Analysis of DIA/SWATH data in Skyline

Note: This tutorial uses data from a Thermo Q-Exactive plus instrument. If you prefer to use the data from a Sciex TripleTOF 6600, see the version of this tutorial entitled Skyline DIA TTOF.

In this tutorial we will learn how to use Skyline to perform targeted post-acquisition analysis for peptide and inferred protein detection and quantification using a SWATH-MS dataset acquired on a QqOrbi instrument (Thermo Q-Exactive plus, Thermo) using an 18 variable width window precursor isolation scheme (Bruderer R. et al. MCP 2015) and a 1 hour gradient.

The data are from samples replicating the [LFQBenchstudy](http://www.nature.com/nbt/journal/v34/n11/full/nbt.3685.html) where quantitative benchmarking samples were created by mixing proteomes of 3 organisms in defined ratios (figure).

Initially, we will set all the parameters in the Skyline session required to work with data-independent datasets and then we will proceed to extract the quantification information from the raw data files. We will import DDA search results to create a spectral library in order to generate peptide query parameters to analyse the DIA data.



[figure adapted from

Navarro, P. et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotech 34, 1130–1136 (2016)

and

Bruderer, et al. Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues.” Molecular & Cellular Proteomics 14, (2015)]

# Getting Started

To start this tutorial, download the following ZIP file:

[zip file link]

Extract the files in it to a folder on your computer, like:

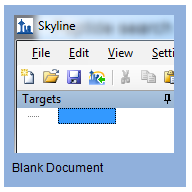
C:\Users\brendanx\Documents

This will create a new folder:

C:\Users\brendanx\Documents\MethodRefine

If you have been using Skyline prior to starting this tutorial, it’s a good idea to revert Skyline to its default settings. To do so:

* Start Skyline.
* From the shaded blue tile on the **Start Page,** click **Blank Document** which looks like this:



* From the **Settings** menu, click **Default.**
* Click **No** on the form to save current settings

The settings in this instance of Skyline have now been reset to the default.

Since this tutorial covers a proteomics topic, ensure that the user interface control is set to the “Proteomics interface”

* Click the user interface control in the upper right hand corner of the Start Page and select **Proteomics Interface** which looks like this:



Skyline is operating in Proteomics mode which is displayed by the protein icon  in the upper right-hand corner of Skyline.

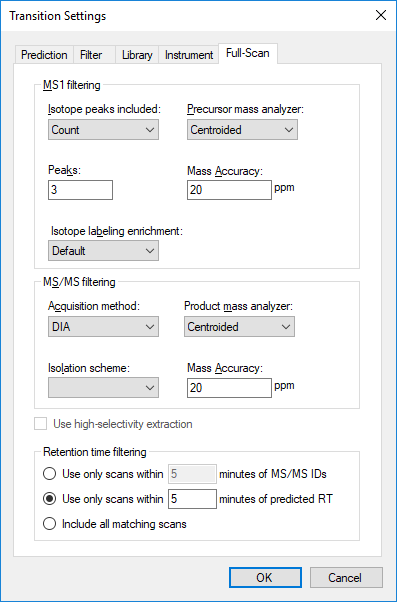
To tune for DIA analysis, we also want to be sure that all our transitions are integrated within Skyline. To do that, click **Settings** and choose **Integrate all.**

# Defining the data-independent acquisition settings and isolation scheme

We will set the parameters for extracting ion chromatograms from both MS1 and DIA-MS2 data in the following steps:

* From the **Settings** menu, click **Transition Settings** and select the **Instrument** tab.
* Change the **Max m/z** value to “1800” m/z
* In the **Transition Settings** dialog box, click the **Full-Scan** tab
* Enter the settings as indicated in the screen shot below.

**Note:** The resolving power depends on the type and settings of the instrument used for data acquisition. The optimum will be slightly different for each dataset. In this analysis we are using centroided data to save space so we will select ‘centroided’ and specify a mass accuracy for extraction. With profile mode data the resolving power of the instrument can be specified.



Now we need to define a new isolation scheme according to the parameters defined on the instrument for data-independent acquisition.

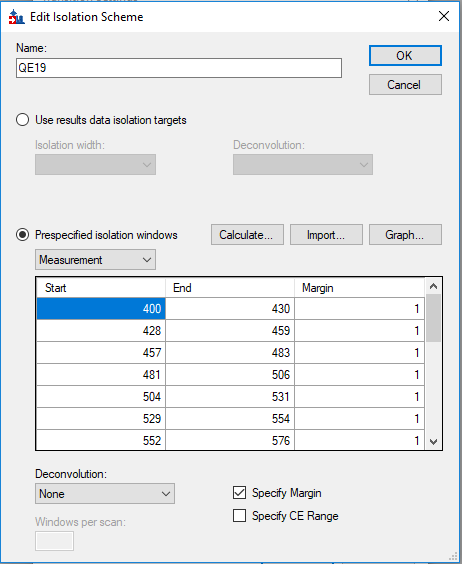
**Note:** In this example, we used 18 variable width windows that covered the range from 400 to 1200 m/z with 2 m/z overlap.

* Still on the **Full-Scan** tab of the **Transition Settings** dialog box, click the **Isolation scheme** menu and select **Add..**
* In the **Name** field of the **Edit Isolation Scheme** form, enter “QE18”.
* Click the **Prespecified isolation windows** radio button.
* At the bottom of the form, check the **Specify Margins** box.
* Ensure the dropdown under **Prespecified isolation windows** that **Measurement** is selected.
* Open the file QE\_DIA\_18var.tsv in Excel (e.g. from *C:\Tutorials\4-DIA\QE\DIA*).

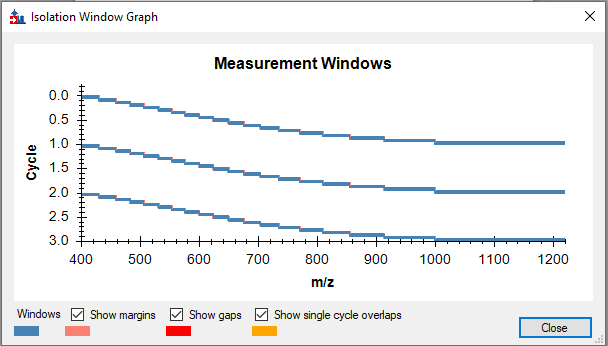
The first column of the table is the start and the second column is the end m/z of each isolation window. The third colum specifies the margin = 1

**Note:** As the quadrupole transmission windows are not perfectly square, the margin option allows to specify how much of the window edges schould not be used for extraction. Skyline will then extract from **start**+**margin** until **end**-**margin**

* Highlight the relevant columns and copy the excel table with *ctrl-c* and paste it with *ctrl-v* into the Skyline table for isolation windows.
* The **Edit Isolation Scheme** window should now look like this:



* Click the **Graph** button to see how the isolation windows cover the specified range.

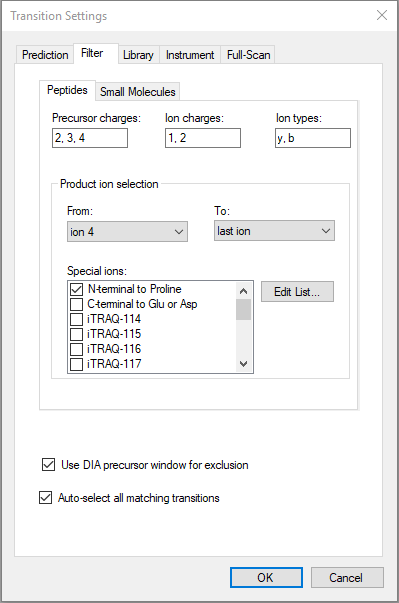


* Click **Close.**

In the **Edit Isolation Scheme** form, click **OK.**

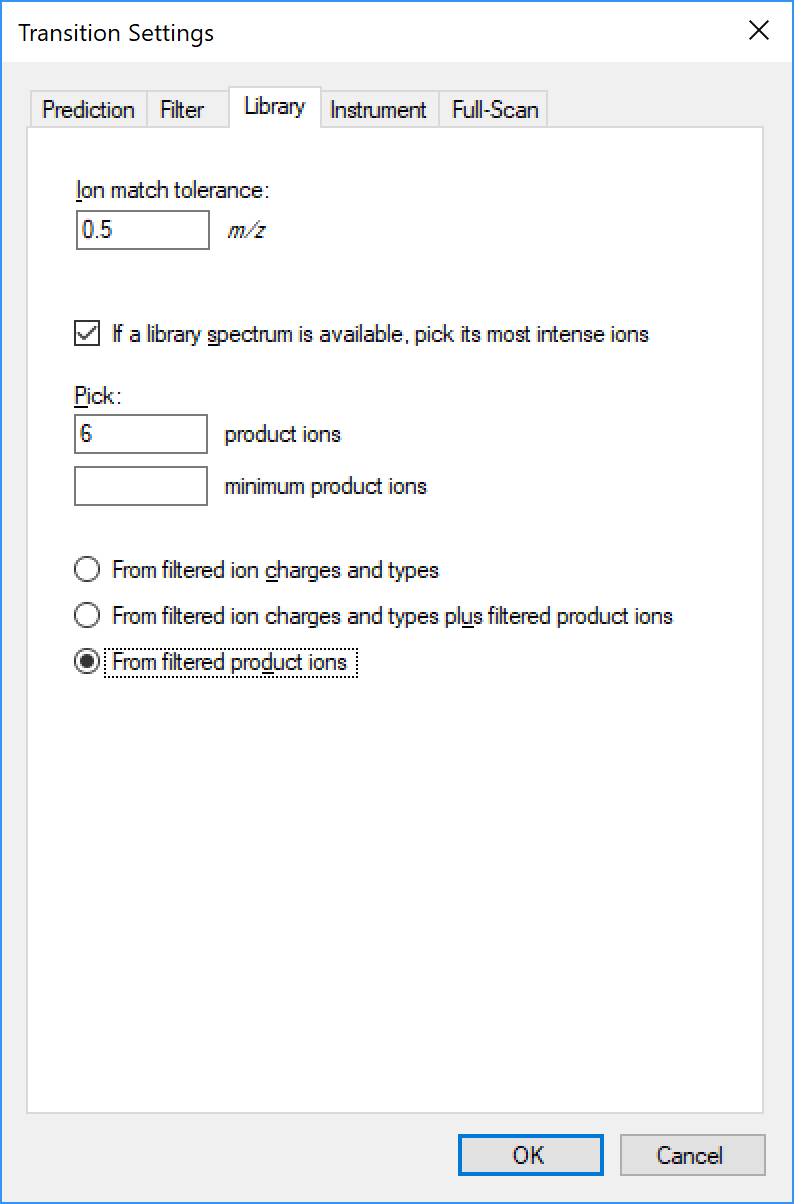
* Make sure that the Isolation scheme that you have just created is selected in the **Isolation scheme** drop down menu in the **Transitions Settings**, **Full-Scan** tab.
* Also, in in the **Full-Scan** tab, make sure that the **Retention time filtering** is set to **Use only scans within 5 minutes predicted RT**. This is used for setting a 10 minutes window around the predicted RT.
* Go to **Filter** tab of the **Transitions Settings** dialog box.

Fill in the options according to the screenshot below:



* Go to the **Library** tab (if closed, the **Transitions Settings** window is under the **Settings** in the main Skyline menu)
* Ensure that the **If a library spectrum is available, pick its most intense ions** is enabled and enter “6” in **Pick: product ions.**”
* Select the **From filtered product ions** radio button.

The **Library** tab of the Transition Settings for should look like this:



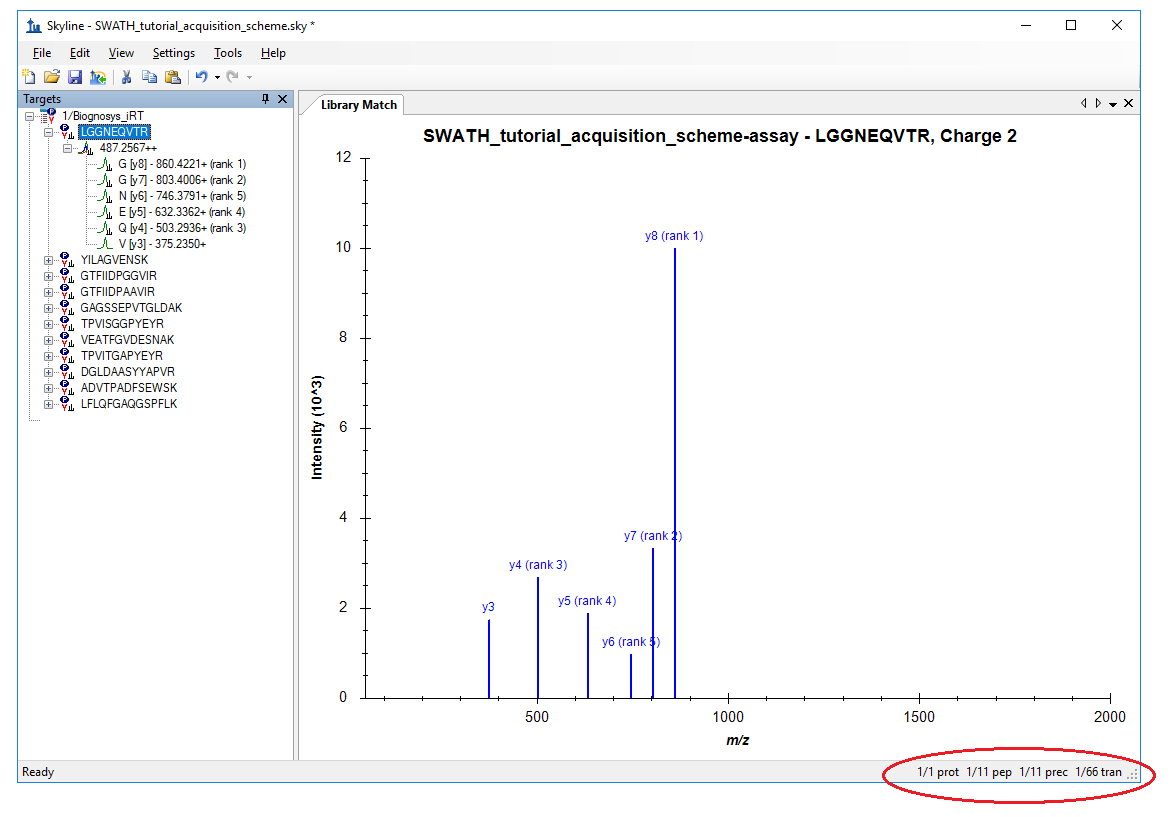
* Click **OK**.
* Save your Skyline document as *DIA\_QE\_acquisition\_scheme.sky*

# Adding query parameters for the iRT peptides

A set of 11 synthetic peptides with well characterized chromatographic behavior have been spiked into the samples for the purpose of making a linear regression of the measured retention times of these peptides and the ‘iRT values’ [see Escher, C. et al. Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics 12, 1111–21 (2012)]. This allows us to project the iRT values for the target peptides onto the retention time space of the DIA runs facilitating accurate predictions of their retention times in each DIA run we will analyze. We will first add the query parameters for the iRT peptides to the Skyline document.

* Open the Biognosys\_iRT\_for\_OpenSWATH\_6tr.tsv file in Excel or a text editor and notice the information contained.
* In Skyline, click **File** and then **Import.**
* From the fly-out menu, select **Transition List**.
* Select the file Biognosys\_iRT\_for\_OpenSWATH\_6tr.tsv and click **Open**.
* Skyline should ask if you want to create new iRT calculator. Click **Skip.** We will create the iRT calculator in a separate step later.
* Skyline should now ask if you want to create a spectral library from the spectral library intensities. Click **Create**.
* You should now see the 11 iRT peptides on the left of the document. Click on the first peptide (**LGG…**) and click on the **+** beside this peptide to expand.

You should see the precursor and fragment ion m/z and the ion type annotations as well as their intensity rank in the spectral library. You should also see the pseudo-spectrum created by Skyline from the fragment ion relative intensities in the **Library Match** window (if you don’t see the spectrum click **View** and then select **Library Match**). There should be 1 protein, 11 peptides, 11 precursors, and 66 transitions as highlighted in the bottom right corner of the screenshot below:



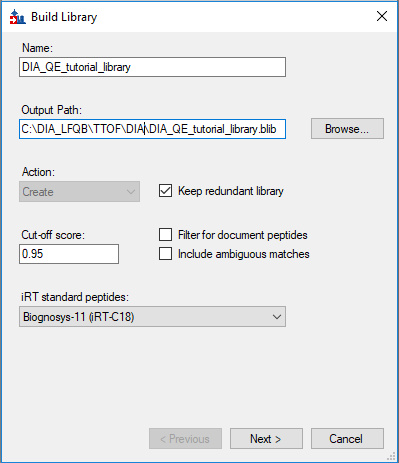
# Creating a spectral library

We will import some search results of DDA analysis of one replicate each of the A and B samples in order to generate a spectral library and peptide query parameters to query the DIA data. The DDA search has already been performed by us using the Comet search engine and post processed using PeptideProphet – see this link for more DDA search engines supported by Skyline <https://skyline.ms/wiki/home/software/Skyline/page.view?name=building_spectral_libraries> ). We will start from the interact.pep.xml file which is the output of PeptideProphet and contains the database search results from both DDA files.

* From the **Settings** menu, click **Peptide Settings** and click the **Library** tab.
* Click the **Build** button and add “DIA\_QE\_tutorial\_library” in the Name.
* You may have to click **Browse** and set the **Output Path** to the current directory.
* Set the “**Cut-off score** set to “0.95” and select **Biognosys-11 (iRT-C18)** from the **iRT standard peptides** drop-down menu

**Note:** 0.95 is the threshold applied to the PeptideProphet probability computed for every peptide spectrum match in the DDA database search – in this particular data set this corresponds to a PSM false discovery rate of 0.2% but this will differ among data sets so a score threshold to achieve the FDR you want to use should be entered here.

Your **Build Library** form should look something link this:



* Click Next

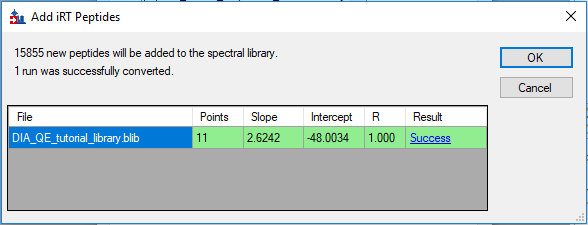
In the **Build Library** form that appears:

* Click **Add Files** and navigate to C:\DIA\_LFQB\QE\DDA\_search in your QE directory
* Clink on **interact.pep.xml** in the **Add** **Input Files** dialog box(Note: you might see **interact.pep,** if the extension is not shown) and then click **Open**.
* Click **Finish** to start the library build process. The library will take 1-2 minutes to build.

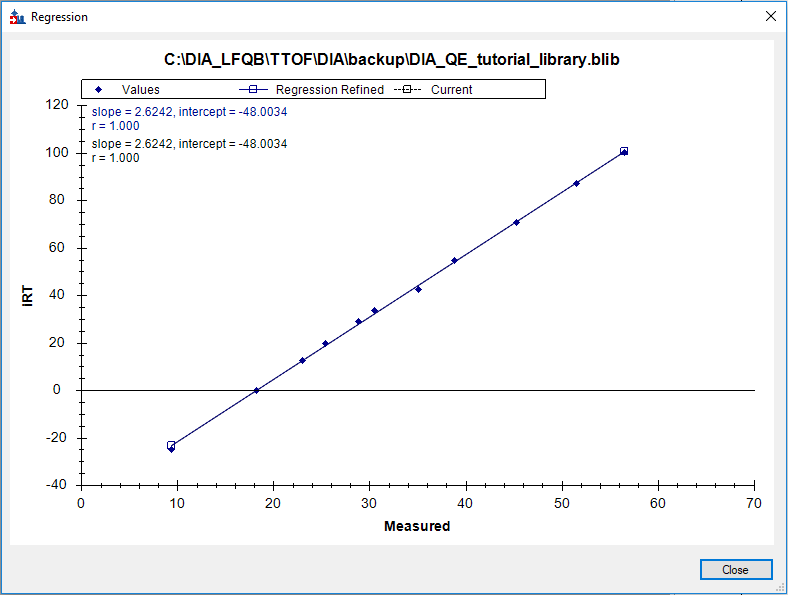
Skyline will report that “The library built successfully. Spectra matching the following peptides had multiple ambiguous peptide matches and were excluded”

* Click “OK”

Skyline should now indicate that it was able to find the iRT peptides and could make a linear regression to generate the iRTs fort the peptides added to the library



* Click on **Success** (highlighted in blue) to see the linear regression

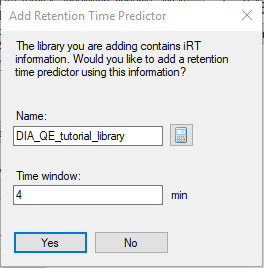


* Click **Close** and **OK.**

Skyline will ask if you want to recalibrate the iRT standard values.

* Click **No.**

Skyline should ask if you want to add a **Retention Time Predictor.**



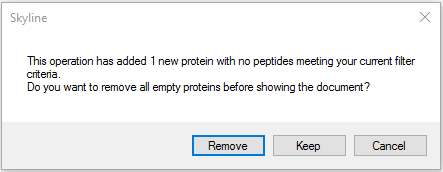
* Change the time in **Time window** to “4” minutes and click **Yes**, then click **OK** to close the form.
* Click **OK** to close Peptide Settings.
* You can visualize the spectral library by selecting **Spectral libraries** from the **View** menu.
* Save the file as “DIA\_QE\_library.sky.”

# Adding Targets

We will add 12 target proteins (4 from each species) to extract from the DIA data by adding the protein sequences in FASTA format to Skyline.

* From the **File** menu, select **Import** and then **FASTA**.
* Select the target\_protein\_sequences.fasta file from the C:\DIA\_LFQB\QE\DIA directory.

Skyline should show you the following indicating that for one of the proteins in the fasta file no peptides were present in the spectral library.



* Click **Remove**.

You should now have a document containing 12 proteins, 148 peptides, 178 precursors, and 1,044 transitions as indicated in the lower right hand corner.

# Adding Decoys

The peptide query parameters only contains target peptides and proteins. Decoys need to be generated in the next step.

* From the **Refine** menu, select **Add Decoys.** 
  + Leave the default number of decoy precursors which is **137**. This is the same as the number of precursors in the document not counting the iRTs.
  + Select **Shuffle Sequence** from the **Decoy Generation Method** dropdown menu.
  + Click **OK.**
* Save the file as “DIA\_QE\_targets.sky”

# **Performing the DIA data analysis**

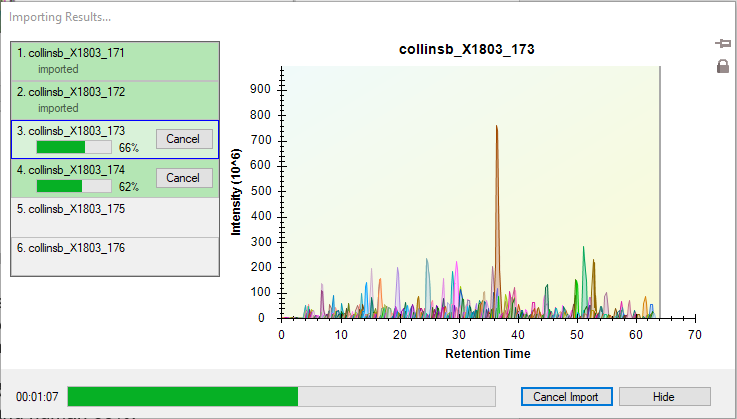
* From the **File** menu select **Import** and then **Results**.
* In the **Import results** form, choose **Add single-injection replicates in files**.
* At the bottom of the form, select **Many** from the “**Files to import simultaneously** menu
* Click the **Show chromatograms during import** checkbox.
* Click **OK** to close the form.
* In the **Import Results Files**, drag over and select all 6 DIA files from *C:\DIA\_LFQB\QE\DIA*
  + collinsb\_X1803\_171.mzXML
  + collinsb\_X1803\_172.mzXML
  + collinsb\_X1803\_173.mzXML
  + collinsb\_X1803\_174.mzXML
  + collinsb\_X1803\_175.mzXML
  + collinsb\_X1803\_176.mzXML
* Click **Open**.

Skyline should ask if you want to remove the file name prefix.

* Click “**Do not remove.**”
* Click **OK**

The next tasks in Sample Annotation, can be performed while the data is importing.

Note: The DIA data should start importing and the target and decoy transitions are extracted. This process can take some time (~5 minutes). For this tutorial, we selected data that was already converted from the raw data format to mzXML and centroiding was performed (using msconvert with vendor centroiding). Using centroided data increases processing speed compared to profile data. Furthermore we use a small peptide query parameter set to make the processing time manageable for the course. In a lot of applications larger peptide query parameter sets are chosen resulting in a longer time for the data extraction process.



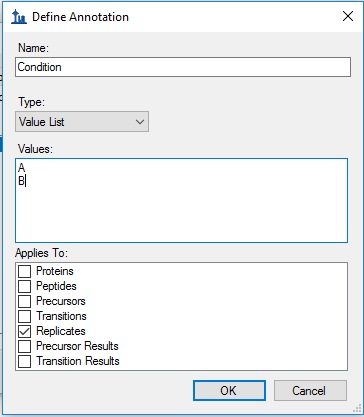
# Sample Annotation

Here we need to define which samples belong to which experimental group

* From **Settings,** select **Document Settings** and then choose **Annotations** tabto bring up a form to add one annotation to the documents.
* Click **Add** to see the **Define Annotation** window.
* In the **Name** field, enter “Condition.”
* In the **Type** field, choose **Value List.**
* In the list under **Applies To,** choose **Replicates.**

In our experiment we have two conditions: Condition A are samples that have chimeric proteome composition of E.coli 20%, yeast 15% and human 65%, and condition B with composition of E.coli 5%, yeast 30% and human 65%.

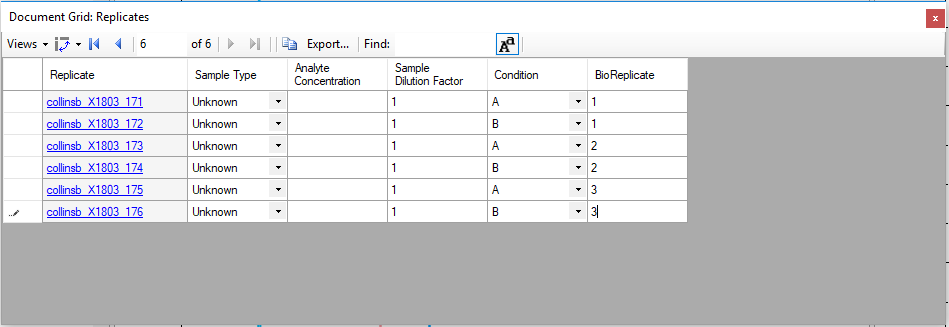
The define Annotation window should like this:



* Click “OK”
* Follow the steps above and create a second annotation named “BioReplicate.”
* Select **Test** in the **Type** menu for the BioReplicate annotation.
* Click **OK**.

Skyline will now return you to the **Document Settings** window and be sure you are still looking at the **Annotations** tab.

* Click the checkboxes to select **Condition** and **BioReplicate.**
* Click **OK.**
* From the **View** menu, choose **Document Grid** which brings up the **Document Grid: Proteins** window.
* In the **Document Grid: Proteins** window, click the **Reports** menu and from the drop-down menu select **Replicates.**
* Annotate the samples as shown in the screen capture below:



* Close the **Document Grid: Replicates** window by clicking on the red X in the upper right hand corner of the form.

For easy viewing we will split the data by condition into 2 windows.

* From the **View** menu, choose **Arrange Graphs** and from the submenu, **Grouped**.
* In the **Arrange Graph Grouped** dialog box, enter “2” for **Group panes**
* Click the radio button for **Distribute graphs among groups**
* Under the **Display** heading, choose **Tiled.**
* The **Sort order** option should read: **Document.**
* Click **OK.**

Note: If the data is still importing at this stage you will need to wait until this is finished to save the document.

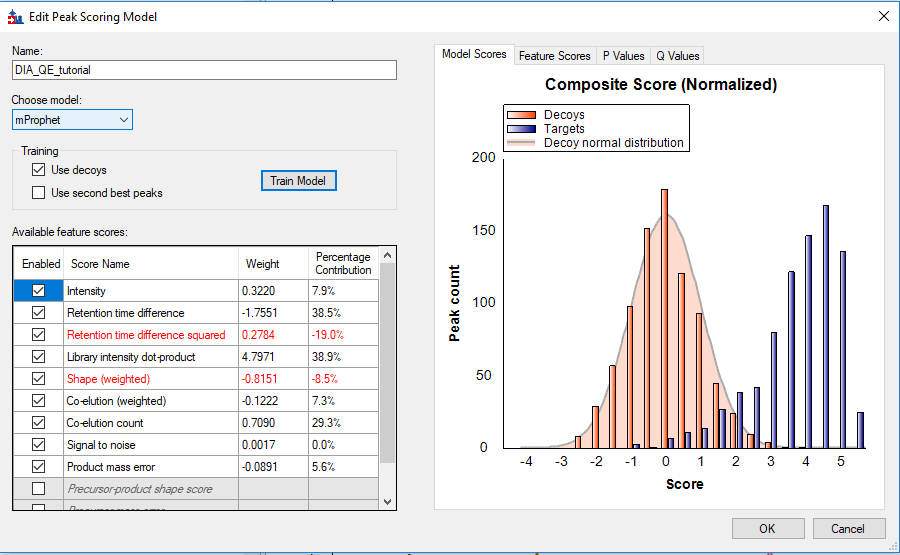
* Save your skyline document as *DIA\_QE\_extracted.sky*

# mProphet

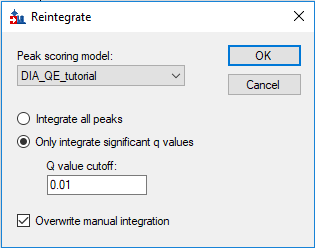
In this section, we will apply the mProphet algorithm to the data. The best scoring peak groups are selected and the q-values are calculated to enable FDR control.

* From the **Refine** menu, choose **Reintegrate**
* In the **Reintegrate** form that appears, select: **Add** in the **Peak scoring model** dropdown.
* Click **Train Model** and inspect the model score distributions
* Enter “DIA\_QE\_tutorial” in the **Name** Field.

The window should look something like this (the score distribution and weights might be slightly different depending on the Skyline version or minor difference in extraction parameters):



* Inspect the target and decoy discriminant score distributions in the “Model Scores” tab.
* Click on the **Feature Scores** tab and then select each of the scores in the table one by one. Notice which scores are best at discriminating between targets and decoys.
* Inspect the **P Values** and **Q Values** tabs. Are the majority of the targets detected?
* Click **OK.**
* Back in the **Reintegrate** window select the following:



* Click **OK.**
* From the **Settings** menu choose **Integrate all** to quantify using all transitions
* Save your skyline document as *DIA\_QE\_mProphet.sky*

# Inspect the data manually

Now we will manually inspect some of the chromatography and underlying spectra. To do that, we need to be sure that the **Retention Times** and **Peak Area** views are showing:

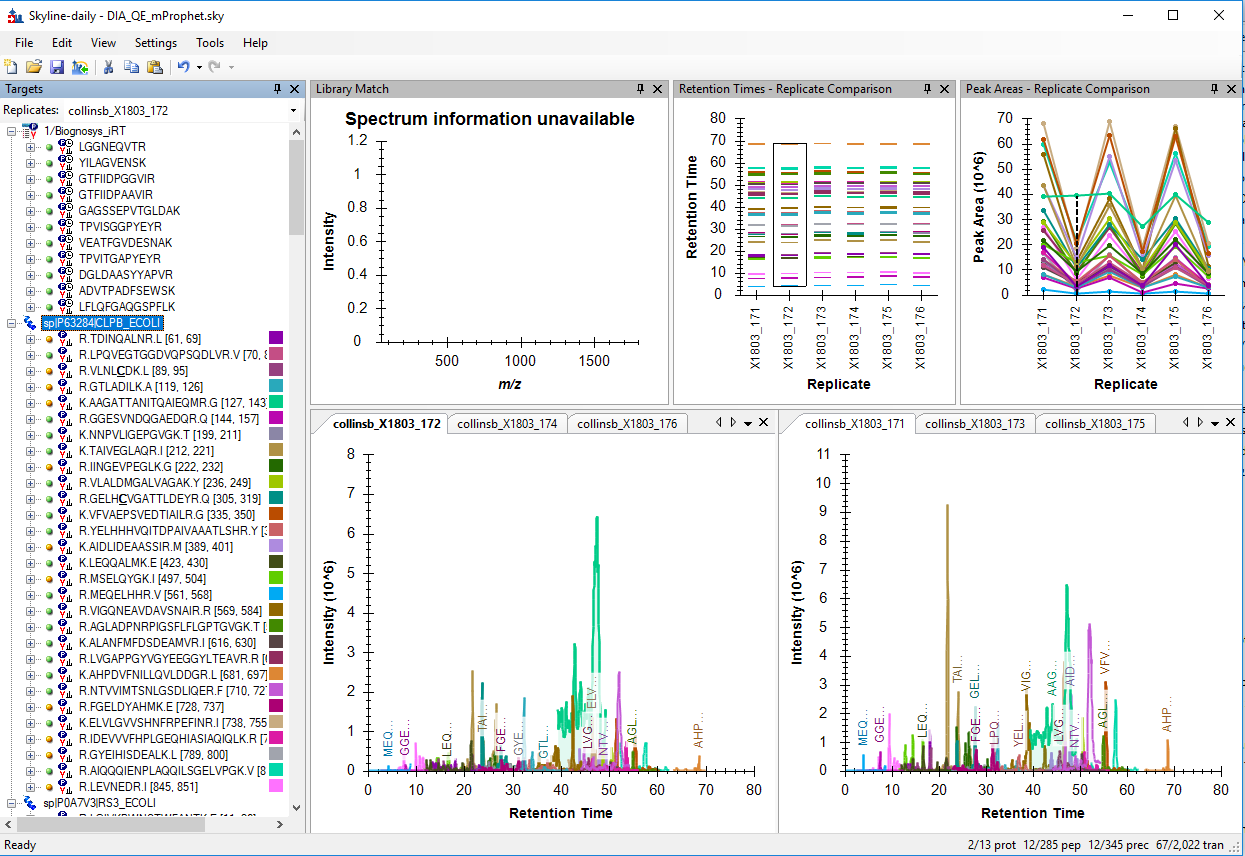
* From the **View** menu click **Retention Times** and then choose **Replicate Comparison.**
* Again from the **View** menu, click **Peak Area** and **Replicate Comparison** from the submenu.

Now both views should be visible. Let’s configure them optimally on the screen:

* If these new windows are floating you can dock them by clicking on the top border of the floating window, holding the left mouse button down, and dragging this window:
* You can also similarly dock the **Library Match** window as shown below.



* Similarly, dock the **Peak Area: Replicate Comparison** and **Library Match** windows so that all information is easily viewable, as below:

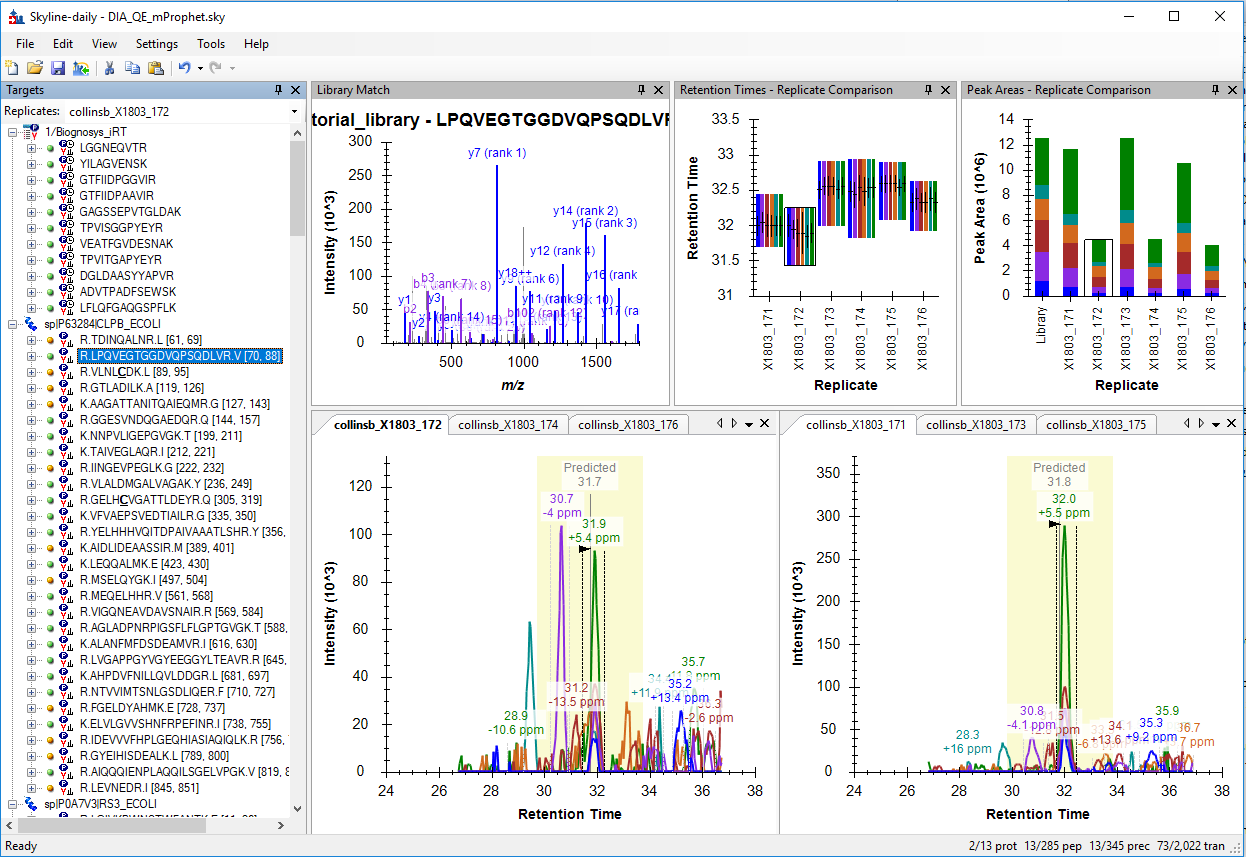


* Click on the first E. coli protein: **sp|P63284|CLPB\_ECOLI**.

You should see all of the peptides for this protein shown on the various plots (XIC chromatrography, peak areas, and retention time replicate graphs). The screenshot above is an example of when one protein is selected and all of the peptides for this protein are summarized in each of these views (except library match window where nothing is shown).

From the peak area replicate comparison, does this protein appear to be differentially regulated?

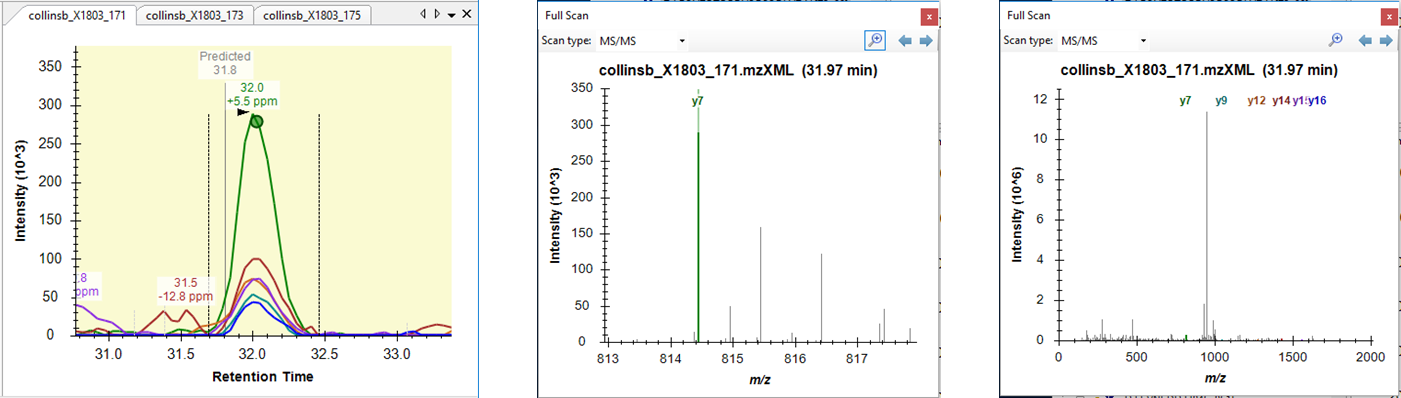
* If you select this fifth peptide (**LPQVEGTGGDVQPSQDLVR**) in this protein, then you get specific information for this peptide in all of these views.



* Examine the peak area patterns for the rest of the peptides belonging to this protein. Is the quantitative pattern for the peptides from this protein consistent with the expected differential regulation pattern?

**NOTE:** If there is more than one precursor charge state for a given peptide sequence these are extracted and scored separately. You can look at these by clicking the **+** next to the peptide sequence and clicking on the individual charge states.

* Click on some of the human proteins in the document. Examine the replicate peak areas from the protein level view and the peptide level view. Are the peak areas consistent with the expected ratios? What about the yeast proteins?
* Click again on the **LPQVEGTGGDVQPSQDLVR** peptide in the first protein
* Notice than if you hover the cursor over the chromatography a circle appears. If you click the circle, Skyline will open the full scan spectrum zoomed on the **y8** ion that you selected. You can click the magnifying glass to zoom out and look at the full DIA scan.

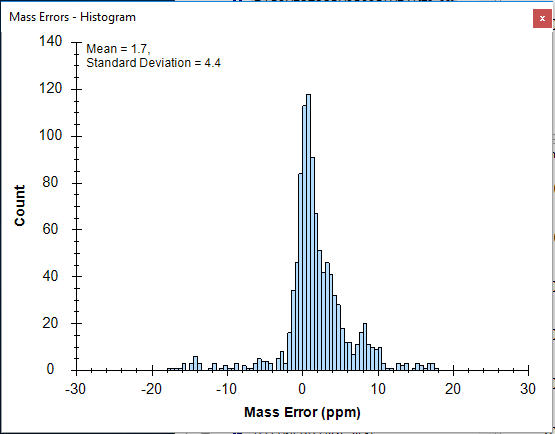


* Explore the data further manually (including some decoys).

# Mass and retention time deviation

We can examine the mass accuracy and retention time prediction accuracy to determine if the optimal extraction parameters have been used.

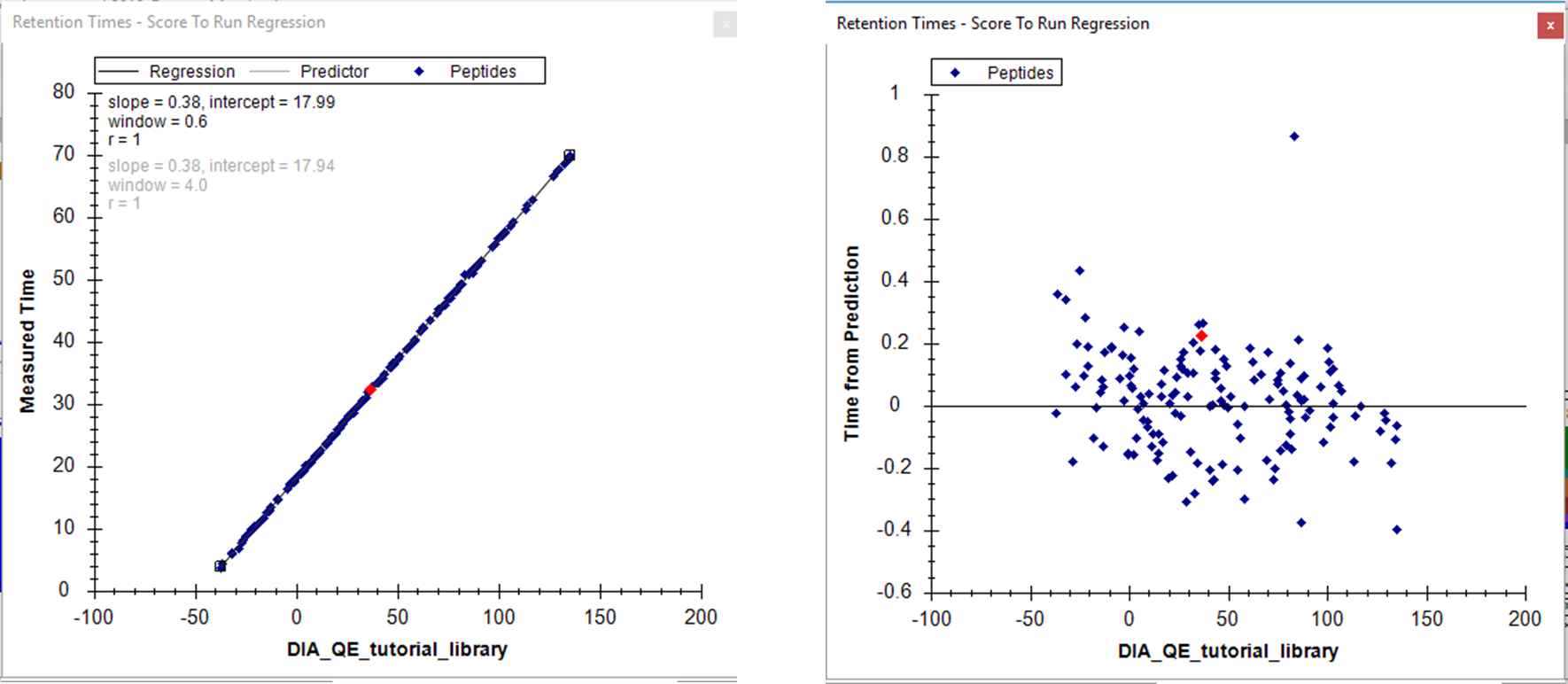
* From the **View** menu, choose **Mass errors** and then **Histogram**. to the see the distribution of mass errors over the data set. Could the extraction window (±20 ppm) have been further optimized?



To the see the linear regression used to predict the target peptide retention times based on the iRT peptides and library iRT values from the target peptides:

* From the **View** menu, choose **Retention Times** and then **Regression.** Lastly choose, **Score to Run**.
* Right click on the graph and select **Plot** and then **Residuals** to see the deviations from the predicted retention times in this data set. Could the extraction window (± 5 minutes) have been further optimized for this analysis?

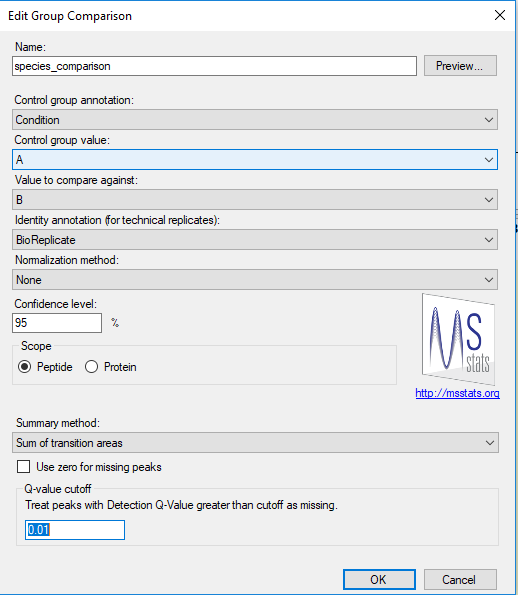
**Note:** As the spectral library for this analysis was constructed from a side-by-side analysis of the same samples. As such, the accuracy of the retention time predictions are very good. Retention times from external spectral libraries acquired on different instruments, at different times, from different samples would lead to larger errors in these predictions.



* Save the file as *DIA\_QE\_inspected.sky*

# Quantitative comparison

* From the **View** menu, select **Group Comparison** and then **Add**.
* Enter the parameters of the group comparison as in the screenshot below (Note: you need to click **Advanced** to see lower options):



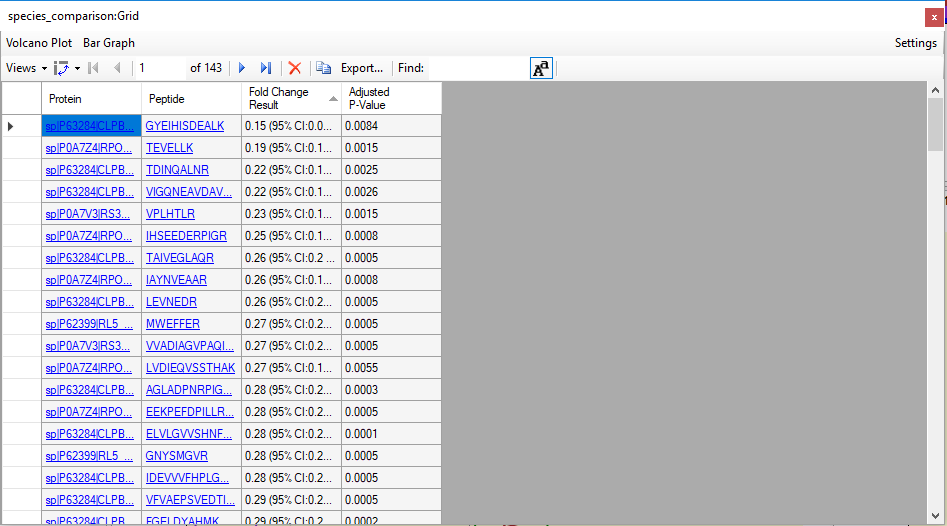
* Click **OK**

From the base screen, click the **View** menu and then select initially **Group Comparison**, followed by **species\_comparison** from the submenu.

A table should appear that shows the peptide level fold-change and adjusted p-value for the comparison between the A and B groups.

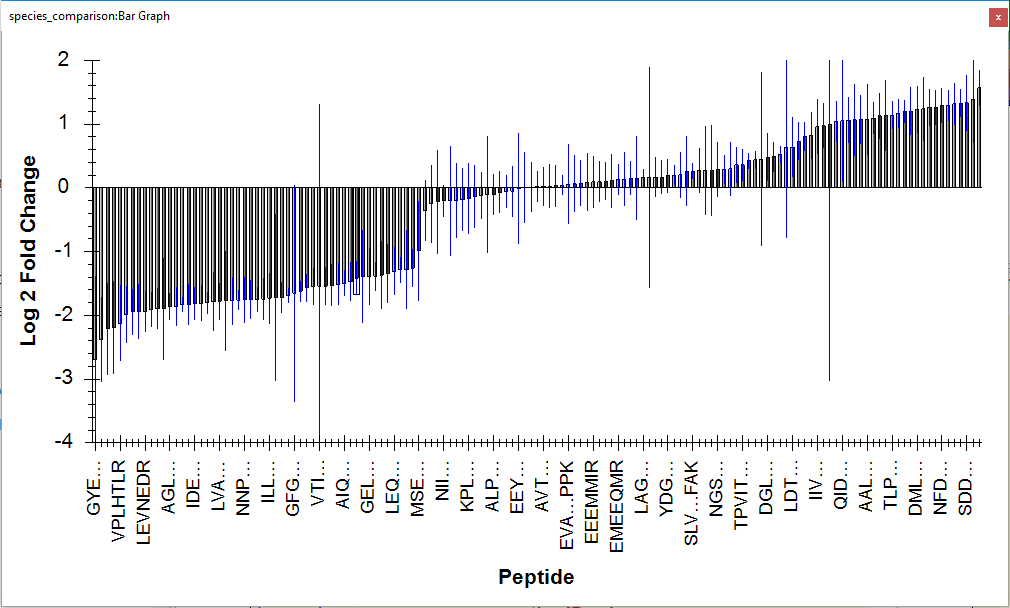
* Expand the width of the **Protein** column header so that you can see the full protein names – including the corresponding species name.
* Right-click on the header of the **Fold Change Result** column and click on **Sort Ascending**.

Inspect the fold changes produced for some of the peptides in the table keeping in mind which species they are from and the expected ratios. What about the adjusted p-values?

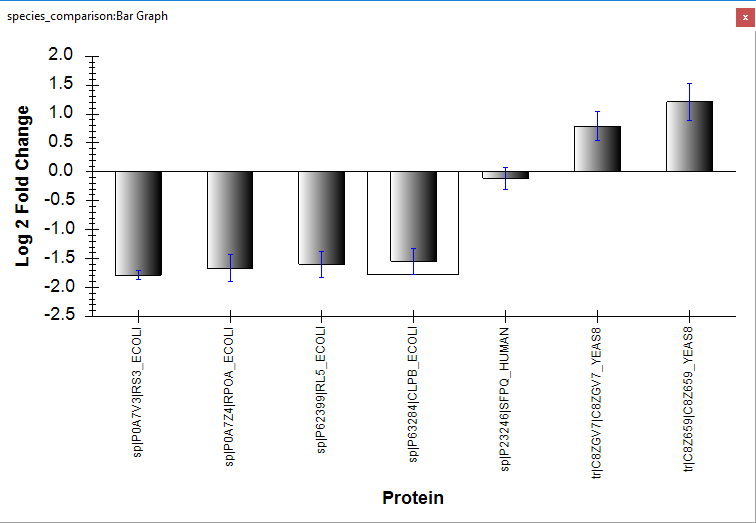


* Click on **Bar Graph** in the top left corner of the table.

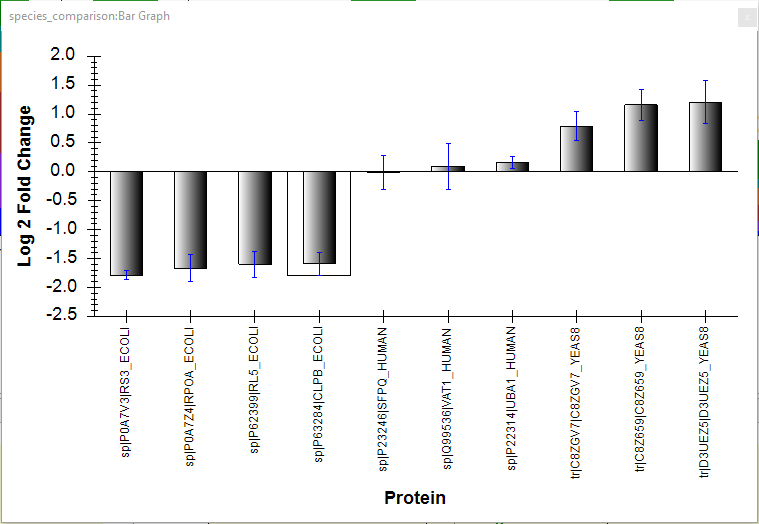
A bar graph of the fold changes for these peptides appears. You can highlight certain peptides in the graph by clicking on various peptides in the **Targets** window of the Skyline document (so that you can confirm which species various peptides in the graph are from).



* Right-click on the graph and select **Settings** to re-open the **Edit Group Comparison** window.
* Change the **Scope** option from **Peptide** to **Protein**.
* Click on **Bar Graph** to see the protein level fold change.
* Now you can see from which species the proteins are from in the x-axis labels. Do the measured ratios fit the expected ratios well? Are there some proteins missing?

ome

* On the **Edit Group Comparison** enable **Use zero for missing peaks** and look
* Close the **Edit Group Comparison** window.



* Save the document as “DIA\_QE\_group\_comparison.sky”
* Export the data for later use in MSstats by clicking on the **File** menu and select **Export** and then choose **Report**.
* From the **External Tools** folder, select **MSstats Input**.

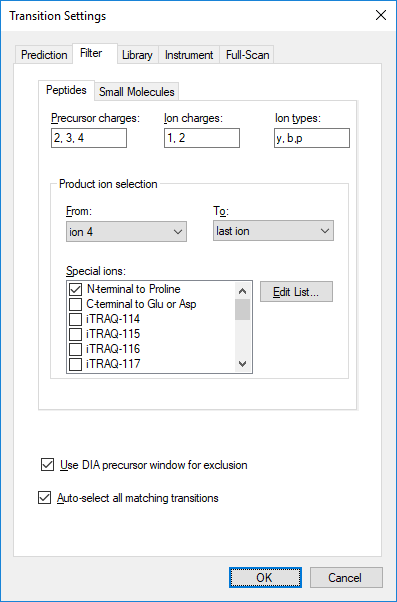
(Note: MSstats must already be installed as an external tool to have access to this report).

* Save as “DIA\_QE\_LFQB\_MSstats\_Input.csv”

# Comparing MS1 and MS2 data

Until now we have only been looking at extracted ion chromatograms from MS2 data (i.e. fragment ions). Now we will compare the signals from MS1 XICs and MS2 XICs.

* From the **Settings** menu, click on **Transition Settings** and then choose **Filter**.
* In the **Ion types** field add “p” to the ion types (indicating precursor).



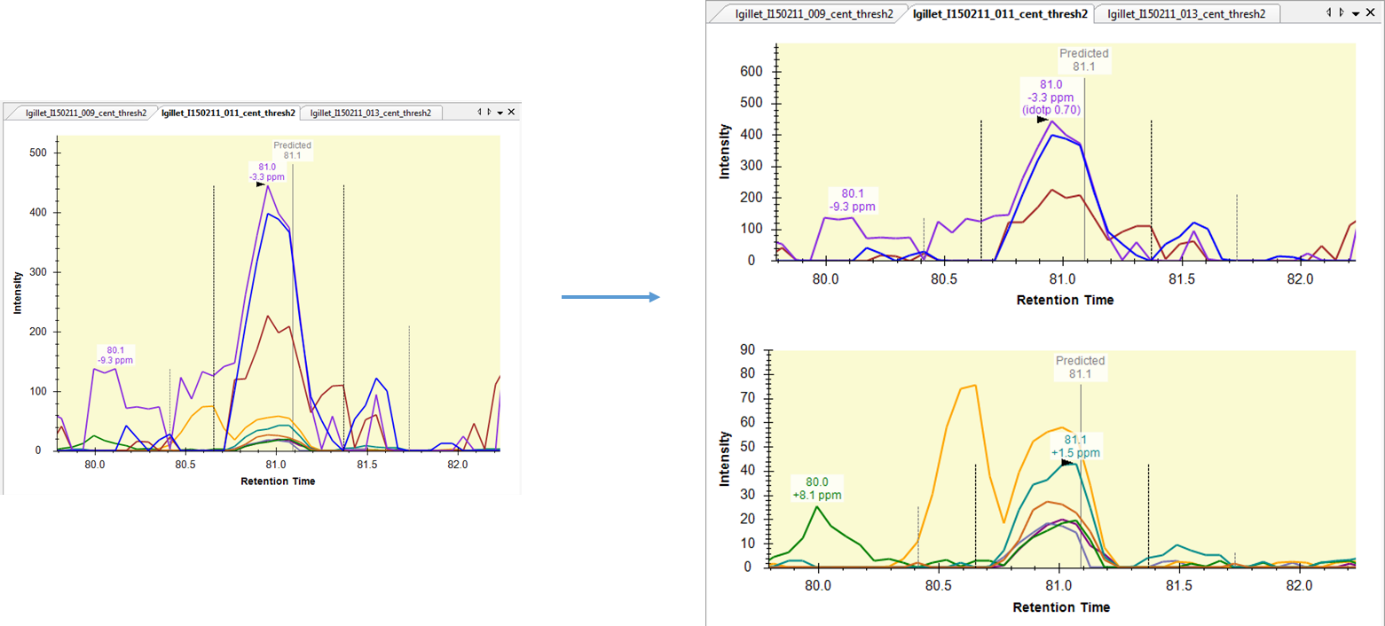
* Click **OK**

The MS1 XICs from the precursor ions will now be overlaid with the DIA MS2 level fragment ion XICs from the fragment ions.

In general, the precursor ions have a much higher absolute intensity than the fragment ions. As such, it is usually easier to view these in separate graphs:

* Click the **View** menu and select **Transitions** and then **Split Graph**.

Now the precursor XICs from the MS1 scans are displayed in the upper graph and the fragment ion XICs from the DIA scans are displayed in the lower graph.



* Search (using *ctrl-F)* for the peptide SDTAAAAVR and examine the difference between the MS1 and MS2 data.
* Browse through some peptides and compare the signals coming from the MS1 XICs and the DIA MS2 XICs. Is there a difference in the selectivity between the MS1 and DIA MS2 data?
* Can you find cases where the DIA MS2 data is higher quality (i.e. better selectivity) than the MS1 data similar to peptide above? Can you find any cases where it is the opposite?
* Save the document as *DIA\_QE\_MS1\_comparison.sky*

Congratulations! You have completed the tutorial ☺