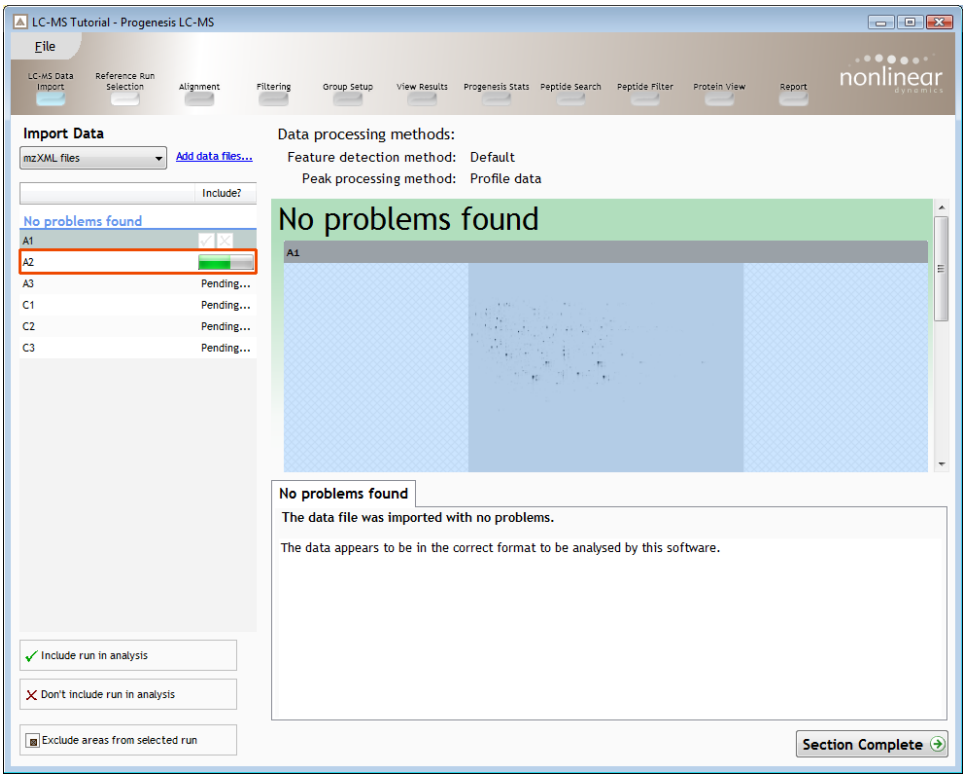


Then press **Import.**

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

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

The LC-MS tutorial will now open at the LC-MS Data Import 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 image reports on the QC of the imported Data files. In this case 'No problems found' with this data file.

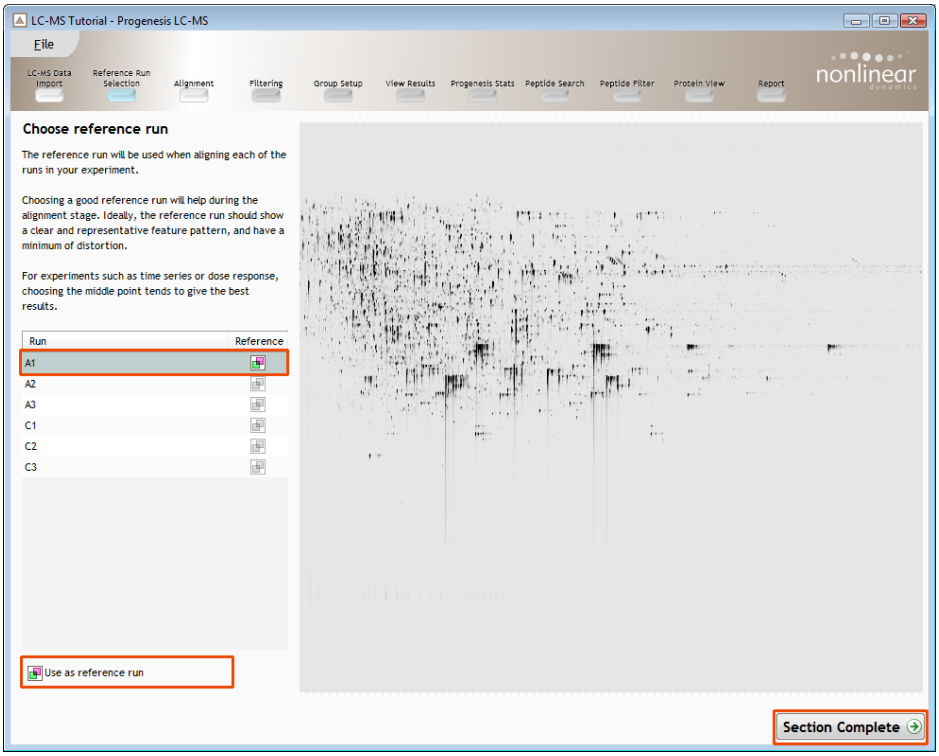
Note: the **'Data Processing Methods'** selected, when the experiment was created, are reported next to the Add data files link (see Appendix 1, page 50).

Note: 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 52).
Not required for this data set.

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.

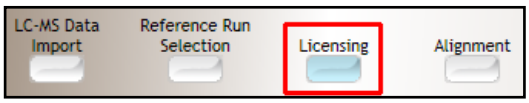


To select a Reference run either click on the run in the list and then click **Use as reference run** or double click on the run in the list.

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 in the LC-MS workflow if you are using 'Unlicensed' data files to evaluate the software and have no dongle attached.

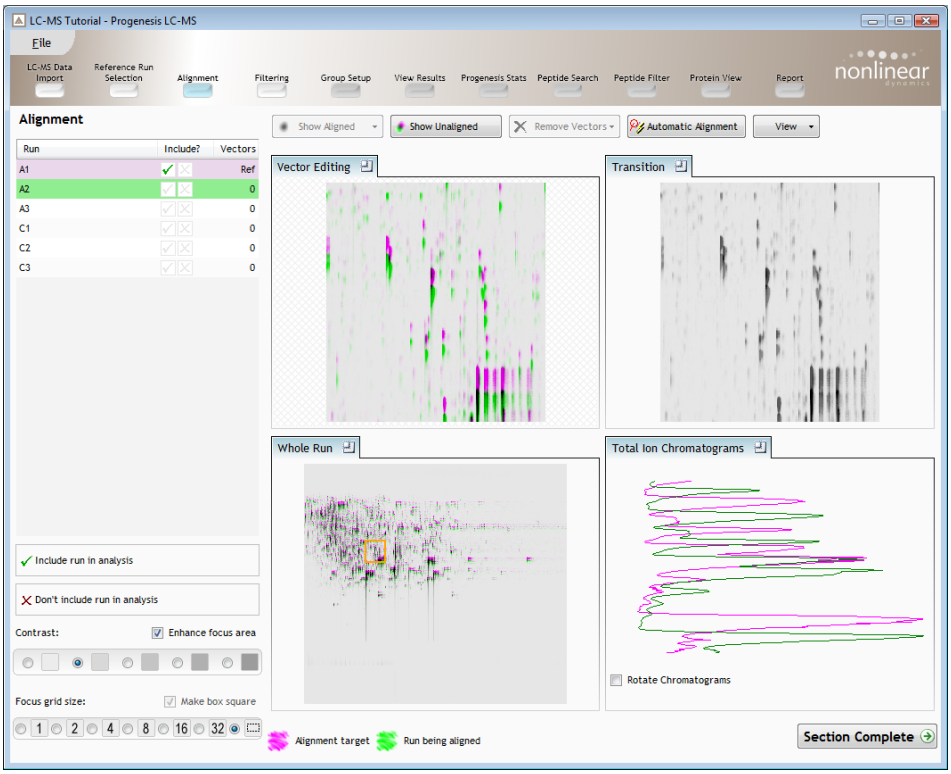


For details on how to use Licensing go to Appendix 3 (page 53)

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

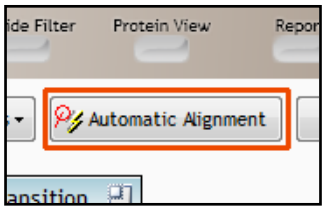
Stage 4: Alignment

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



Generation of alignment vectors

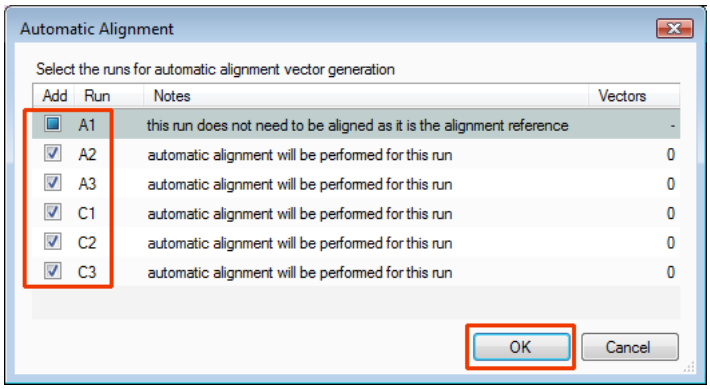
The alignment of LC-MS runs is required in the LC (retention time) direction, this is key to correcting for the variable elution of peptides during the chromatographic separation.



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**.



If applying alignment automatically now move to page 16

The following pages in this tutorial explain in more detail the views and functions of the Alignment stage in the Progenesis LC-MS Alignment, with details on:

- The Program layout
- Adding vectors manually

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 and will help with the refinement of your alignment.

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.

In this example try placing some manual vectors before generating the automatic vectors.

The following sections describe the layout of the Alignment Stage and how to manage the placement of manual vectors on your LC-MS runs

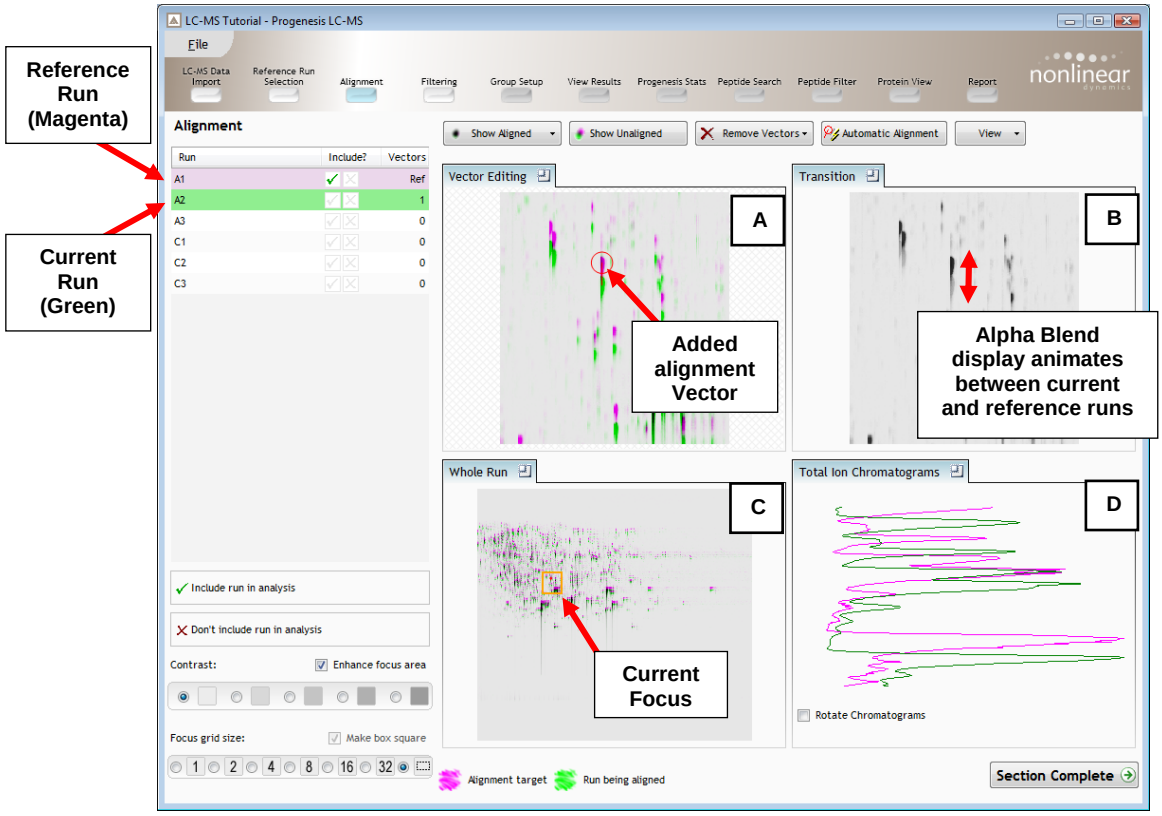
Layout of Alignment

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:

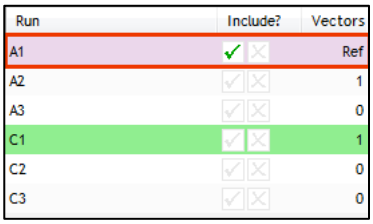
- Click on the features shown in the current focus (orange rectangle) in Window C, this will update windows A,B and D as shown below.
- In window A **click and hold** the left mouse button on a green feature.
- If the green and magenta features (immediately above) have not aligned automatically then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

The **Ref** run for any experiment is the run that you chose, in this case **A1** highlighted in magenta.



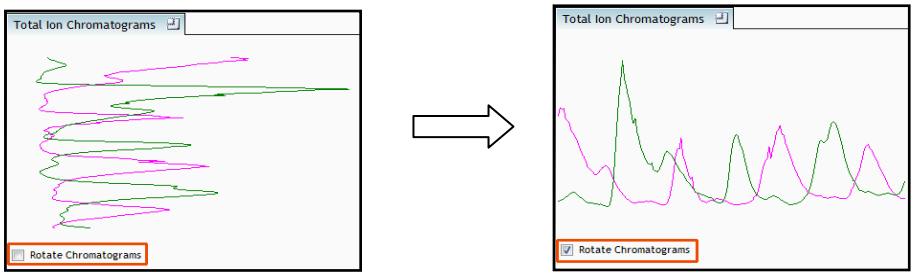
Vector Editing (Window A): is the main alignment area and displays the area defined by the current **focus** rectangle shown in Window C. The current image is displayed in green and the chosen reference image is displayed in magenta. Here is where you place the alignment vectors.

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.

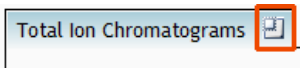
Total Ion Chromatograms (Window D): shows the current **total ion** chromatogram (green) overlaid on the Reference chromatogram (magenta). As the features are aligned in the **Vector Editing** view the chromatograms become aligned. The retention time range displayed is the vertical dimension of the Focus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

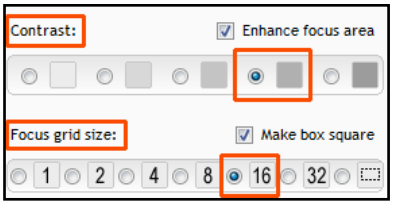
Note: the icon to the right of the 'Window' titles expands the view



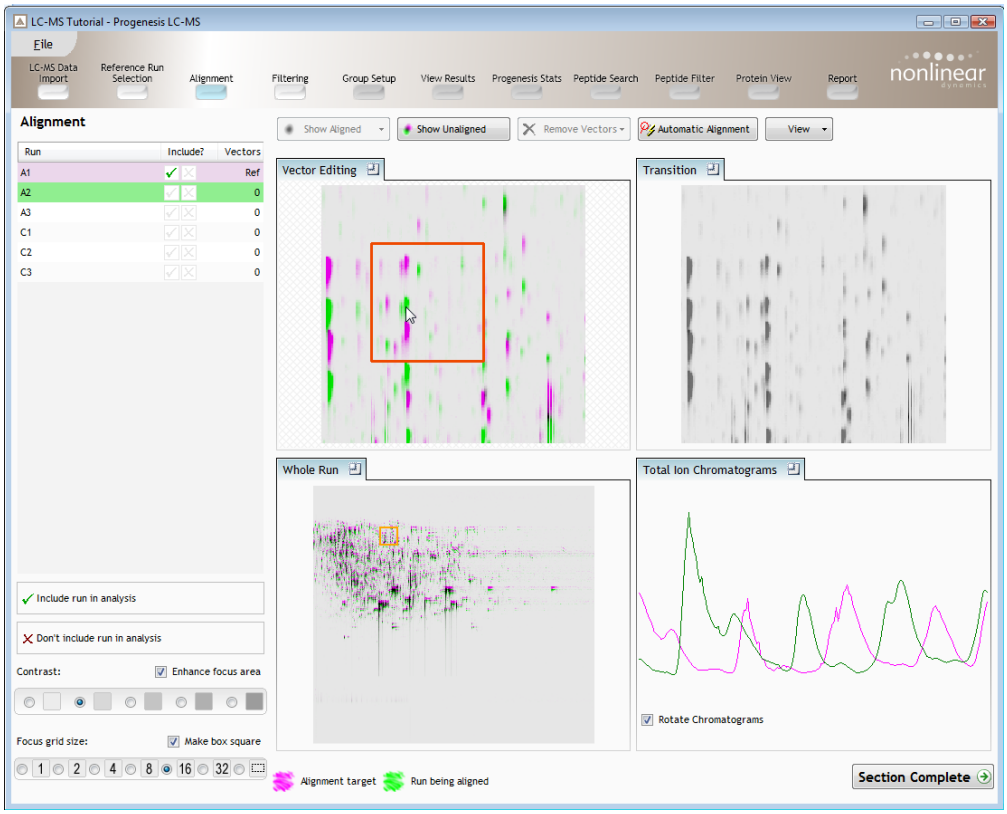
Approach to alignment

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.



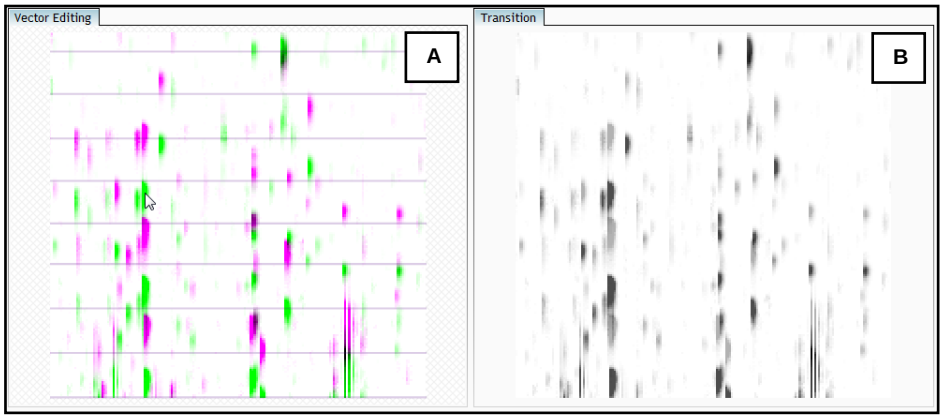
4. Click on an area (see below) in the **Whole Run** window (C) to refocus all the windows. Adjust Contrast as required.



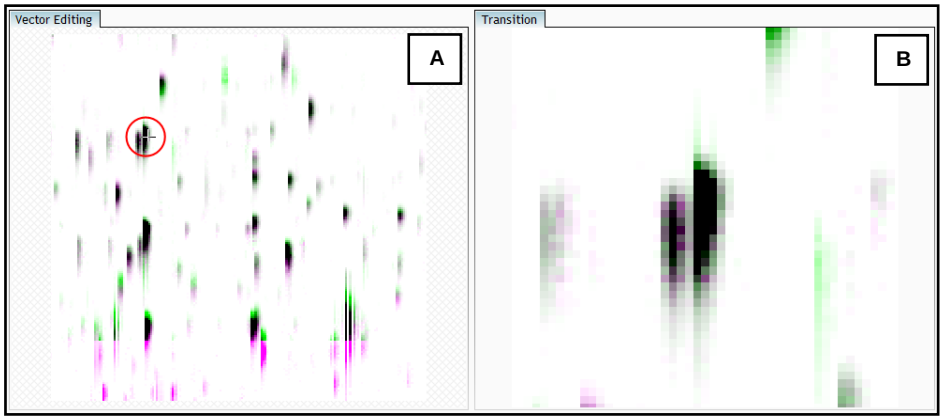
Note: the features moving back and forwards between the 2 runs in the **Transition** view indicating the misalignment of the two LC-MS runs

Note: The **Total Ion Chromatogram** 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** view

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

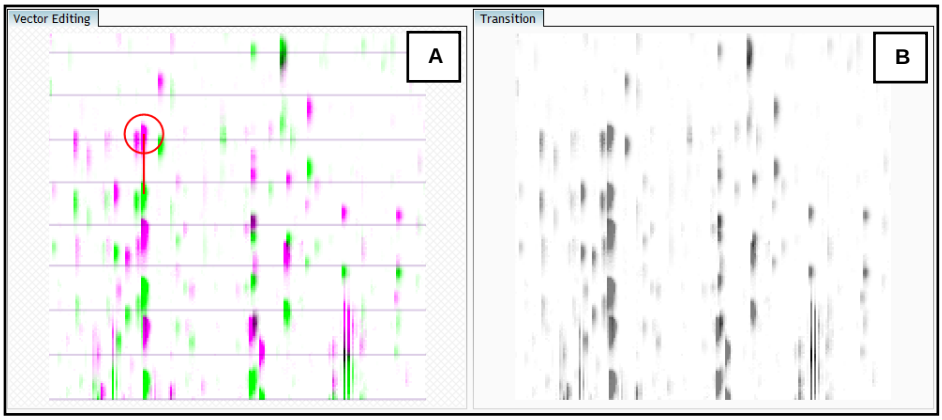


6. As you are holding down the left mouse button drag the green feature over the corresponding magenta feature of the reference image. The vector will appear as shown below as a red circle with a 'cross hair' indicating that a positional lock has been found for the overlapping features.



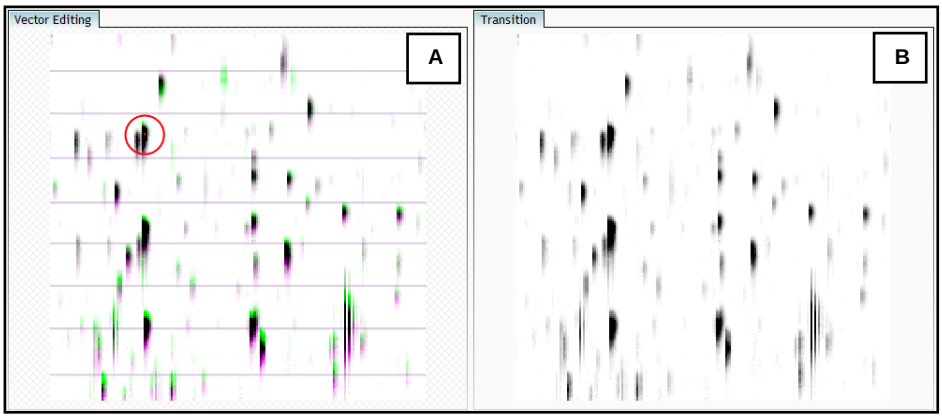
Note: as you hold down the mouse button, window B zooms in to help with the alignment.

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

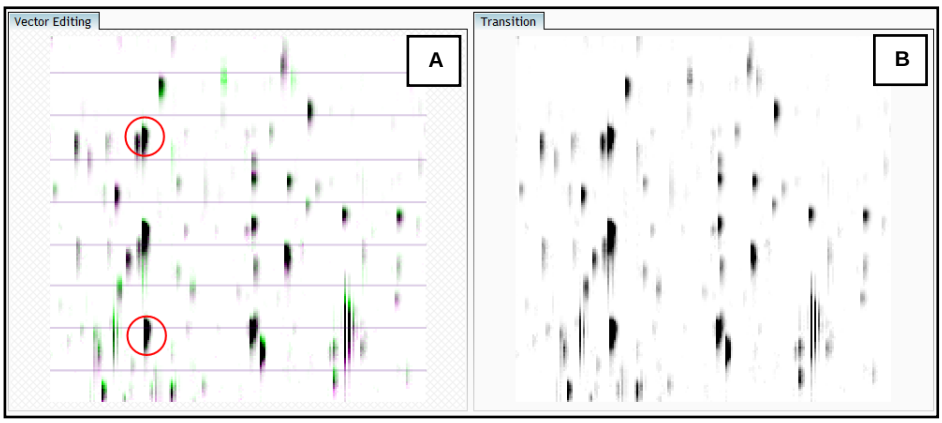


Note: an incorrectly placed vector is removed by right clicking on it in the **Vector Editing** window

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

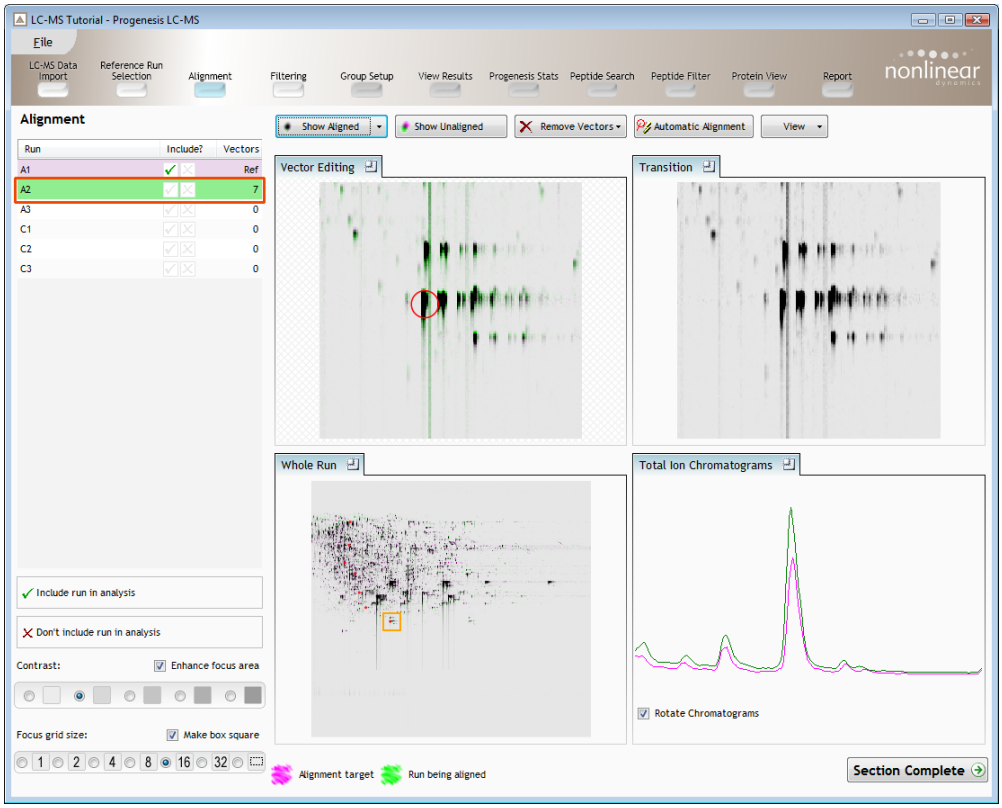


9. Additing an additional vector will improve the alignment further. **Note** this time as you click to add the vector it 'jumps' automatically to the correct position using the information from the existing alignment vector



10. 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

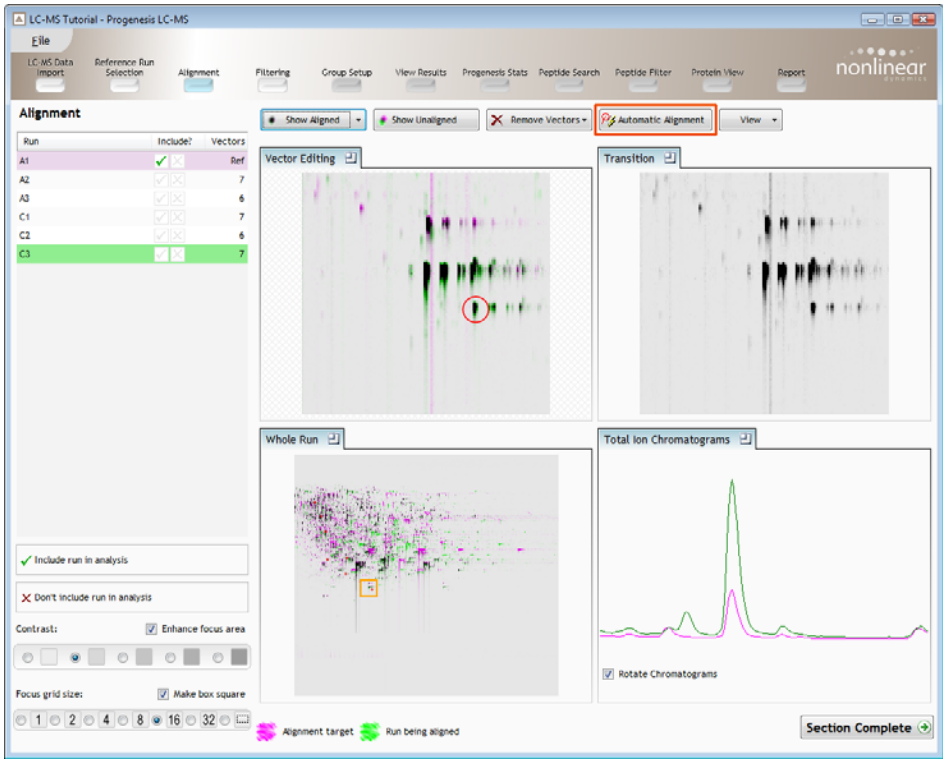


11. Now move on to the next run to align and repeat the addition of a few manual vectors

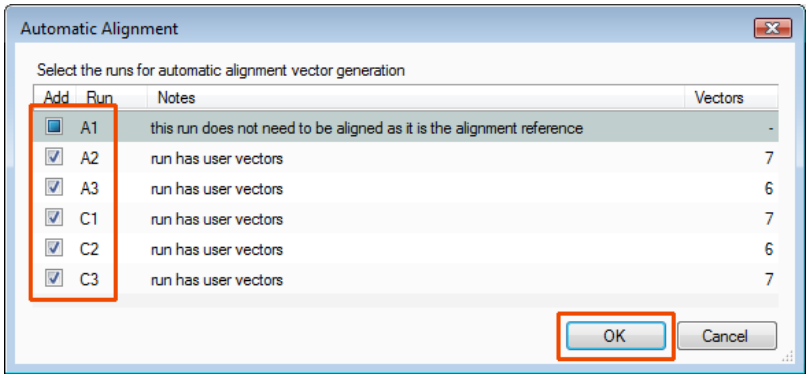
The number of manual vectors that you add at this stage is dependant on the misalignment between the current run and the Reference run. In many cases only using the Automatic vector wizard will achieve the alignment.

Also the ‘ease’ of addition of vectors is dependant on the actual differences between the LC-MS runs being aligned

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



13. Then select Automatic vectors and click **OK**.

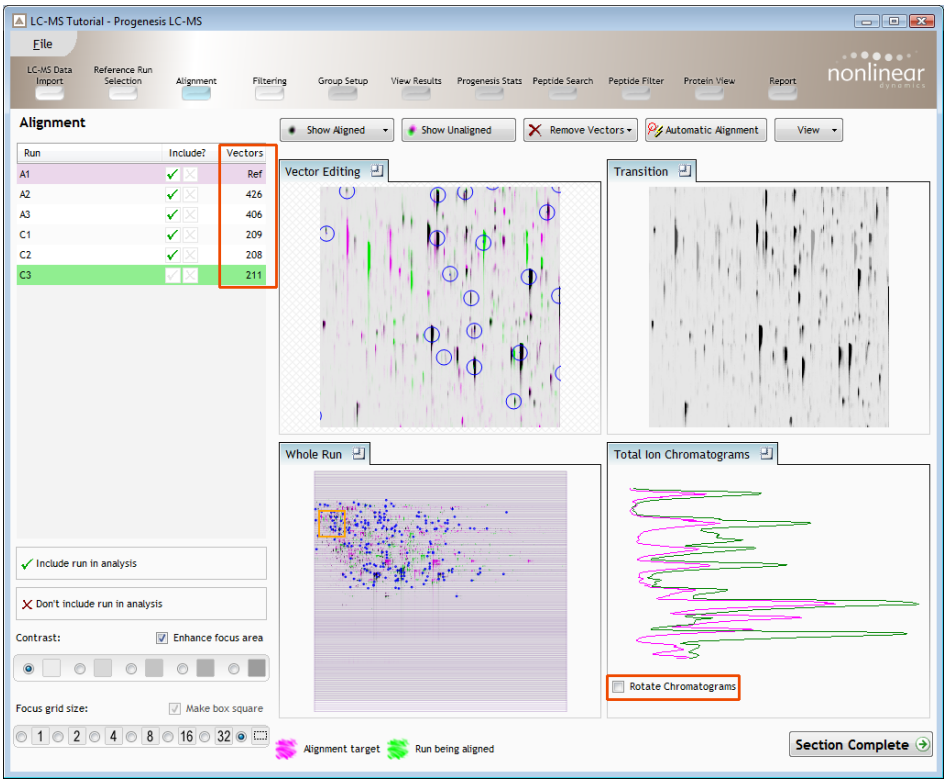


Note: the tick boxes next to the 'Run name control' which control whether vectors will be generated for each run.

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 image.

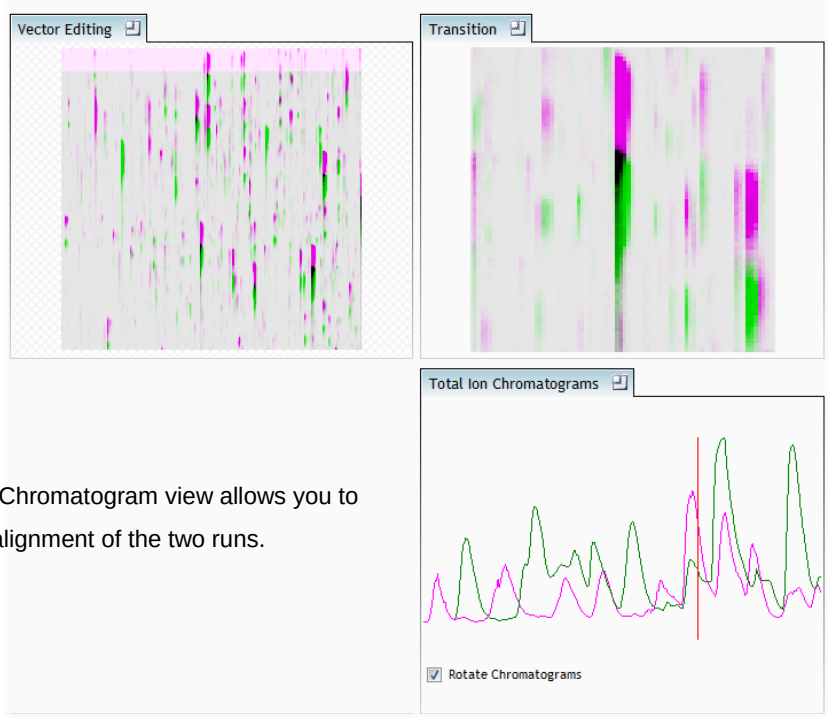
If the alignment has worked well then in Windows A and C the grid lines should show minimal distortion, Window B will show features pulsing slightly but not moving up and down.



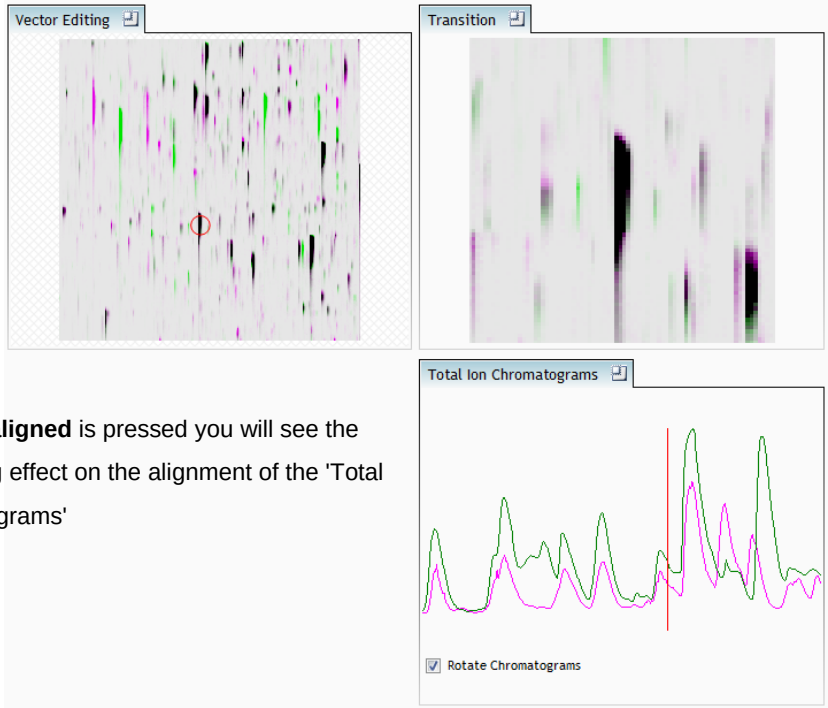
At this point, you should check the automatically placed (blue) vectors. This will be easier with a larger grid size. Make sure the grid size is set to 4 using the **'Focus grid size'** control at the bottom left of the window.

In each square, you can, if required edit the vectors to improve the image alignment (for more information refer to section on Approach to Alignment page 11

Note: the orientation of the TIC view can be changed according to individual preference using the Rotate Chromatograms option

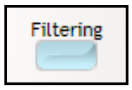


The Total Ion Chromatogram view allows you to further verify alignment of the two runs.



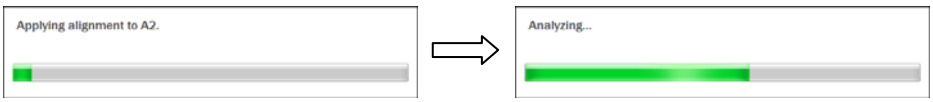
When **show aligned** is pressed you will see the corresponding effect on the alignment of the 'Total Ion Chromatograms'

Stage 5: Filtering

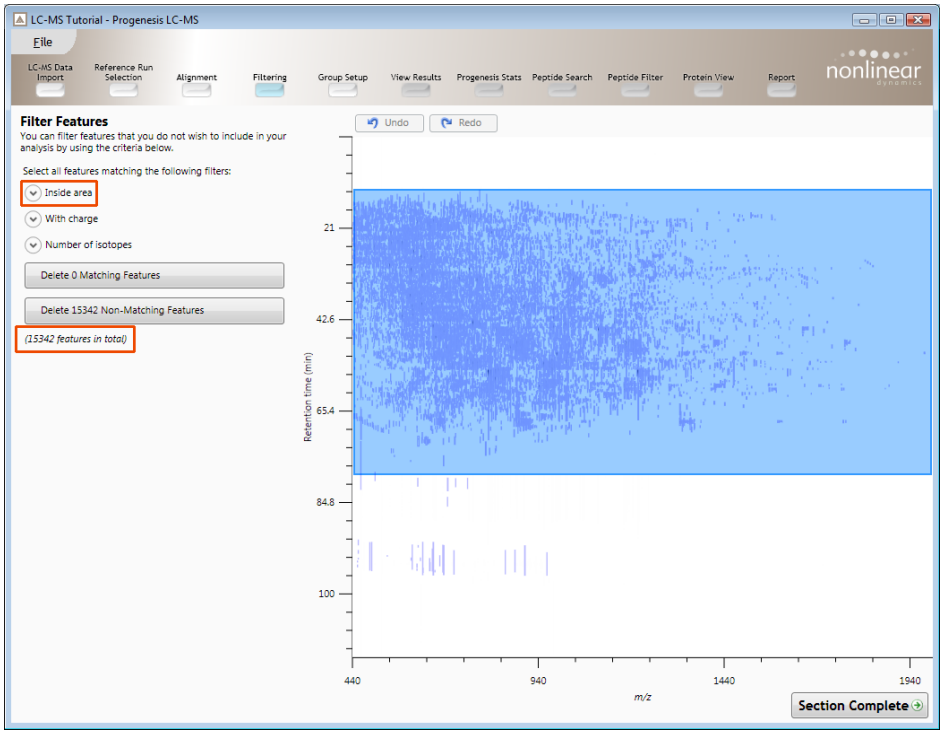


Now that you have reviewed your aligned Runs, you are ready to analyse them. Move to the **Filtering** stage, by either clicking on **Section Complete** (bottom right) or on Filtering on the workflow.

During the few minutes that the automatic analysis requires, a progress bar will appear telling you first that it is applying alignment to the Runs and then that it is Analysing.

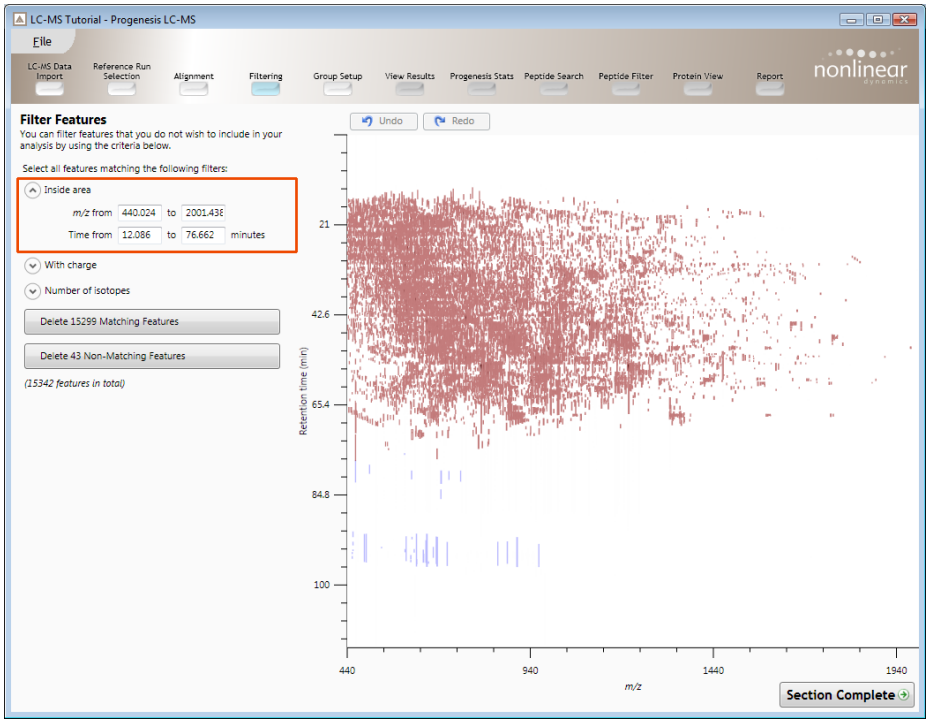


On completion of analysis the Filtering stage will open displaying the number of features detected in this example, 15342. If required you can remove features based on position, charge state, number of isotopes or combinations of these feature properties.



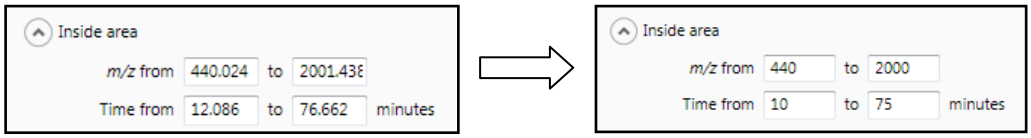
For example, to delete features with early and late ‘Retention times’ drag out an area as shown.

All features contained within the mask will be selected.



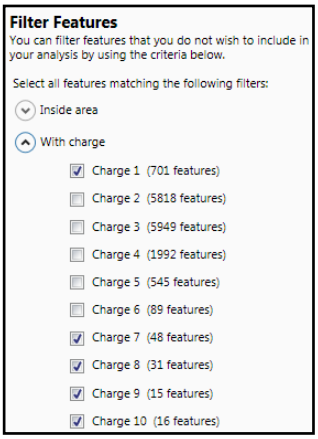
As you release the mouse button the ranges for the masked area will appear on the top left

Note: the limits can be adjusted by entering the required values in the boxes



To remove the (in this case 47) features outside of the selected area, press the **Delete 47 Non-Matching Features** button

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

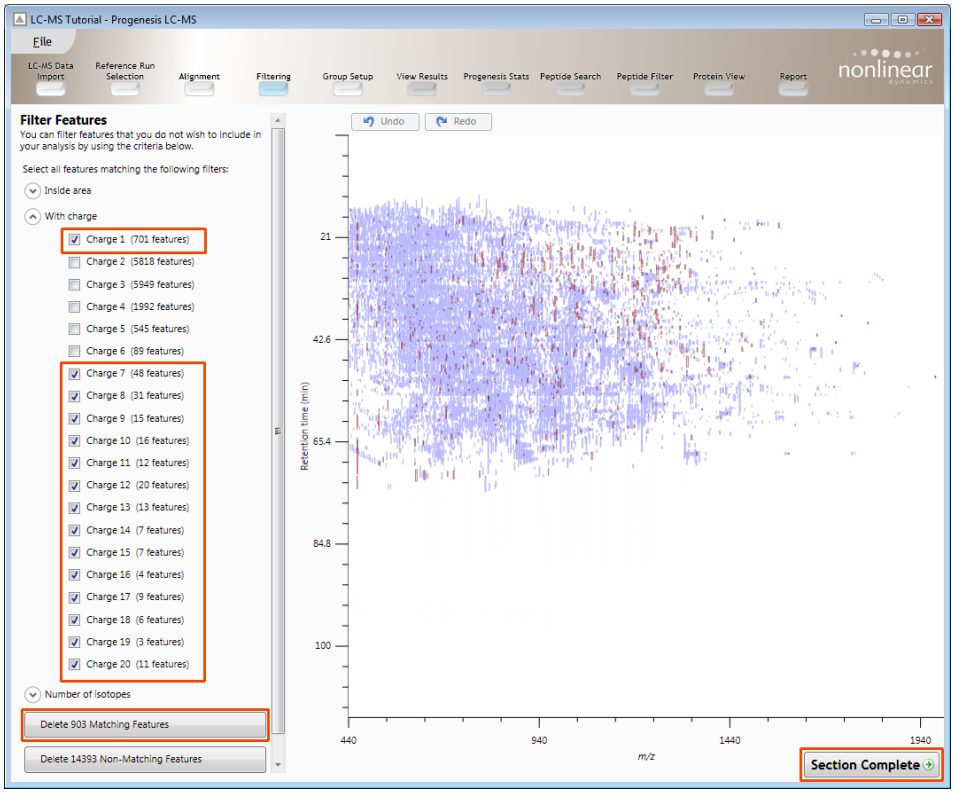


For example: when charge state is selected the number of features present at each charge state is displayed, these can be selected accordingly.

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

For this tutorial, we will filter the area as shown and remove features with a charge state of 1 and 7 and above.

We will now delete a further 903 features with a charge state of 1 and 7 and above by ticking the various options.



Hence all features with a charge state of 1 and 7 and above will appear red (see above).

To remove these features press **Delete 903 Matching Features**

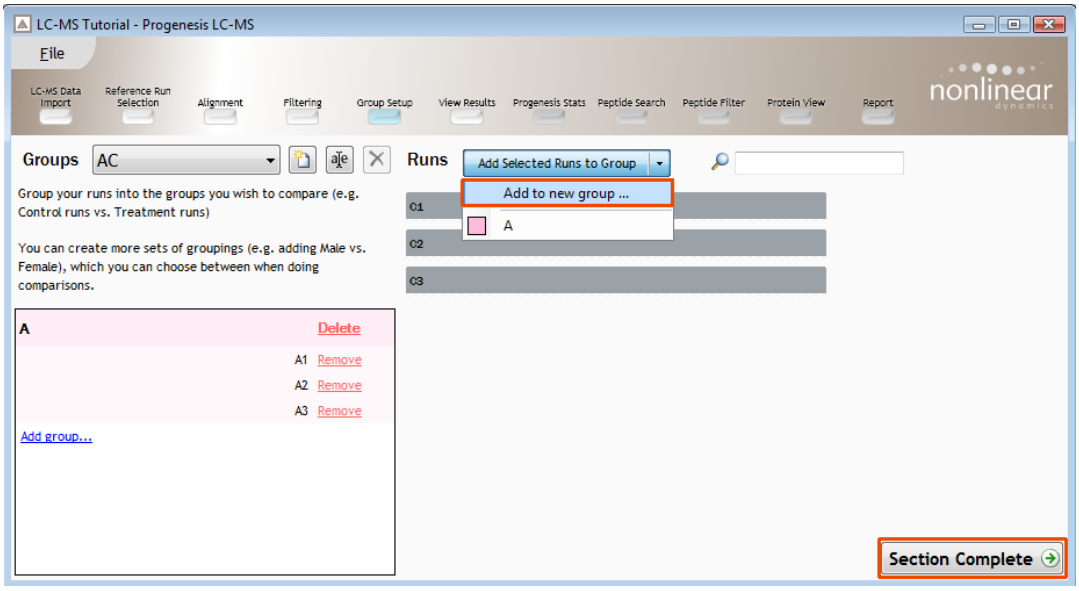
You can use the **Undo** button to bring back deleted features, however, when you move to the next section you will lose the capacity to undo the filter.

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

Stage 6: Group Setup for Analysed LC-MS runs

At this stage in the workflow you can setup one or more groupings of your sample data.

For this example, group the analysed LC-MS runs to reflect the Biological groupings in the original study. This tutorial contains 2 groups: A and C, with 3 replicates each.



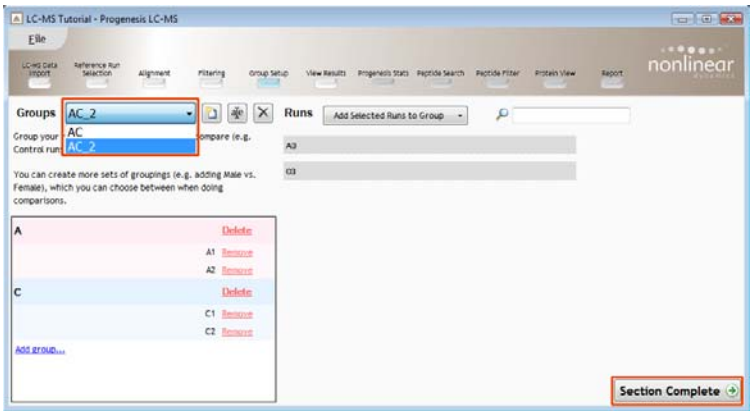
Creating a group

1. Click on all of the runs you wish to include in a group
2. Press the 'black triangle' next to the **Add Selected Runs to Group** button on the main toolbar.
3. Select **Add to new group…** from the **drop down menu**.
4. A new group will appear in the **Groups** pane on the left panel
5. Rename the group by over typing the new group name (e.g. A)
6. Repeat steps 1 to 5 until all the Runs are grouped.

In the example shown the grouping has been renamed "AC" using the rename button

To create another Group Setup, for example comparing only 2 replicates for A and C groups, click on Create a new group setup (see right).

Give it a new name (i.e. AC_2). The Runs will reappear in the main window. Create the new groups as described above.



Note: the **Group set up** drop down will now contain both setups and the ungrouped data files (A3 and C3) will remain in the main window.

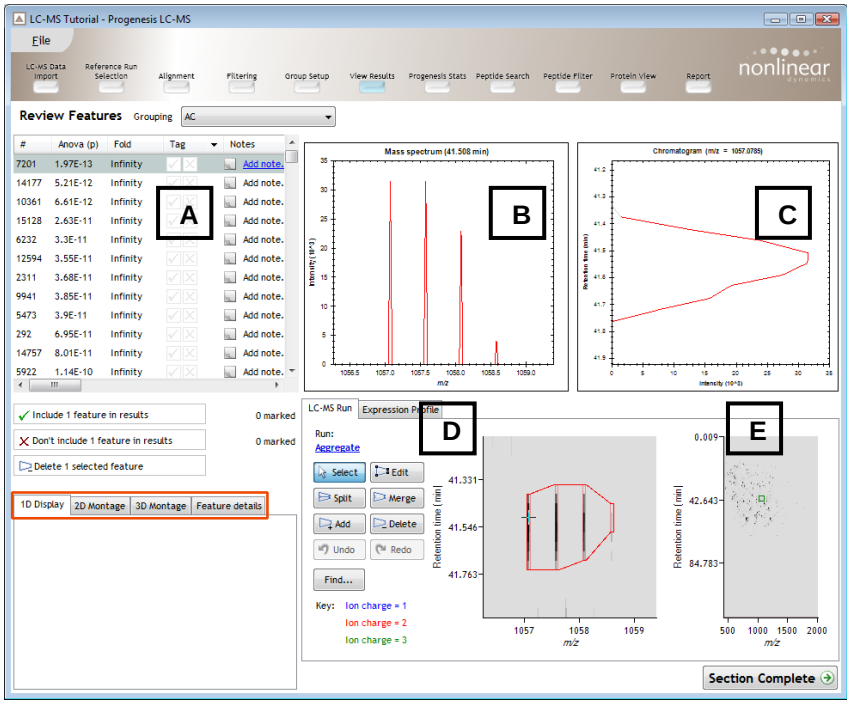
Switch back to the AC Group Setup.

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

Stage 7: Validation, review and editing of results

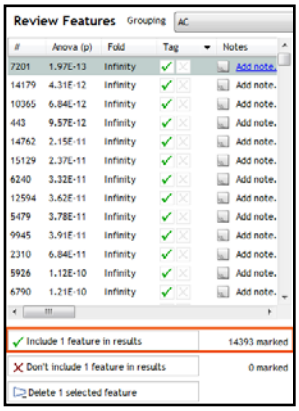
The purpose of this stage in the Workflow is to review the list of features using the visual tools provided and edit features if required.

The review stage has 4 display modes: 1D, 2D, 3D and Feature Details controlled by the tabs on



the bottom left of the display. 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



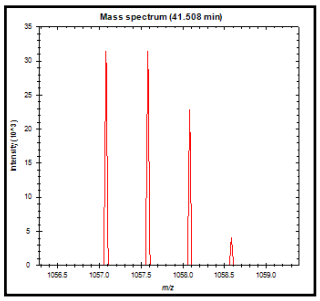
The 1D Display

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

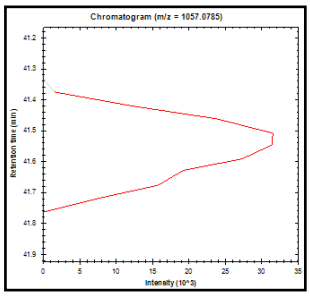
Note: A value of 'Infinity' in the **Fold** column indicates 'Presence/Absence'

To include a feature in the selection for the next section of the analysis, click on the **Include features in results** button at the bottom of the table. On clicking the button it will move on to the next feature on the list.

To select a group of features drag out a selection on the table and click on the **Include feature in results** button (see right)

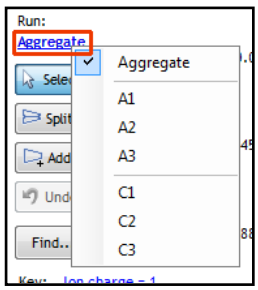


Window B: displays the Mass spectrum for the current feature on the selected Run (in window D).



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

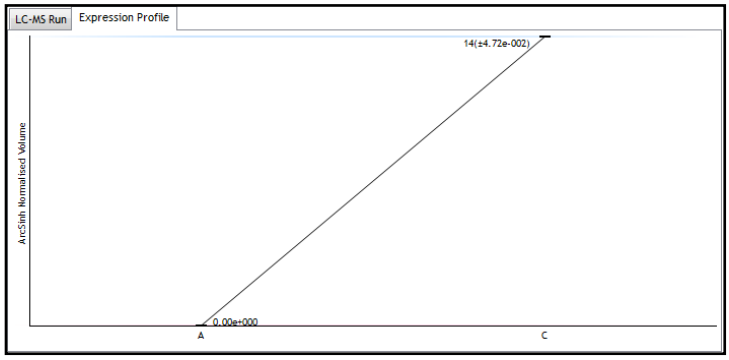
Window D: displays the details of the currently selected run. By default the selected run is an Aggregate of all the aligned runs.



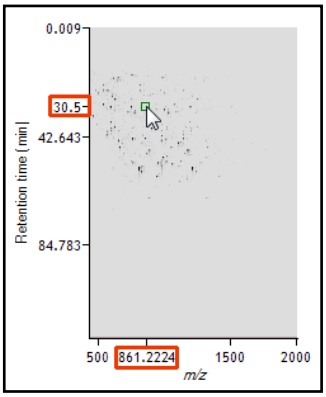
Details of individual runs can be viewed by using the 'Run' link and selecting the run you wish to view.

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.



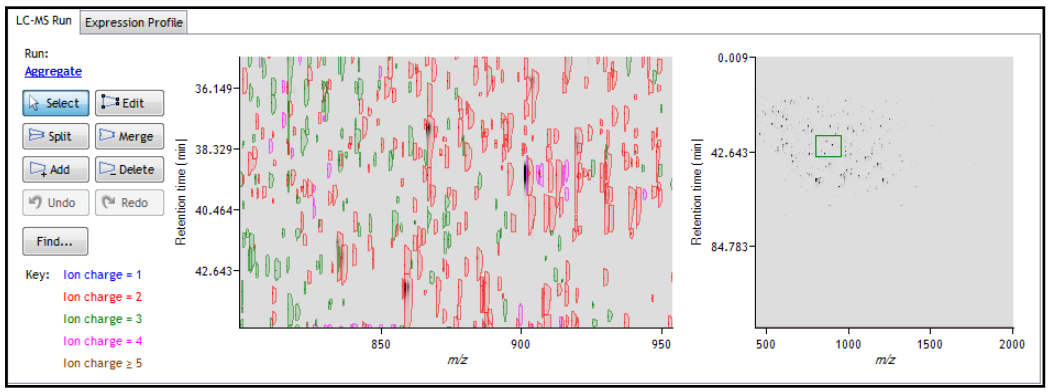
Window E: shows where the current feature is located on the LC-MS run by means of the 'Green' rectangle.



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.

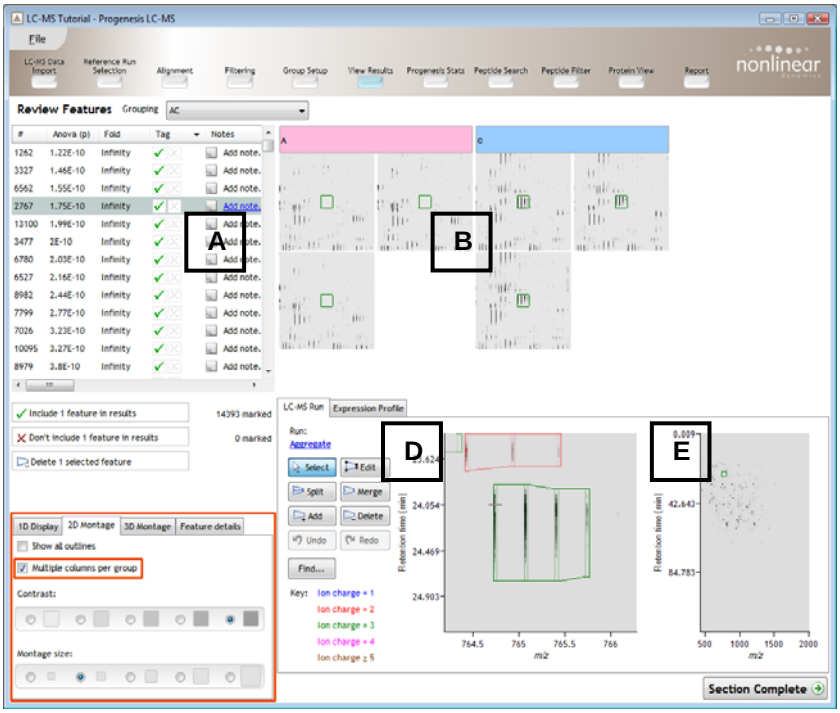
You can also drag out a larger area on this view that will refocus the other windows



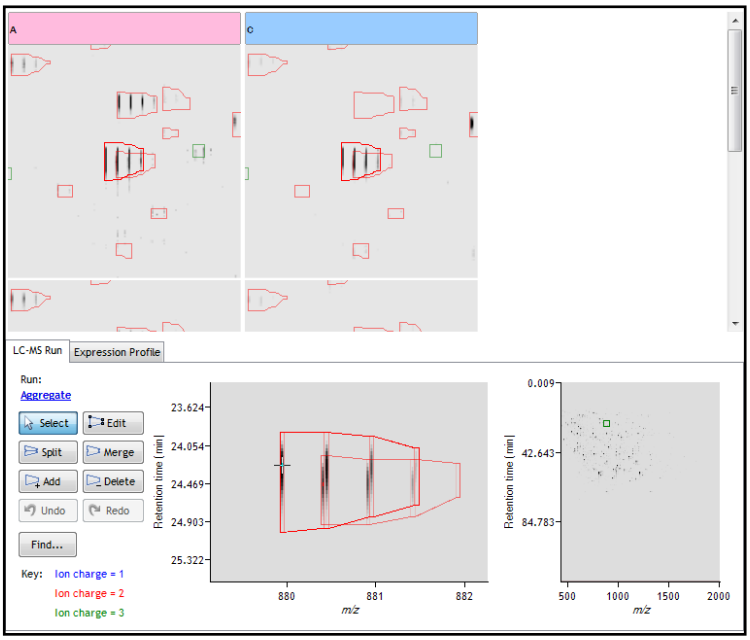
The 2D Display

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



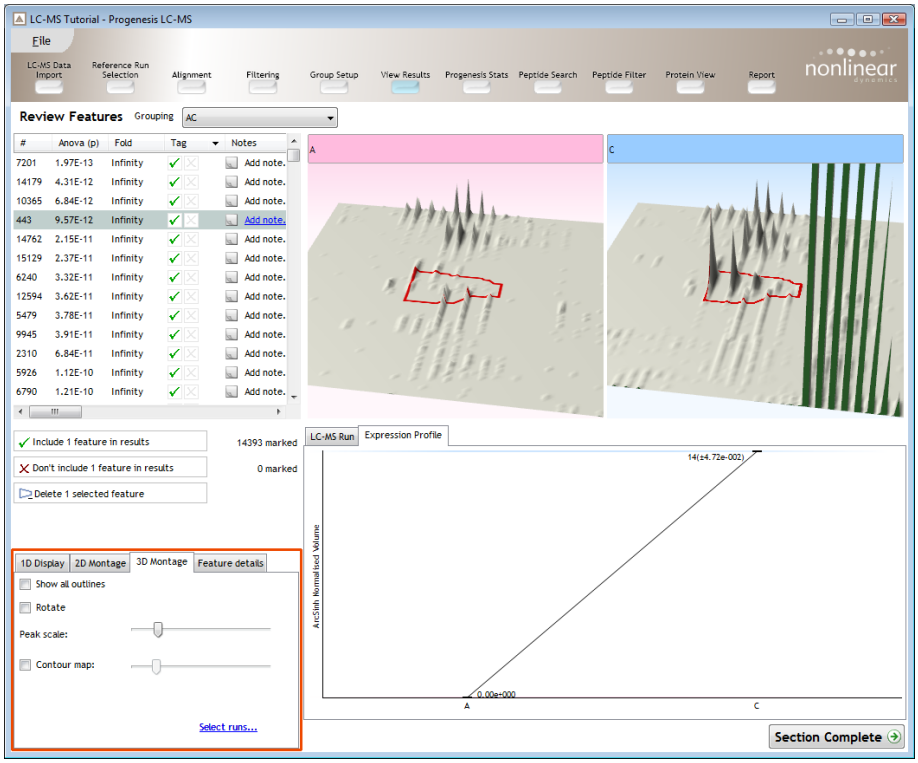
The appearance of the Montage (window B) is controlled by the panel on the bottom left of the display.



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

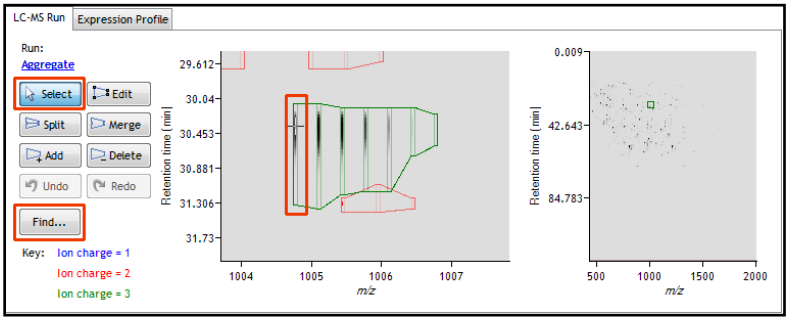
Window B changes into a 3D view by selecting the 3D Montage tab on the bottom left of the display.



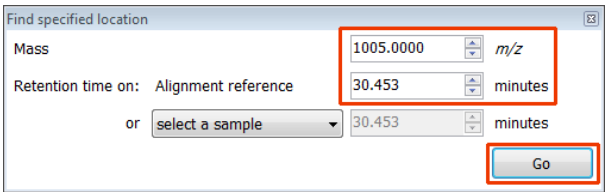
The number of 3D views displayed in the montage is controlled using the Select runs link on the 3D Montage tab. The images can be set to **Rotate** automatically or you can rotate them manually by clicking and dragging them with the mouse.

Editing of features in the View Results stage

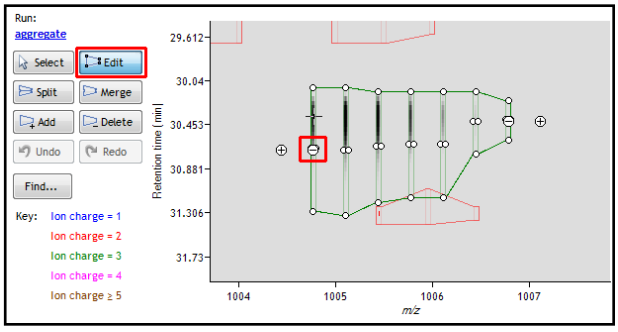
As an example of using the editing tools which are located on the left of the LC-MS Run view, we will remove and add back the 'monoisotopic peak' for the detected feature selected below. A feature can be selected from the 'Features' list or located using the various image views .



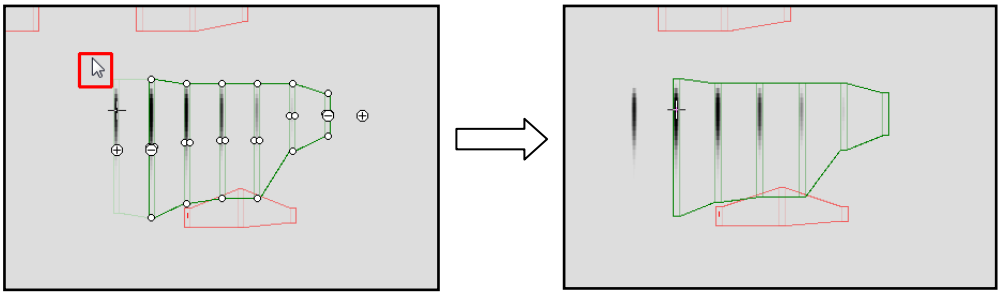
1. Locate the feature at approx 1005 m/z and 30.453 min using the **Find** tool.



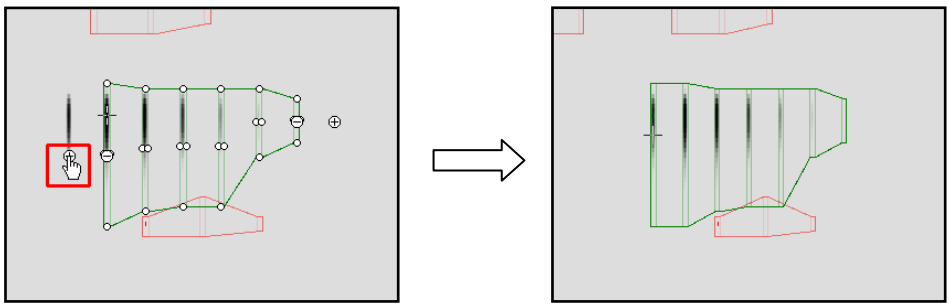
2. Select the **Edit** tool and click on the feature to reveal the 'edit handles'



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.



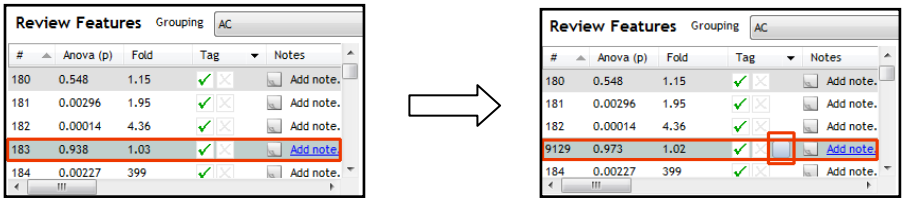
5. To add a peak to an existing feature, ensure that **Edit** is selected then click inside the feature to reveal the handles.



6. Click on the 'plus' handle on the peak to add it.

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

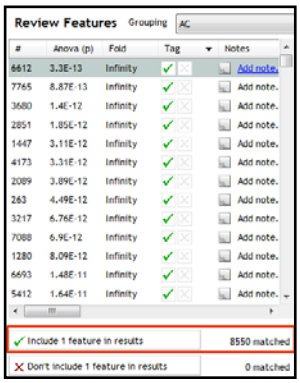
8. **Note**: If you are not satisfied with the editing use the **Undo** button and retry.



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.

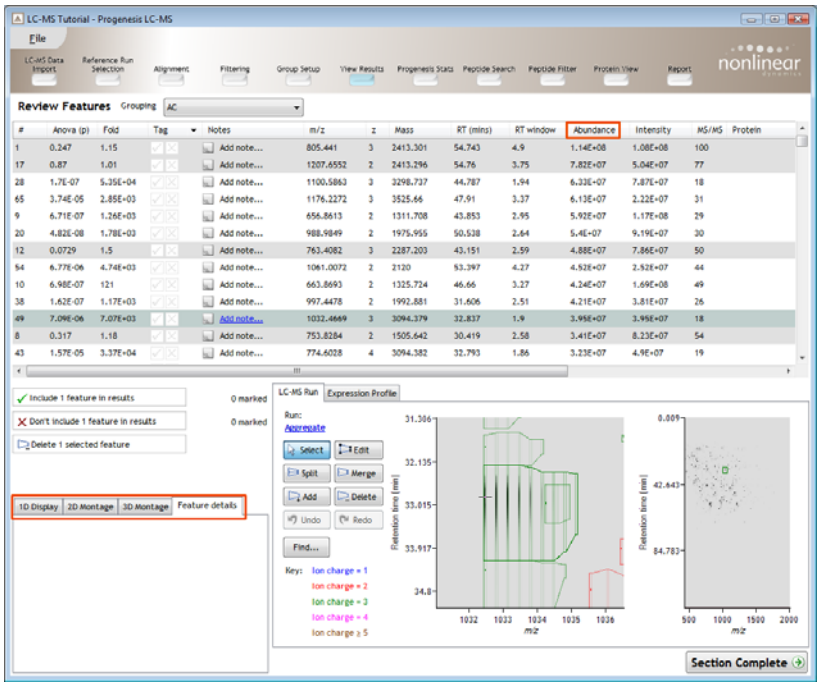
The other tools: split, merge, add and delete behave in a similar fashion and their use can be combined to achieve the desired results.

Selecting and tagging features for Progenesis Stats

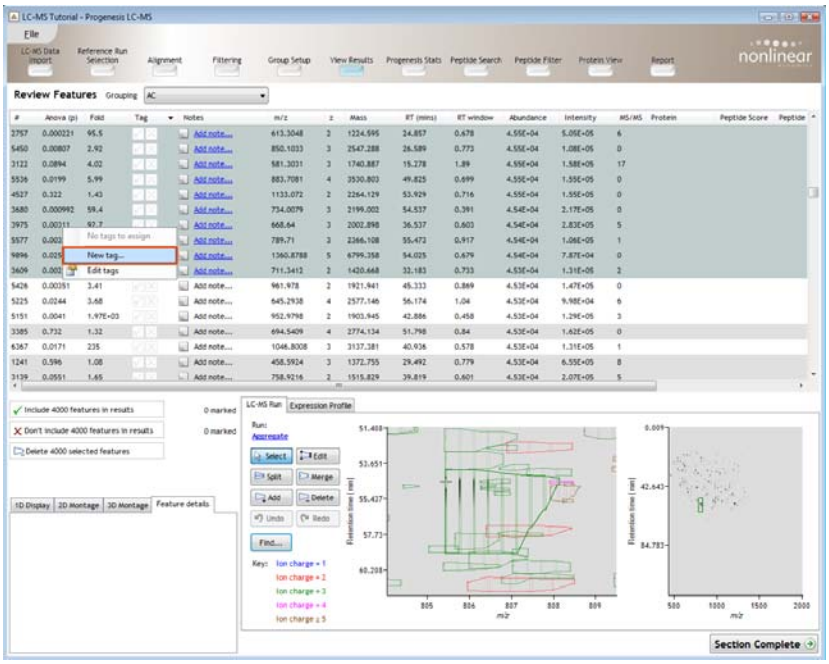


There are a number of ways to 'refine' your 'Ranked List' of analysed features before examining them with the Statistical tools in Progenesis Stats. These make use of simple 'Selection' and 'Tagging' tools that can be applied to the various Groupings created in Stage 6 (page 21). An example is described below.

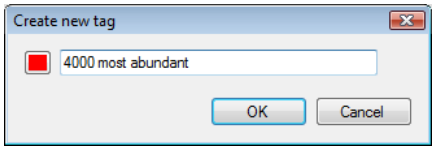
First expand the 'Features' table to show all the details by clicking on the 'Feature details' tab on the bottom left of View Results. Then order on Abundance and select the top 4000 features



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



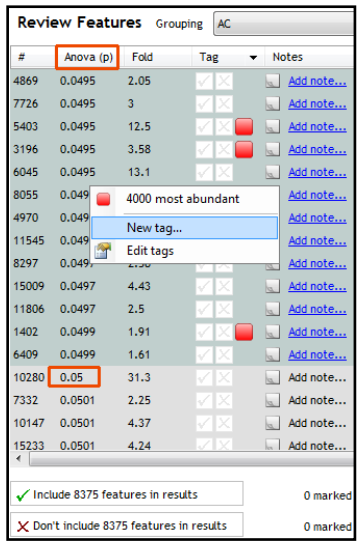
Give the Tag a name. i.e. **'4000 most abundant'**.

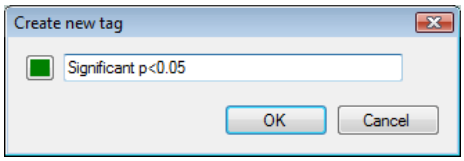


On clicking **OK** the Tag is added to the features highlighted in the table (signified by a coloured square).

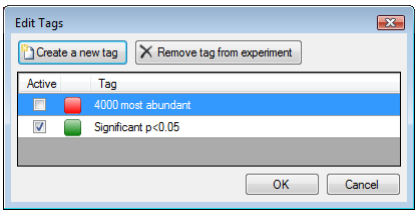
Now re-order the table based on Anova p value and highlight all values less than 0.05.

Right click and call the new tag **Significant p<0.05**



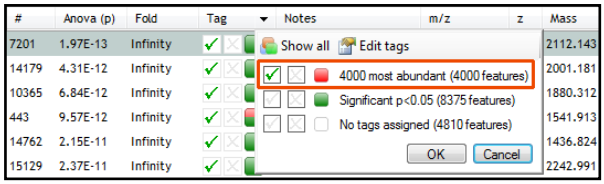


To delete and/or create additional Tags click on **Edit tags** and Create/Remove Tags as required.



Now select all the features currently displayed in the table click on a feature and press Ctrl_A. Then make sure all the features are ticked by pressing **Include 14393 features in results**

To view the Tags and also control the number of features displayed in the table, click on the drop down selection on the right of the Tag column header.



Now tick the tag for the **4000 most abundant** features

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

Stage 8: Multivariate Statistics on Selected Features

The tutorial displays the functionality of the Multivariate Statistics. This section is only available if Progenesis Stats is licensed.

Progenesis Stats opens calculating the Principal Components Analysis (PCA) for the active 'tag' in this case the **4000 most abundant** features.

For this tutorial we will start by examining the behaviour of the **4000 most abundant** features from the previous stage, **View Results**.

The statistical analysis of the selected data is presented to you in the form of interactive graphical representations 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:

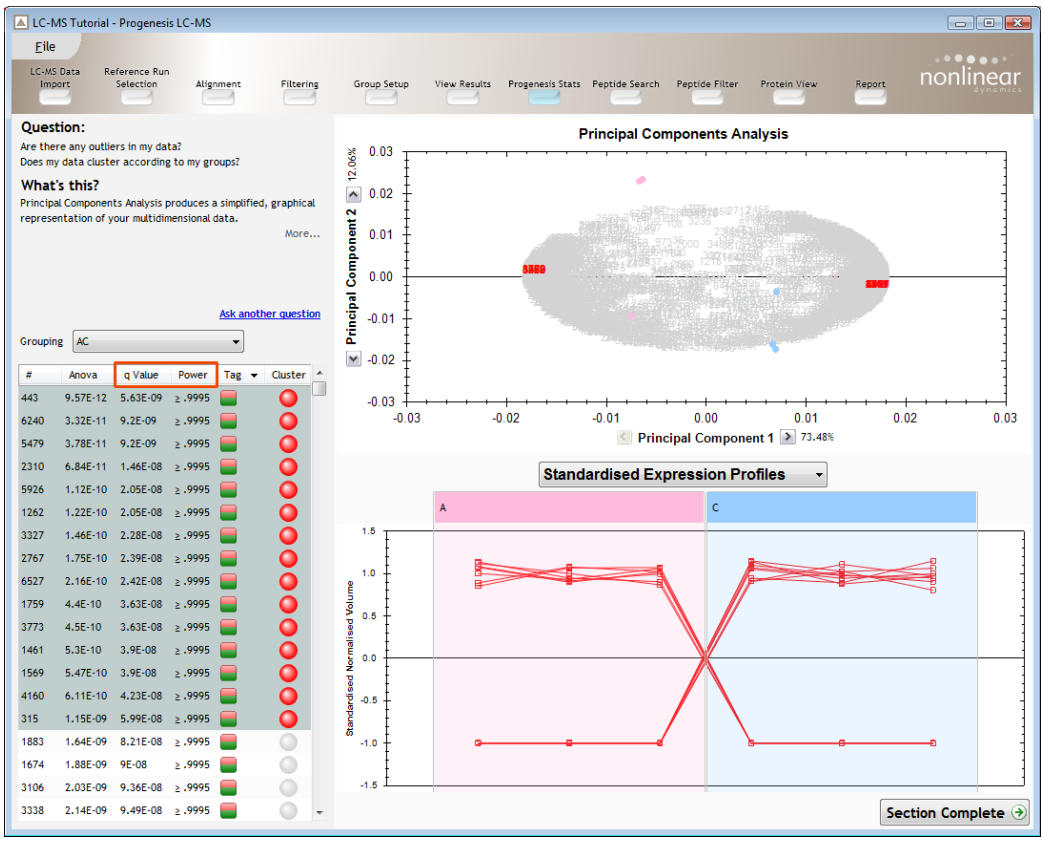
Are there any outliers in my data? And does my data cluster according to my groups?

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 groupings and conclusions can be drawn regarding possible outliers in your data.

Selecting spots in the table will highlight the spots on the 'Biplot' and their expression profiles will appear in the lower panel.



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

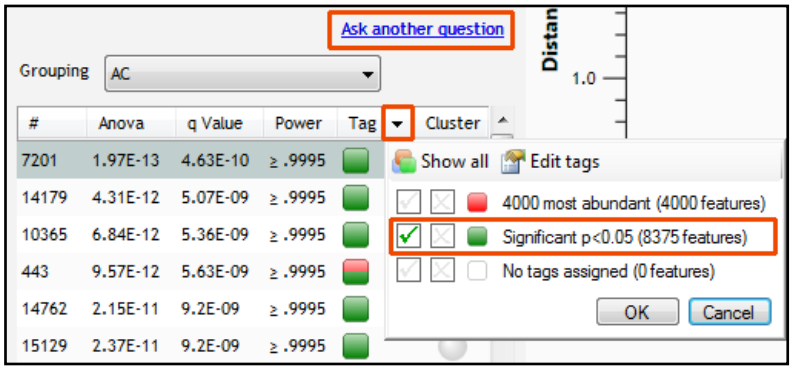
q value: tells us the expected proportion of false positives if that feature's p-value is chosen as the significance threshold

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 4 (page 54)

Correlation Analysis

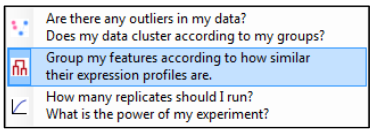
Use the tags created in View Results to filter the features in the table. We are going to explore the Correlation Analysis for all the features that were tagged at the view results stage for having an Anova p<0.05.



On pressing OK the PCA will recalculate using these 8375 features, you can (to save time) stop this calculation by pressing **Cancel calculation** and then set up Correlation Analysis for the 8375 features.

To set up the Correlation Analysis using this filtered data set click on the link 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'

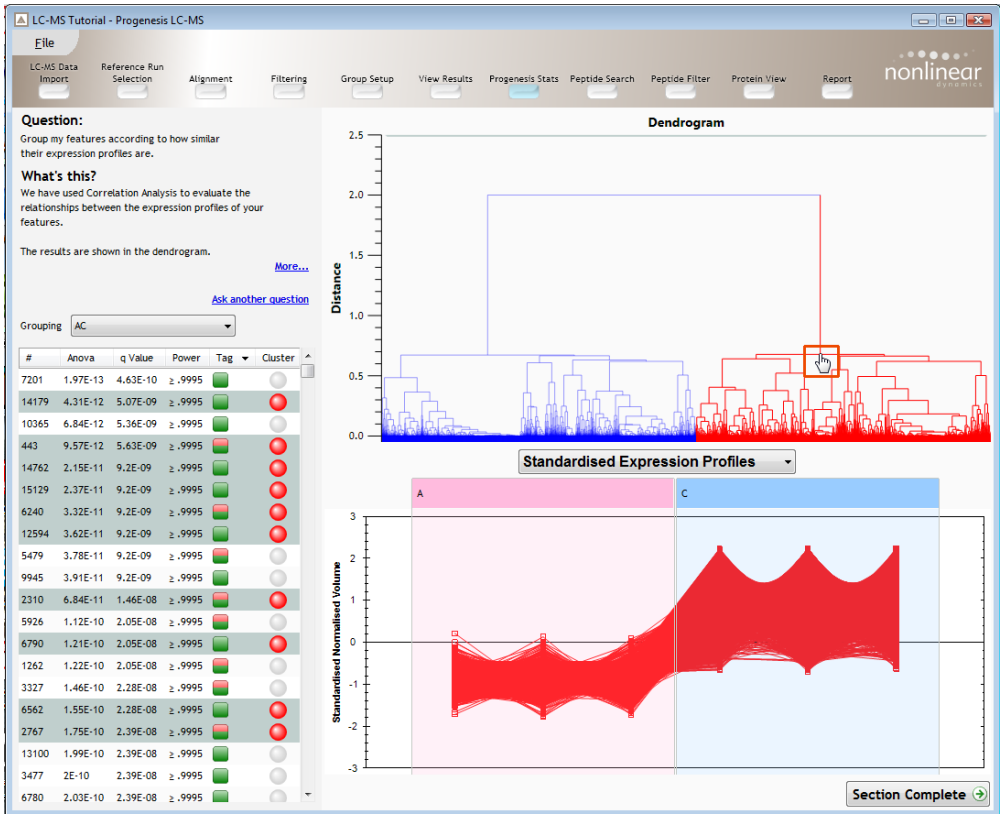
This time the statistically based question(s) being asked is:

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

The question is answered by:

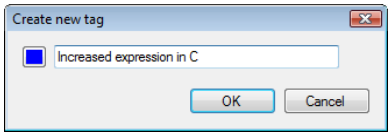
'Using Correlation analysis to evaluate the relationships between the (selected) features' expression profiles'.

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.

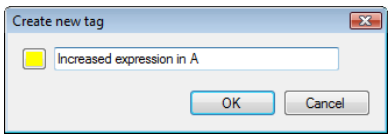


To highlight all the features demonstrating **Increased expression in the C** group click on a 'node' for a branch of the Dendrogram (as shown above). As before create a Tag for these features.

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



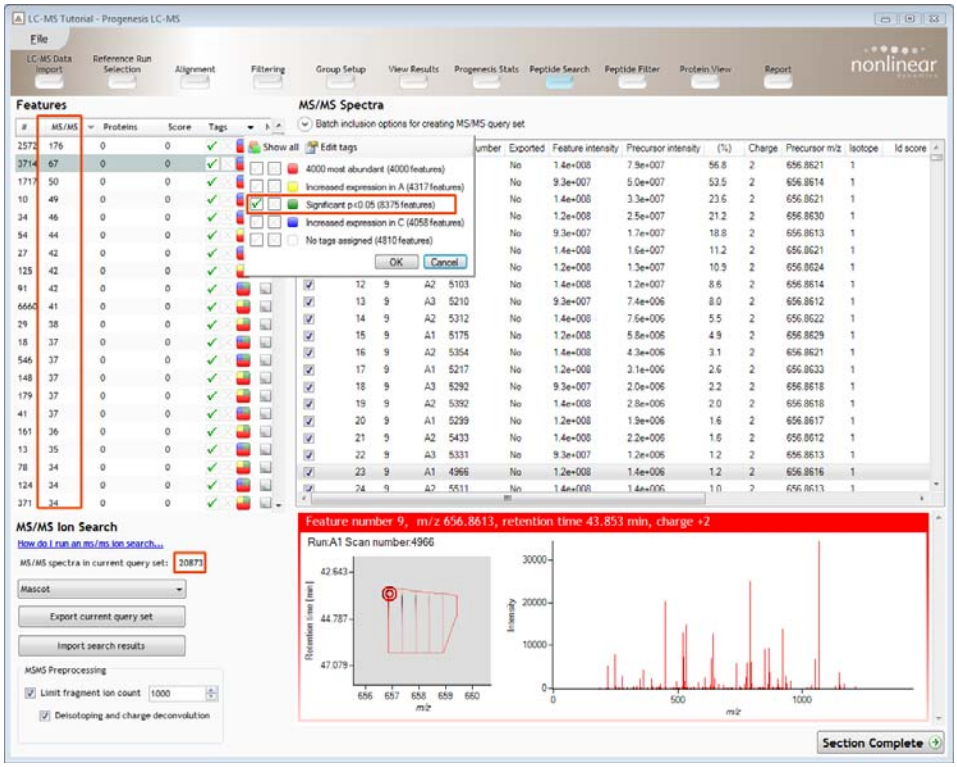
Also create a tag for those features showing **Increased expression in A** by first clicking on the other 'main' node then right click on the highlighted features in the table and creating the New tag



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

Stage 9: Peptide Search

Determining protein ID is dependant 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 dependant on thresholds and parameters set prior to the acquisition of the LC-MS run.



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

The query set can be searched using all the spectra, however the query 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.

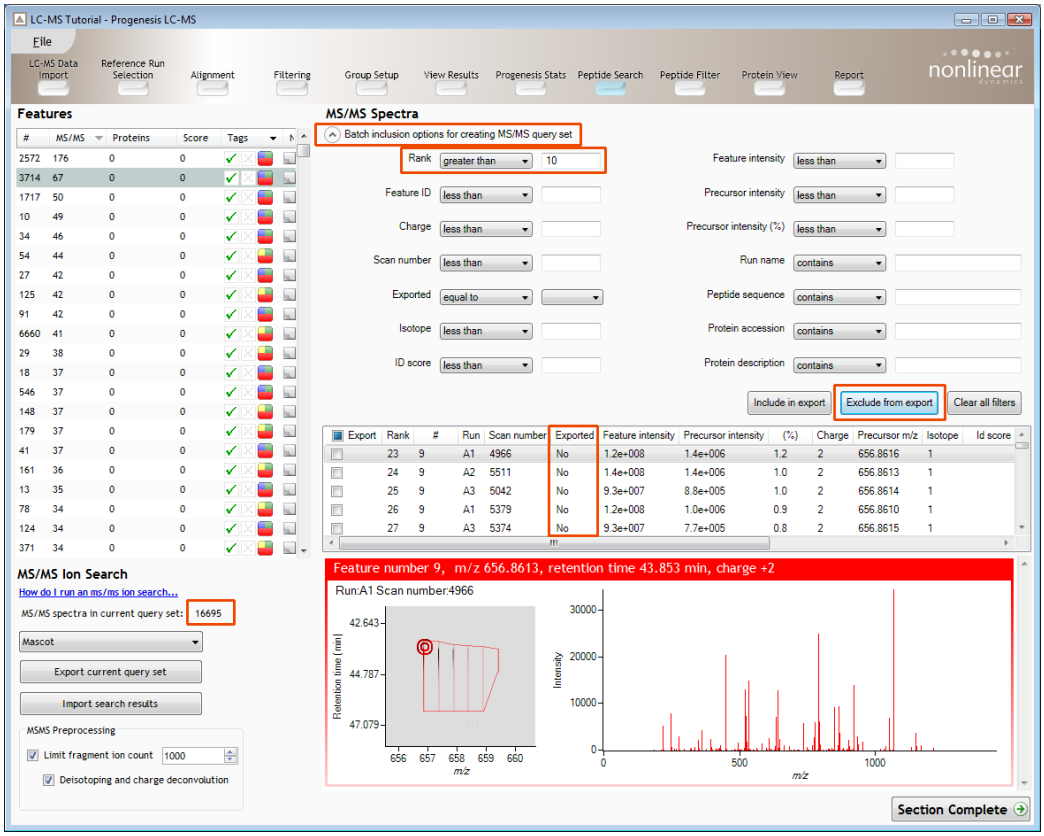
Note: by default the table is ordered on the number of MS/MS spectra available for each feature.

The total number of spectra included in the query set for this data set is 20873

Before exporting peak lists, the query set can be further refined.

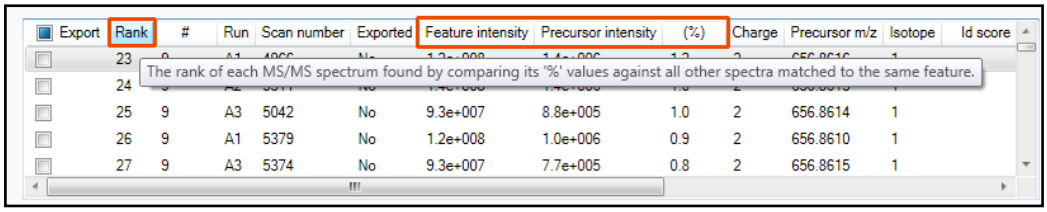
Note: many features have a large number of spectra associated with them.

To control the number of spectra for each feature, expand the **Batch inclusion options** to reduce the size of the query set.



For example: We will make use of the 'Rank' value to reduce the number of Spectra being used for each feature in the query set to a maximum of 10.

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

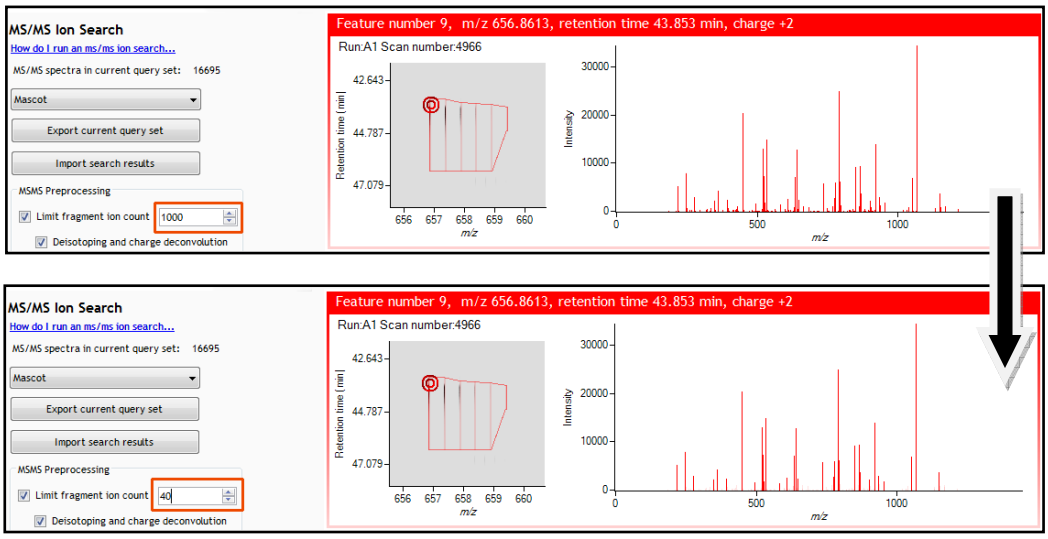


Note: the **% value** for each spectra is the **Precursor intensity as a percentage of the Feature intensity**

Set the Rank filter to 'greater than' 10 and click **Exclude from export**. This will reduce the query set to 16695.

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

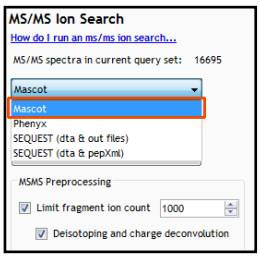
For example for the current spectra, reduce the FIC from 1000 to 40.



Note: the effect this has on the number of peaks in the spectra. This 'limitation' is applied to all the spectra being exported.

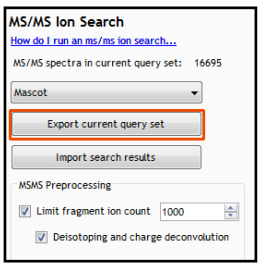
For this tutorial we will limit the fragment count to 1000.

Performing an MS/MS Ion Search



Having filtered the query set to 16695 spectra 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

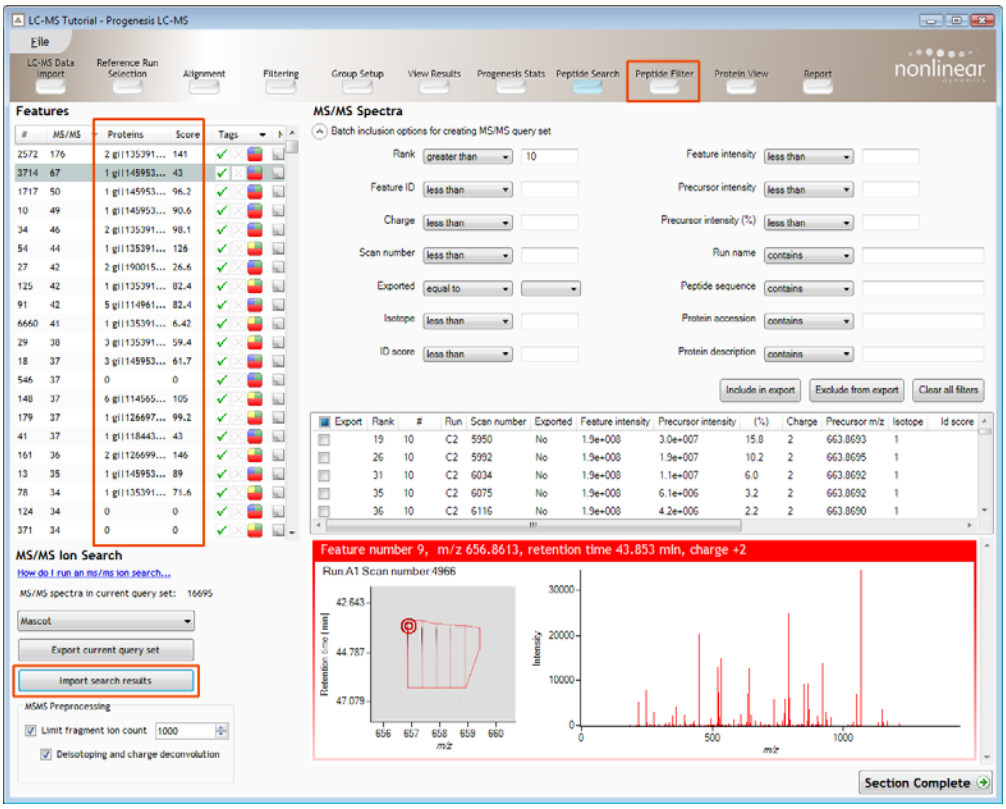


Please refer to Appendix 5 (a and b) (pages 56 and 57) for details of the 'Search Engine' parameters

Note: the blue link tells you the appropriate formats for exporting ID results

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_v2.5.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, **Peptide Filter**.

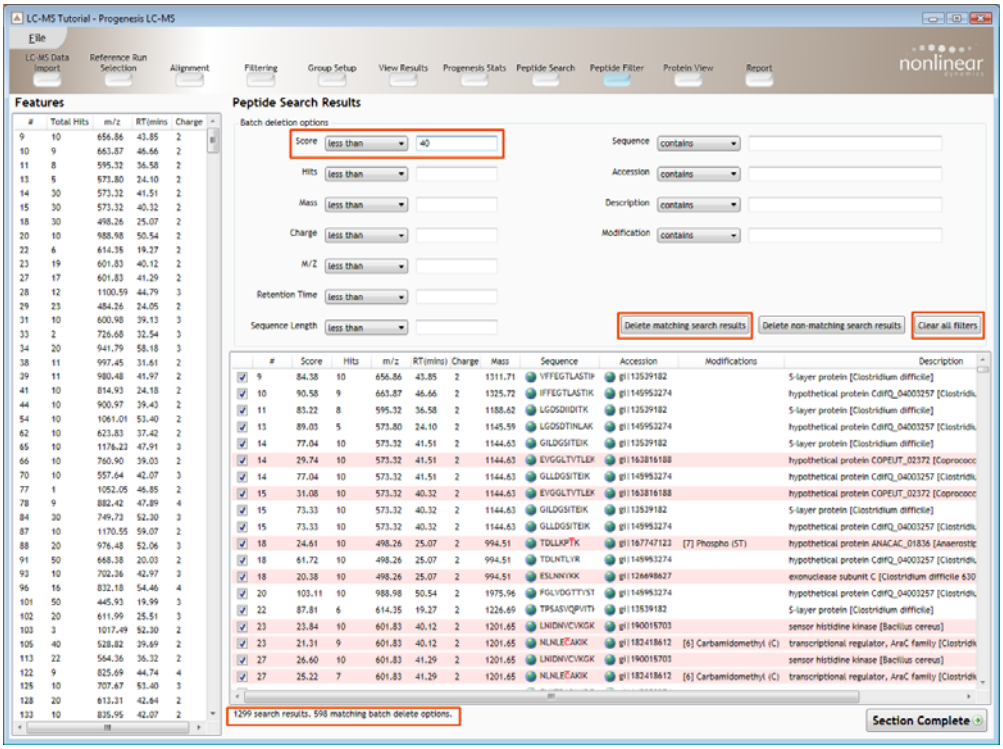
Stage 10: Peptide Filter

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

In this tutorial as an example Acceptance Criteria on which to base the sequential filtering of the Peptide results, the following thresholds can be applied:

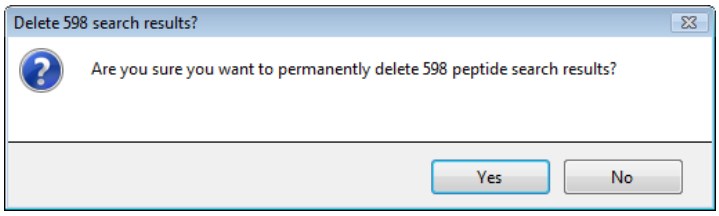
- Remove identifications with a Score less than 40
- Remove identifications where less than 2 hits were returned
- Remove all identifications where the Protein Description **Contains** 'hypothetical'
- Remove all identifications where the Protein Description **Doesn't contain** 'Clostridium difficile'

Set the Score to less than 40, then **Delete matching search results**.



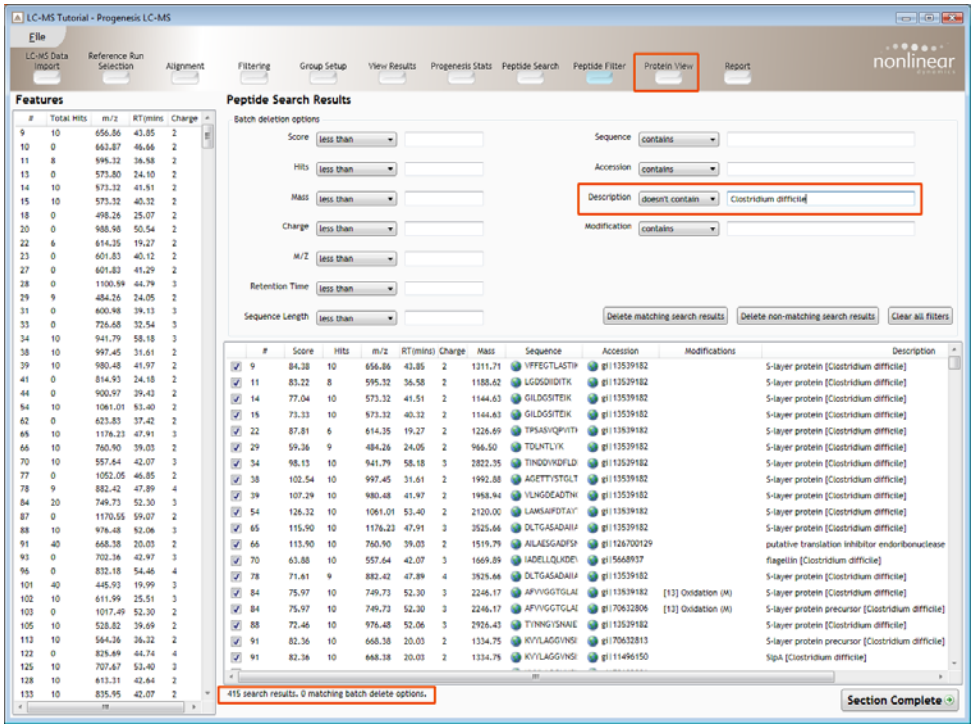
Note: the search results matching the filter criteria turn pink and the total is displayed at the bottom of the table (598 matching out of 1299)

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 39)

Having applied all 4 filters the **Peptide Search Results** should be reduced to 415.



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

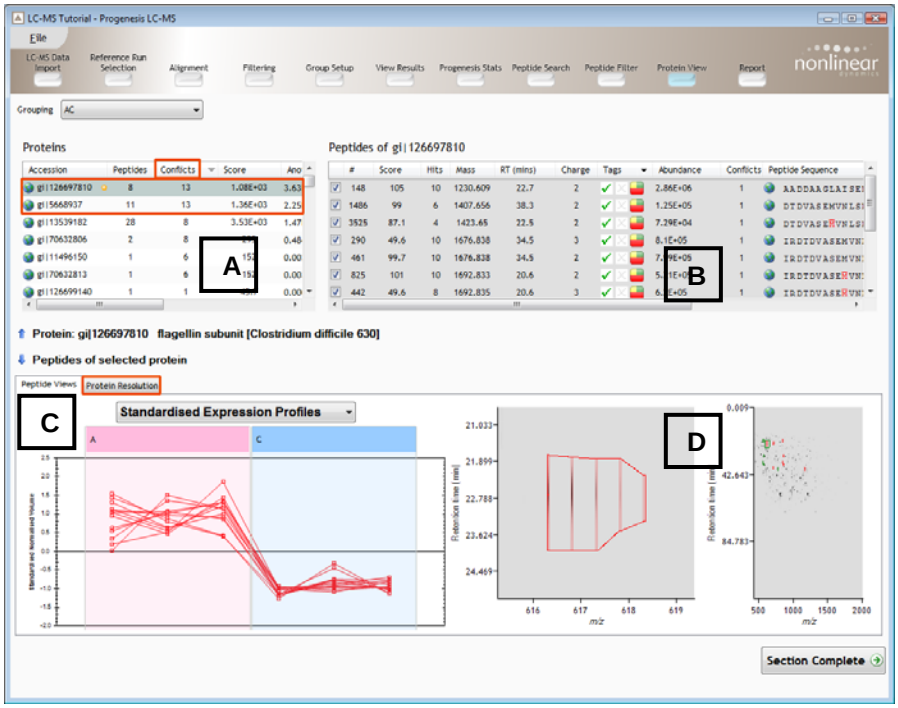
Stage 11: Protein View

The Protein View combines the quantitative LC-MS data with the qualitative MS/MS results at the protein level, highlighting proteins of interest between experimental groups. This stage allows you examine the behaviour of the assigned peptides and resolve any conflicts for the various peptide assignments at the protein level.

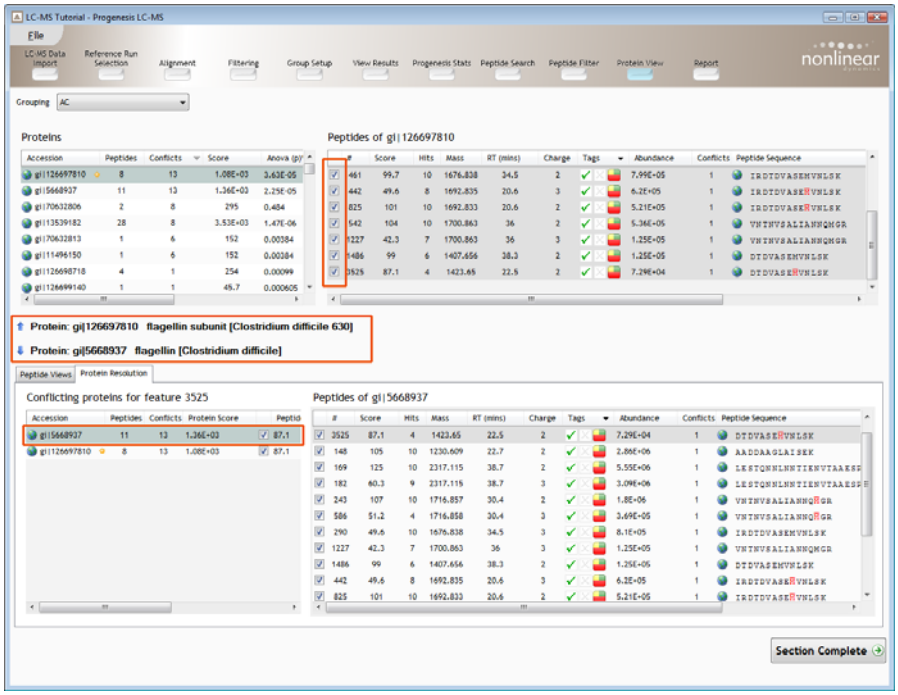
The Protein view provides a number of interrelated graphical and tabular views to assist you in the validation of the peptides that have been assigned to proteins and also to review the relevance of the data returned from the search.

When you open the Protein View order the data in the Proteins table (A) on the basis of **Conflicts**.

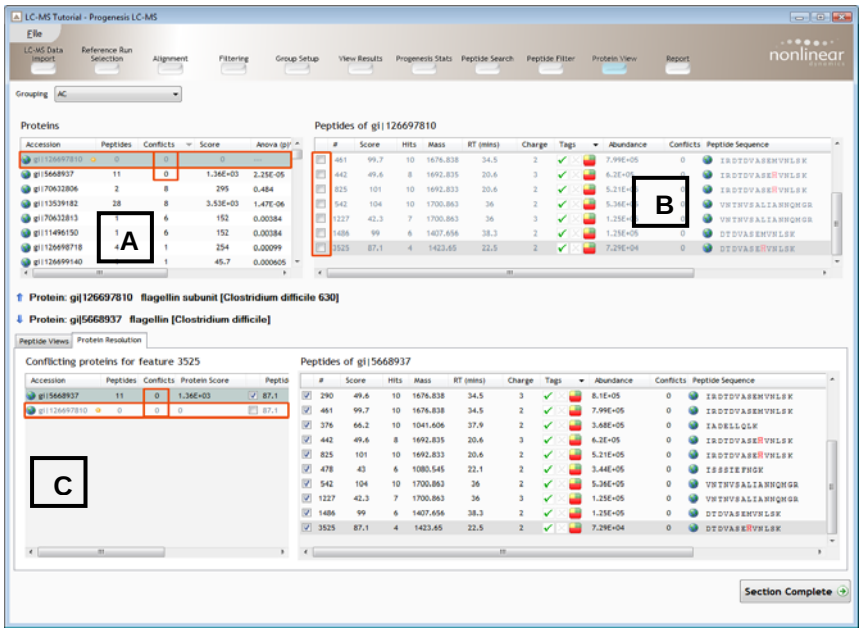
Note: the look of the tables (with regards to ordering) in the following section may vary slightly.



Depending on the ordering, make **'flagellin subunit'** the current protein by clicking on it in Window A (a circular orange symbol indicates current protein). **Flagellin subunit** has 8 peptides assigned (window B) which have a total of 13 conflicts. To view the conflicting assignments click on the **Protein Resolution** tab (window C) and then step through each assignment on window B.

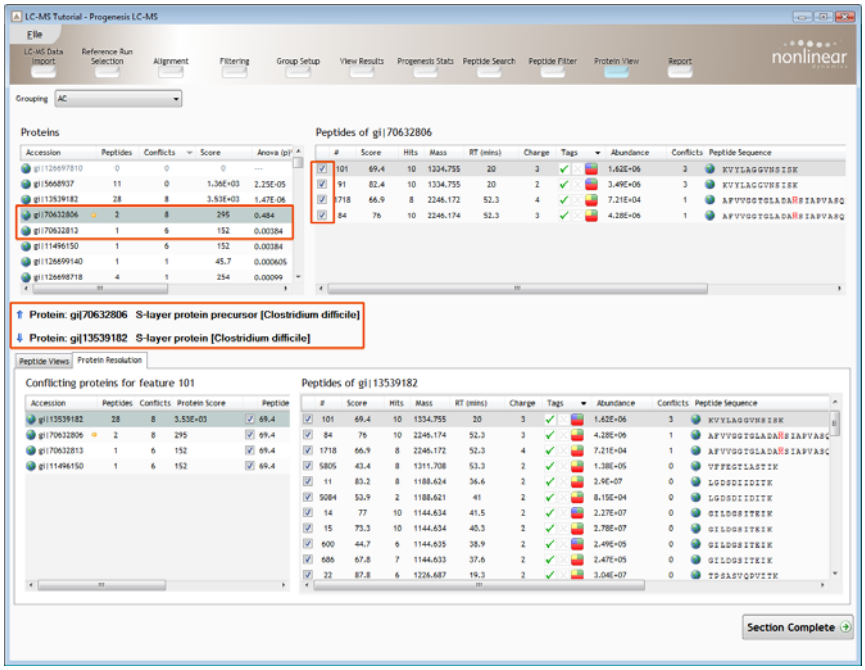


In this case the conflicting peptide assignments are with the main **flagellin** protein which contains 3 additional peptides. To resolve this conflict un-assign all the peptides assigned to the **flagellin subunit** protein, by selecting and then unticking all the peptides in window B.



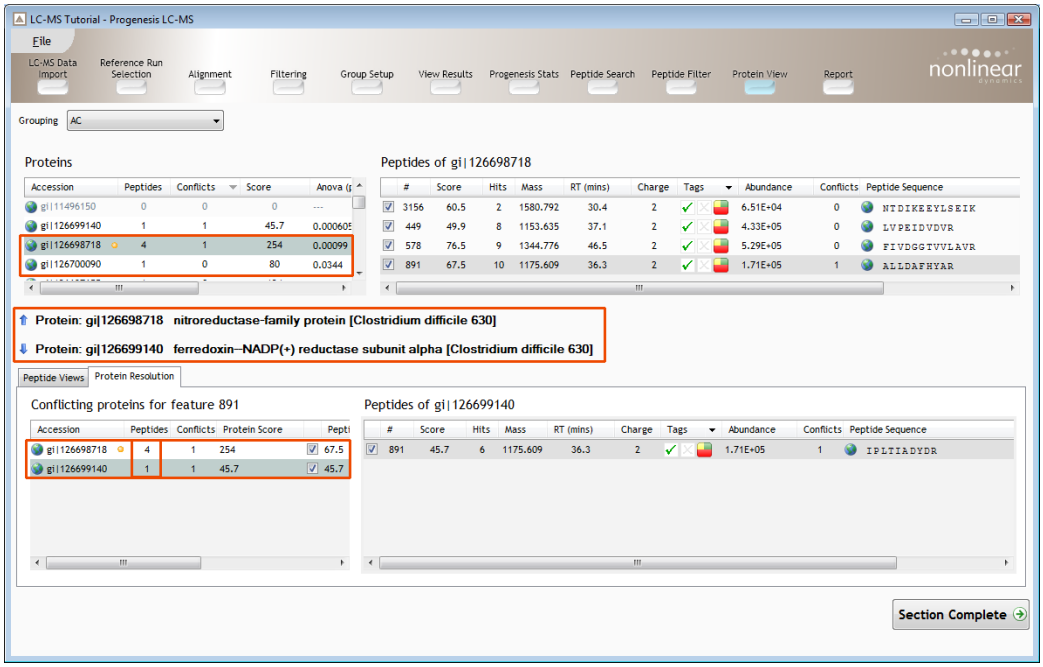
Note: as you un-assign the peptides the number of conflicts update 'on the fly' in all the windows

A similar argument can be applied to the next set of conflicting assignments



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

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 4 peptides.

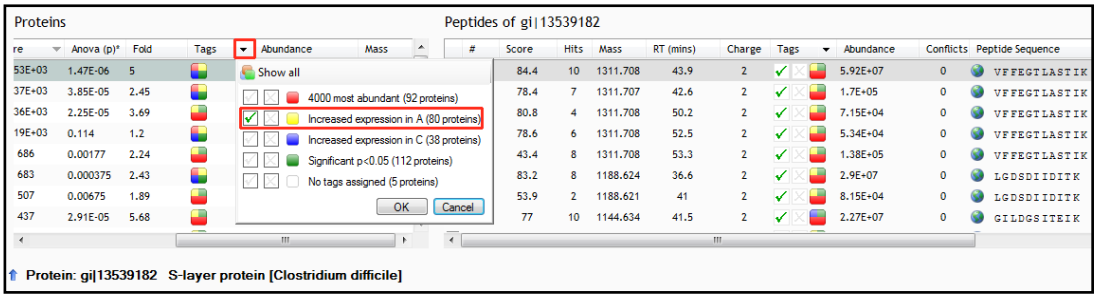
Note: the number of conflicts you have to resolve will depend on the scope and stringency of the filters you apply at the **Peptide Search** stage

For this tutorial use of the 'suggested' **Peptide Filter** criteria, (page 39), results in a low number of conflicts requiring 'subjective' resolution. A less stringent application of filters at the Peptide Filter stage will increase the time spent resolving conflicts.

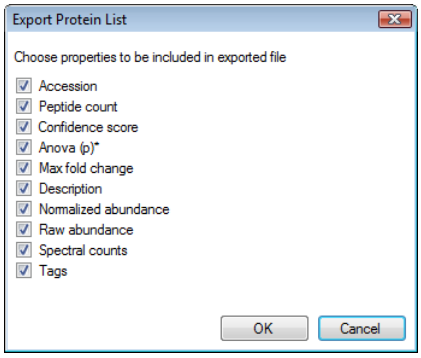
Finally order the Protein table (A) using descending score, and then scroll to the right to locate the ‘tags’ column.

You can now select proteins on the basis of the tagged features.

For example you can filter the list to show only these proteins that contain features with **Increased expression in A**



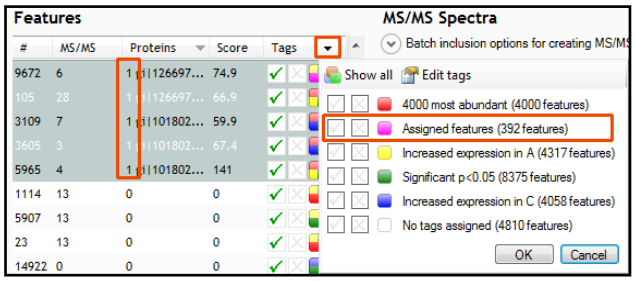
This will filter the Protein list so that it now only displays the 80 proteins containing peptides that show I**ncreased expression in A**



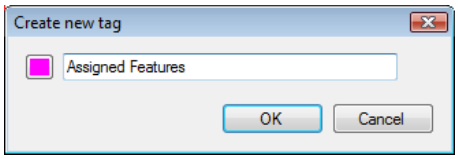
You can export this filtered Protein list (csv format) by selecting this option from the **File** menu. You can control the data output required, using the dialog provided.

Now return to the **Peptide Search** stage by clicking on the icon in the **Workflow** at the top of the screen.

Note: in the Features list, if you have resolved all the conflicts, there will only be one protein assigned to each feature.



Create a new tag for the selected features and call it **Assigned Features**



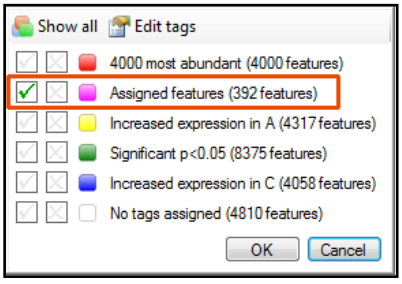
Now move to the Report section by clicking on **Report** icon on the workflow at the top of the screen.

Stage 12: Reporting

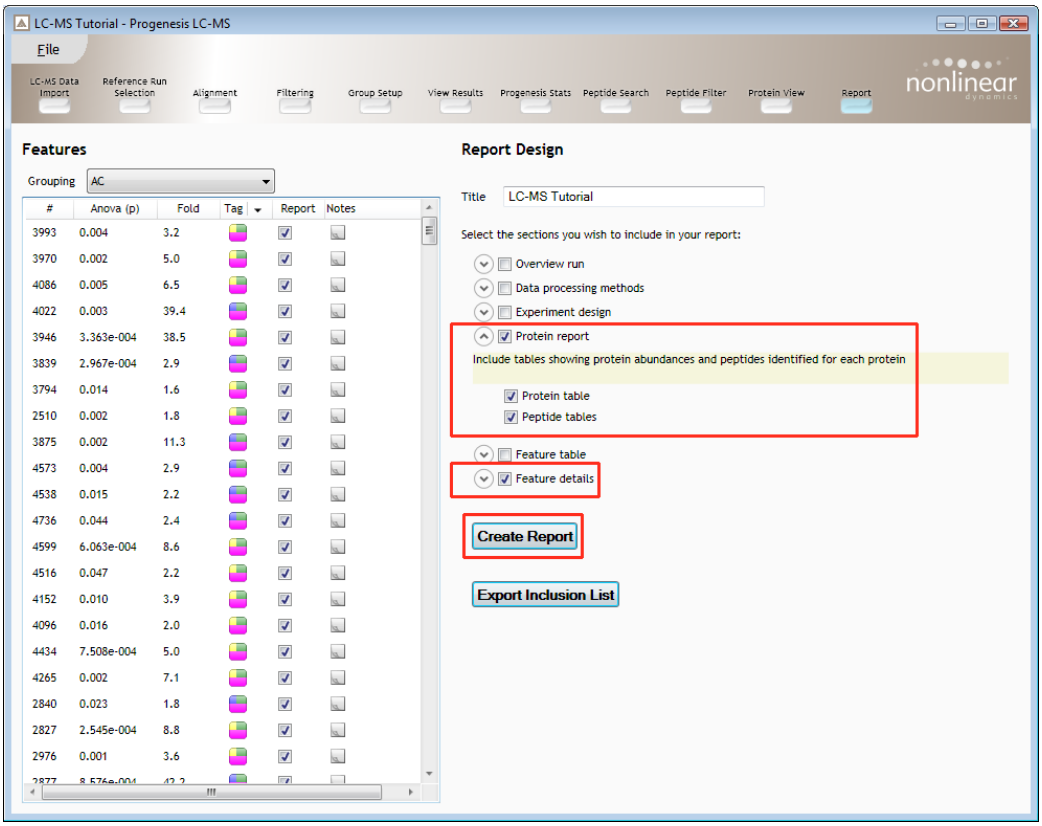
The **Report Design** stage allows you to select what views you want to include in a report based on the list of **currently selected features.**

As an example we will create a report for **only** the features with identified proteins and showing an 'Infinite fold' difference between the groups AC.

1. First reduce the features to report on by selecting the 'Assigned Features' tag. In this example it reduces the number of features in the table to 392.
2. Expand the various Report Design options (by default they are all selected)
3. Un-tick as shown below

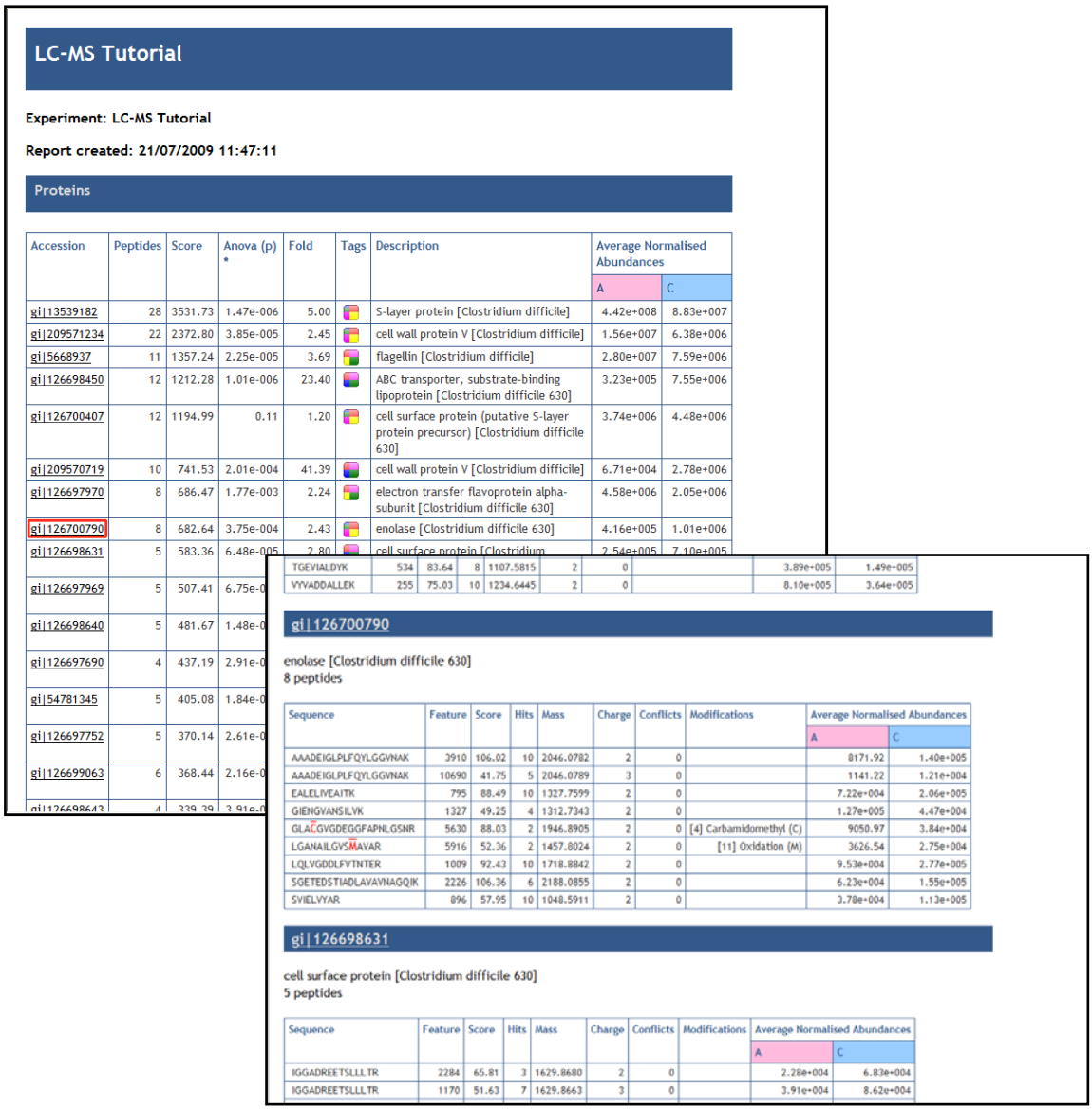


4. Click **Create Report**



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



LC-MS Tutorial

Experiment: LC-MS Tutorial

Report created: 21/07/2009 11:47:11

Proteins

Accession	Peptides	Score	Anova (p) *	Fold	Tags	Description	Average Normalised Abundances	
							A	C
gi|13539182	28	3531.73	1.47e-006	5.00		S-layer protein [Clostridium difficile]	4.42e+008	8.83e+007
gi|209571234	22	2372.80	3.85e-005	2.45		cell wall protein V [Clostridium difficile]	1.56e+007	6.38e+006
gi|5668937	11	1357.24	2.25e-005	3.69		flagellin [Clostridium difficile]	2.80e+007	7.59e+006
gi|126698450	12	1212.28	1.01e-006	23.40		ABC transporter, substrate-binding lipoprotein [Clostridium difficile 630]	3.23e+005	7.55e+006
gi|126700407	12	1194.99	0.11	1.20		cell surface protein (putative S-layer protein precursor) [Clostridium difficile 630]	3.74e+006	4.48e+006
gi|209570719	10	741.53	2.01e-004	41.39		cell wall protein V [Clostridium difficile]	6.71e+004	2.78e+006
gi|126697970	8	686.47	1.77e-003	2.24		electron transfer flavoprotein alpha-subunit [Clostridium difficile 630]	4.58e+006	2.05e+006
gi|126700790	8	682.64	3.75e-004	2.43		enolase [Clostridium difficile 630]	4.16e+005	1.01e+006
gi|126698631	5	583.36	6.48e-005	2.80		cell surface protein [Clostridium	2.54e+005	7.10e+005
gi|126697969	5	507.41	6.75e-0					
gi|126698640	5	481.67	1.48e-0					
gi|126697690	4	437.19	2.91e-0					
gi|54781345	5	405.08	1.84e-0					
gi|126697752	5	370.14	2.61e-0					
gi|126699063	6	368.44	2.16e-0					
gi|126698643	4	339.39	2.91e-0					

TGEVIALDYK	534	83.64	8	1107.5815	2	0		3.89e+005	1.49e+005
VYVADDALLEK	255	75.03	10	1234.6445	2	0		8.10e+005	3.64e+005

gi|126700790

enolase [Clostridium difficile 630]
8 peptides

Sequence	Feature	Score	Hits	Mass	Charge	Conflicts	Modifications	Average Normalised Abundances	
								A	C
AAADEIGLPLFQYLGGVNAK	3910	106.02	10	2046.0782	2	0		8171.92	1.40e+005
AAADEIGLPLFQYLGGVNAK	10690	41.75	5	2046.0789	3	0		1141.22	1.21e+004
EALELIVEAITK	795	88.49	10	1327.7599	2	0		7.22e+004	2.06e+005
GIENGVANSILVK	1327	49.25	4	1312.7343	2	0		1.27e+005	4.47e+004
GLACGVGDEGGFAPNLGSNR	5630	88.03	2	1946.8905	2	0	[4] Carbamidomethyl (C)	9050.97	3.84e+004
LGANAILGVSMAVAR	5916	52.36	2	1457.8024	2	0	[11] Oxidation (M)	3626.54	2.75e+004
LQLVGDDLFVTNTER	1009	92.43	10	1718.8842	2	0		9.53e+004	2.77e+005
SGETEDSTIADLAVAVNAGQIK	2226	106.36	6	2188.0855	2	0		6.23e+004	1.55e+005
SVIELVYAR	896	57.95	10	1048.5911	2	0		3.78e+004	1.13e+005

gi|126698631

cell surface protein [Clostridium difficile 630]
5 peptides

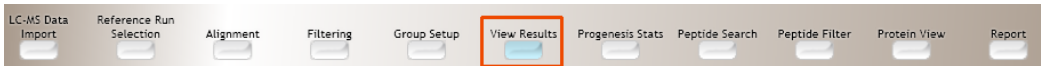
Sequence	Feature	Score	Hits	Mass	Charge	Conflicts	Modifications	Average Normalised Abundances	
								A	C
IGGADREETSLLLTR	2284	65.81	3	1629.8680	2	0		2.28e+004	6.83e+004
IGGADREETSLLLTR	1170	51.63	7	1629.8663	3	0		3.91e+004	8.62e+004

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

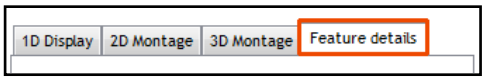
Note: you can also copy and paste all or selected sections of the report.

Creating an Inclusion list

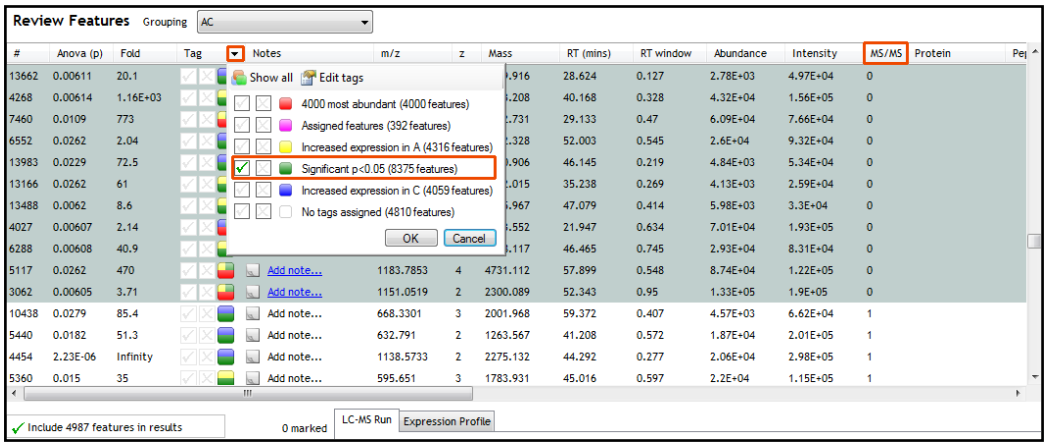
As an example of creating an **Inclusion list** you are going to create an inclusion list for all the features that show a Significant difference between groups A and C (Anova p<0.05) and have no MS/MS spectra.



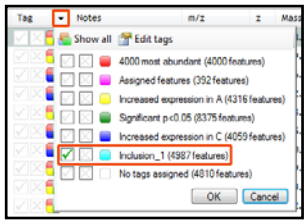
First return to **View Results** using the Workflow icons. Then select the **Feature details** tab to expand the table.



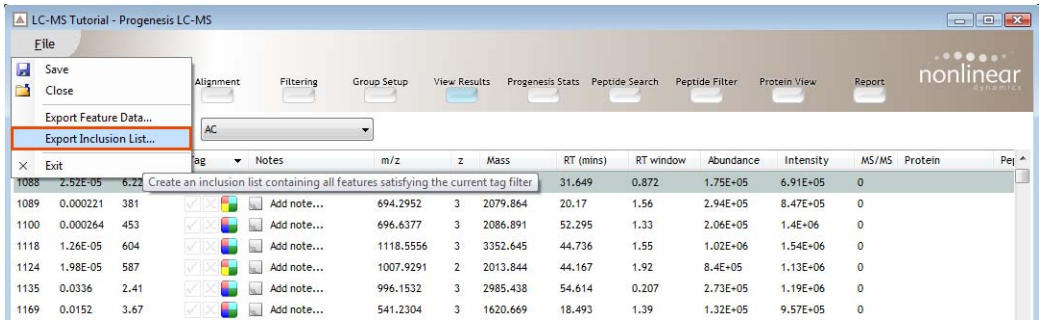
Filter the table so that it is only showing features with a **Significant p<0.05** then order the table on ascending MS/MS. Highlight all features with **No** MS/MS spectra and create a new tag for them called **Inclusion_1**.



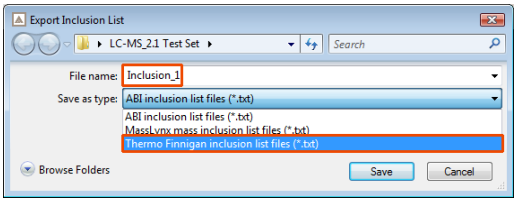
Now use the new tag to filter the table to display only those features that show a Significant Change and DO NOT have any MS/MS spectra.



Now select **Export Inclusion List...** from the file menu



Finally export the inclusion list in the appropriate MS machine format to use in the acquisition of additional MS/MS spectra from new LC-MS run(s).



Note: The new LC-MS runs can then be added to the original experiment to increase the MS/MS coverage using the **Add data 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

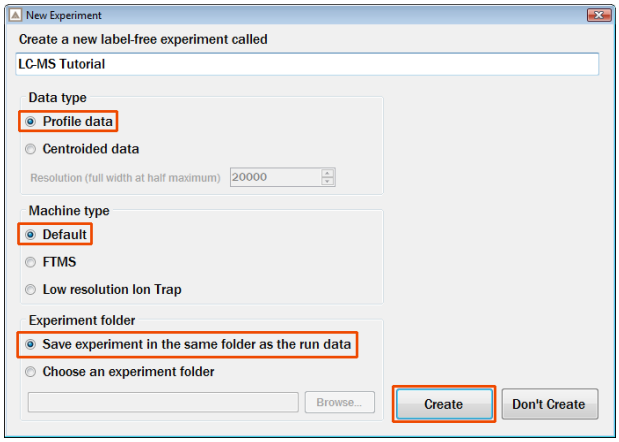
If you would like to see the benefits of running Progenesis LC-MS using your own data and explore the LC-MS Data Import module as well as the rest of the workflow please go to the next section.

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, alternatively convert them to mzXML 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'.

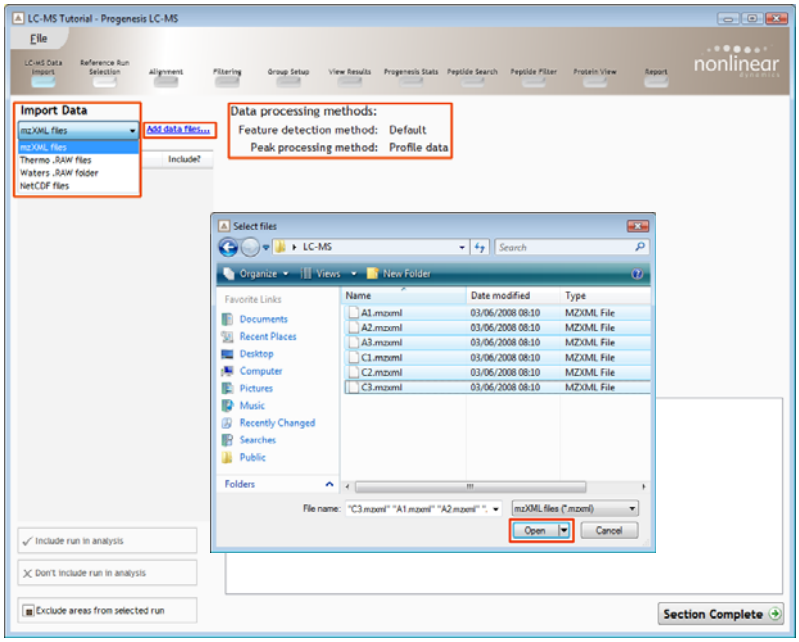
Note: if you have converted or captured the data as centroided then select Centroided data and enter the Resolution for the MS machine used.



Click Create 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 the Add data files... link.

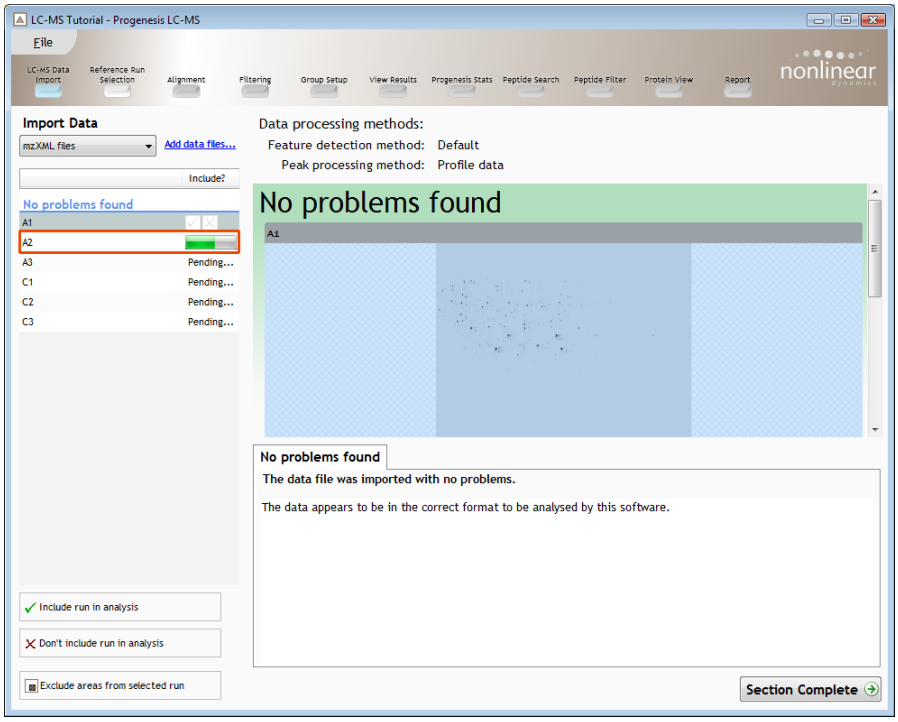


Locate and select all the Data files (A1 to C3).

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 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.

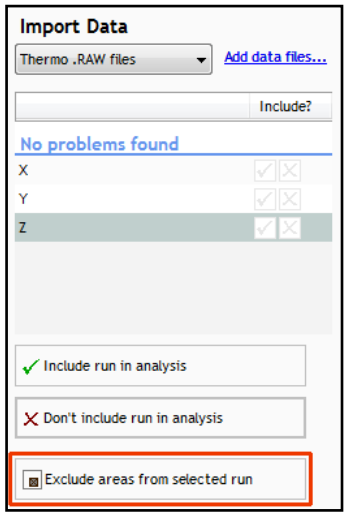


Note: as each data file is loaded the progress is reported in the **Import Data** list. The dialog below the image reports on the QC of the imported Data files. In this case ‘No problems found’ with the this data file

Now move to the next stage in the workflow (page 6 in this tutorial) 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'.

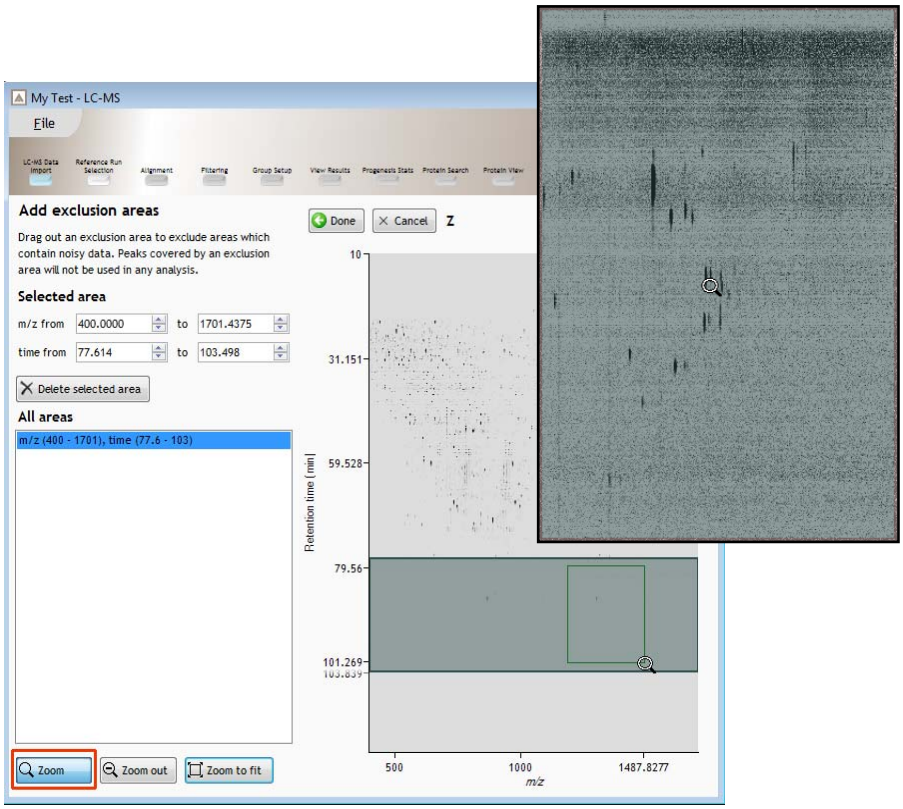


For example if the run is part of a 'replicate set' of runs it is possible to exclude such areas on the noisy run by applying a mask to the area. By doing so this area is excluded during the initial part of the detection process in order that it does not 'interfere' with the detection of the features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

To do this select Exclude areas from selected run on the bottom left of the Software.

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

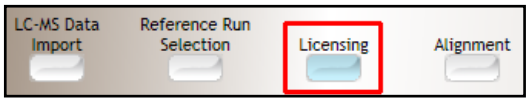
Note: if you now zoom into the masked area using the **Zoom** tool you will see the isotopic features in the noise.



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 LC-MS with unlicensed Runs then the licensing page will open after **Reference Run Selection**



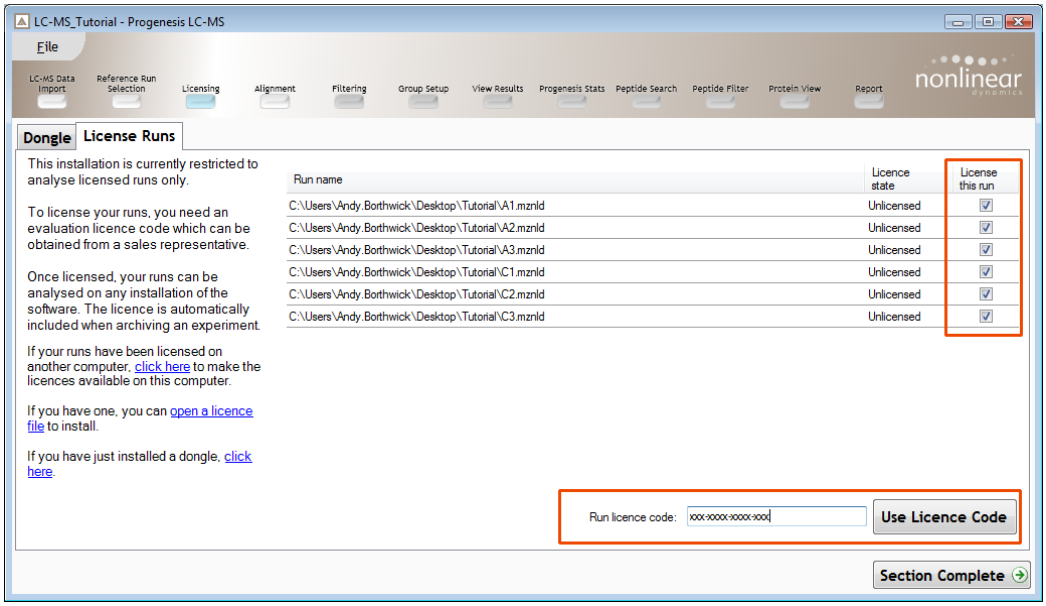
If you already have a programmed dongle attached to your machine then the following License Images page will not appear.

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

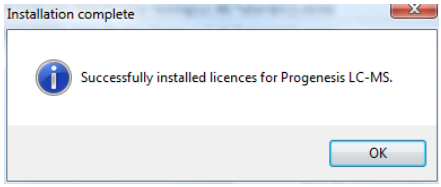
Each code will allow you to license a set number of Runs.

The Runs you wish to License will be listed as shown below.

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



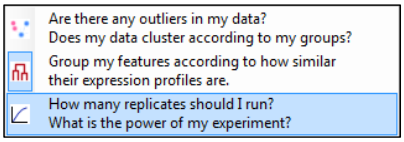
A message confirming successful installation of your image licenses will appear.



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

Appendix 4 Power Analysis (Progenesis Stats)

To explore the third Statistical analysis of the data click on the blue link Ask another question at the top of the table. The selection of 3 tools will appear in the form of questions.



Select the third option to explore the number of replicates required and obtain a measure of the 'power' of the current experiment.

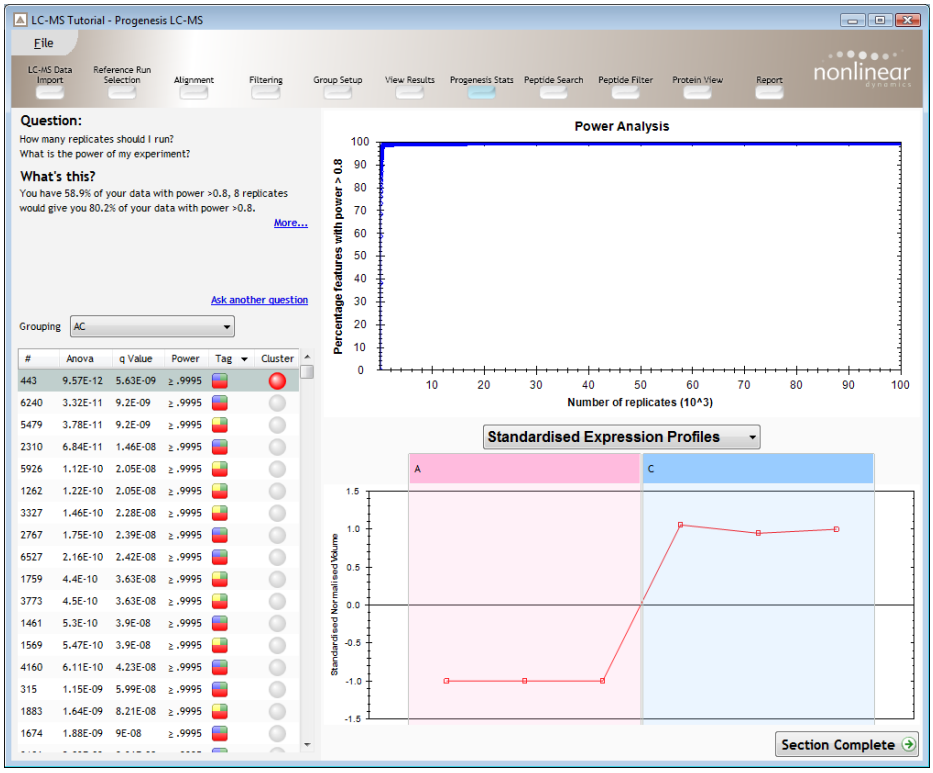
This time the statistically based question(s) being asked of the application is:

'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 all **4000 most abundant features** view the power analysis.

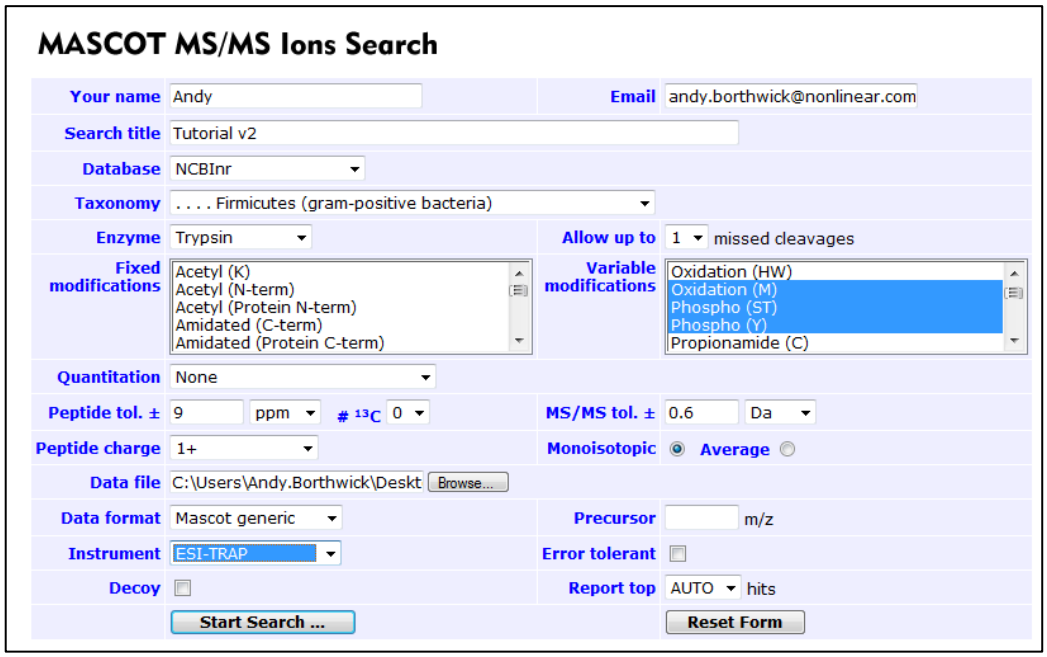


This is displayed graphically showing that 58.9% of the 4000 features have a power of 80% or more and therefore 9 replicates would be required to give you 80% of your data with power > 0.8.

- **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 5 (a): Search engine parameters (Stage 9) Mascot

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



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

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

Peptide Tol: 9ppm

Instrument: ESI-Trap

Appendix 5 (b): Search engine parameters (Stage 9) Phenyx

The parameters applied to the Phenyx search that yielded the search results used in this example tutorial 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] PHOS[variable, <=4] Cys_CM[variable, <=4]
Enzyme	Trypsin_(KR_noP) miss. cleav. 1 cleav. mode. normal
Parent tol.	0.01Da
Pept thresholds	length>=6 score>=6.0 p-value<=1.0E-6
AC Score	6.0
Conflict resolution	yes
Turbo scoring	tolerance=0.5Da coverage >=0.2 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