Step-by-Step Tutorial for Integrating Contaminant FASTA and Spectral Libraries in Various DDA and DIA Proteomics Software Platforms

Mass spectrometry-based proteomics is challenged by the presence of contaminant protein background signals. During data analysis, contaminant FASTA libraries allow the search algorithm to distinguish between peptides with similar retention times and *m/z*. Here, we generated universal contaminant FASTA and spectral libraries that can be used for both data-dependent acquisition (DDA) and data-independent acquisition (DIA) proteomics, available to download at: https://github.com/HaoGroup-ProtContLib, and ProteomeXchange (#PXD031139) These new contaminant libraries have been shown to reduce false identifications, increase protein IDs, and do not influence protein quantification for both DIA and DDA workflows. We modified the contaminant FASTA library to contain a "Cont" prefix before each UniProt identifier, simplifying the process of searching and removing contaminant proteins prior to statistical analysis.

In this tutorial, we describe how to use our new contaminant FASTA and spectral libraries with various DDA and DIA software platforms.

Please cite our publication:

Ashley M. Frankenfield, Jiawei Ni, Mustafa Ahmed, Ling Hao, "Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics", Journal of Proteome Research, 2022. https://doi.org/10.1021/acs.jproteome.2c00145

(PDF is available on BioRxiv and the Hao Lab Website: https://blogs.gwu.edu/haolab/research/publication/)

Table of Content:

1. **Brief Description of Contaminant Libraries**
2. **Removing Protein Contaminants from Result File in Excel**
3. **Using Contaminant FASTA in DDA Software**
   - 3.1 Proteome Discoverer for DDA
   - 3.2 MaxQuant for DDA
4. **Building Contaminant Protein Spectral Libraries**
   - 4.1 Building a Spectral Library in Spectronaut
   - 4.2 Building a Spectral Library in MaxQuant
5. **Using Contaminant FASTA and Spectral Libraries in Library-based DIA**
   - 5.1 Spectronaut for library-based DIA
   - 5.2 DIA-NN for library-based DIA
   - 5.3 Skyline for library-based DIA
   - 5.4 MaxDIA for library-based DIA
6. **Using Contaminant FASTA in Library-free DIA**
   - 6.1 Spectronaut for library-free DIA
   - 6.2 DIA-NN for library-free DIA
   - 6.3 PECAN for library-free DIA

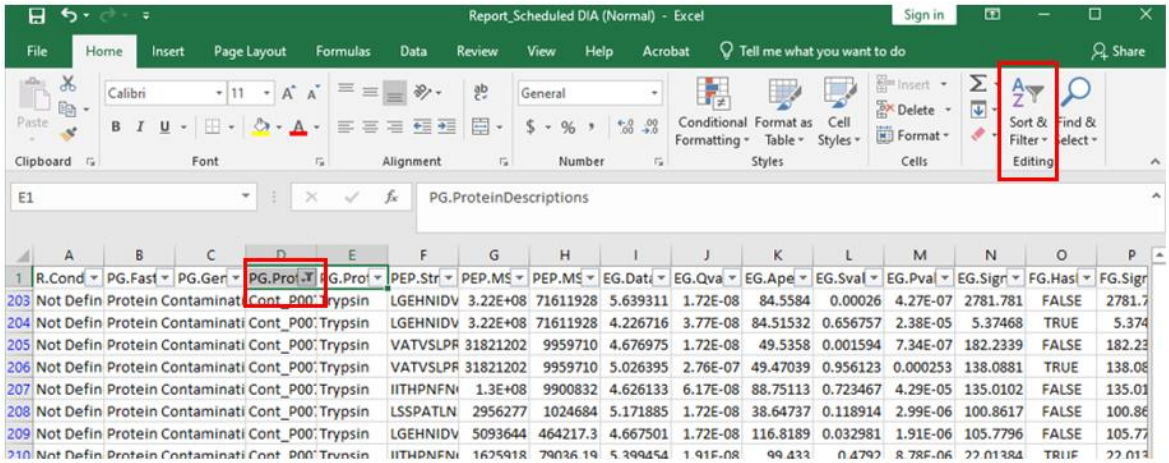
1. Brief Description of Contaminant Libraries

Exogenous contaminant proteins originated from reagents and sample handling are often shared in most bottom-up proteomic experiments. Although widely used for DDA proteomics, the list of common protein contaminants from MaxQuant and cRAP list have not been updated for years, containing deleted Uniprot IDs, sample-specific interference proteins that are incorrectly listed as contaminants, and commercially available human protein standards which are not contaminant proteins. Therefore, we first built a new contaminant FASTA library by manually merging the available contaminant lists online, updating their Uniprot entry IDs, deleting noncontaminant proteins, searching for new contaminant proteins on Uniprot, and combining them into a new FASTA file. Our new contaminant FASTA library contains 381 contaminant proteins including all human keratins and skin-derived proteins, common bovine contaminants from cell culture and affinity columns, various proteolytic enzymes, affinity tags, and other contaminants. When compared to the MaxQuant and cRAP contaminant lists, our new FASTA library is up-to-date for all Uniprot IDs and contains an additional 166 contaminant proteins. This new FASTA library can be used for both DDA and DIA proteomics. We also added a "Cont_" prefix in each contaminant entry in the FASTA library, allowing contaminant proteins to be easily filtered and removed in the result files.

2. Removing Contaminant Proteins from Result Files.

1.1. Launch the results file in Microsoft Excel. In the "Home" tab, click on "Sort & Filter" and then "Filter".

1.2. Navigate to the Protein ID column and type in "Cont_".

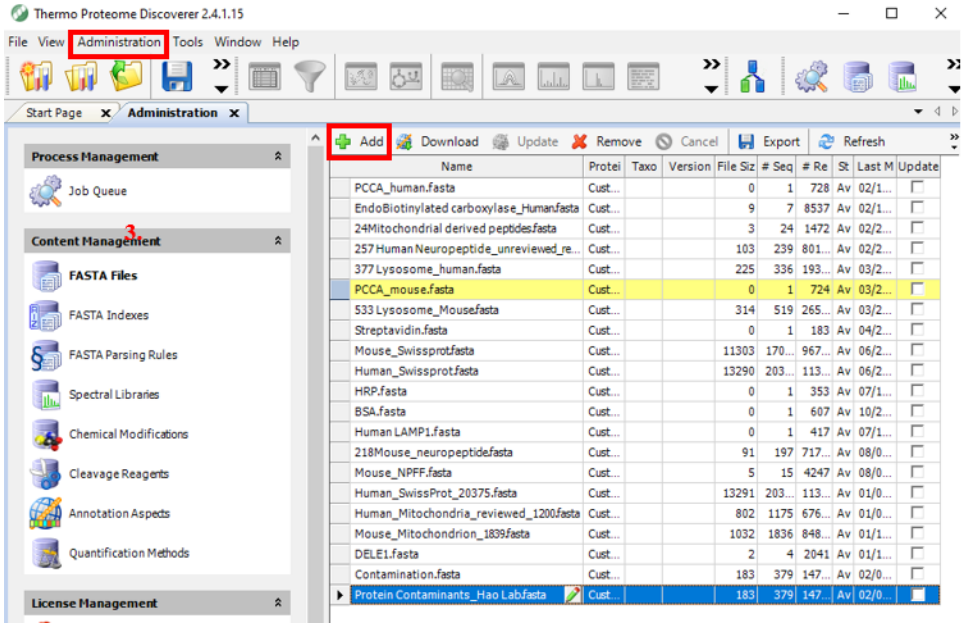


1.3. This will select all contaminant proteins. Evaluate the selected proteins to ensure that they are not biologically relevant based on custom sample types. Remove contaminant proteins prior to statistical analysis.

3. Using Contaminant FASTA in DDA Software

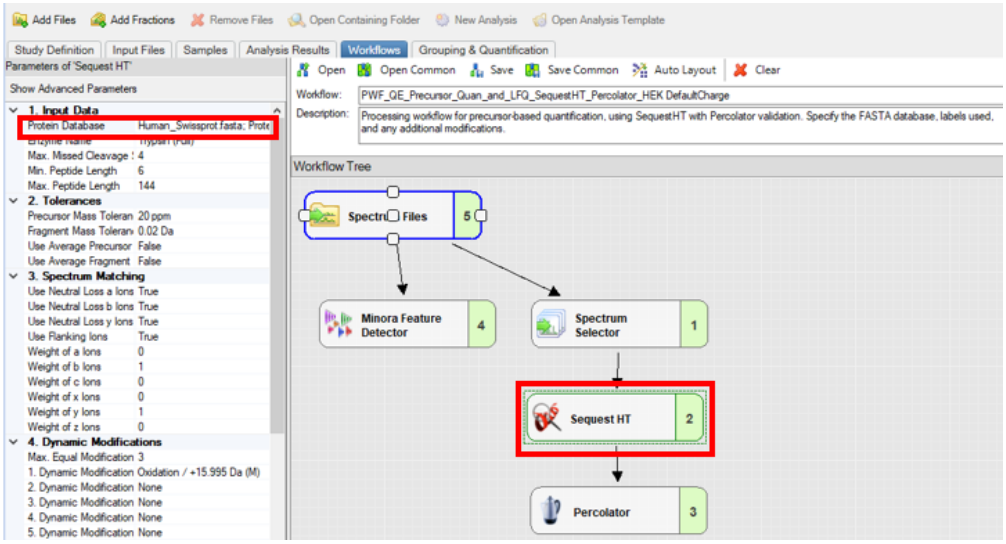
3.1. Proteome Discoverer for DDA

3.1.1. Click the “Administration” tab and select “Maintain Fasta Files”. Click “Add” and then select “Protein Contaminants_Hao Lab.Fasta”.

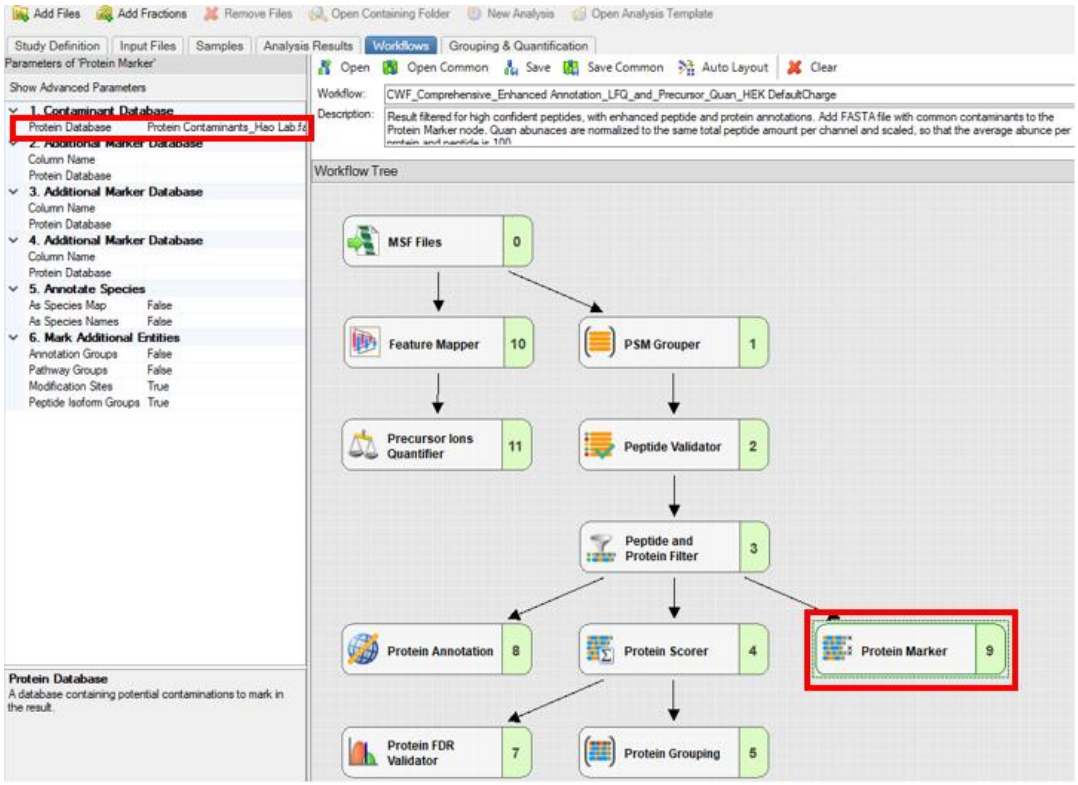


3.1.2. Click on the “Sequest HT” tab in the processing workflow in a study file. For protein database, select both the “Protein Contaminants_Hao Lab” and organism FASTA for your sample.

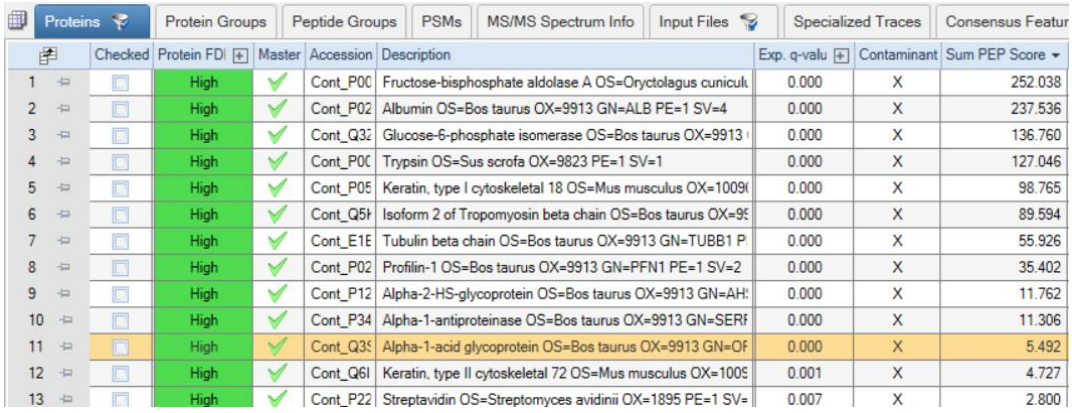
NOTE: The protein contaminant FASTA file must be included in data processing step to reduce protein/peptide false identifications.



3.1.3. Select your consensus step workflow. Under the “Protein Marker” tab, include the contaminant FASTA in the contaminant database. This will create a separate column in the result file marking contaminant proteins.



3.1.4. Contaminant proteins can be filtered using the accession column or separate contaminant column.



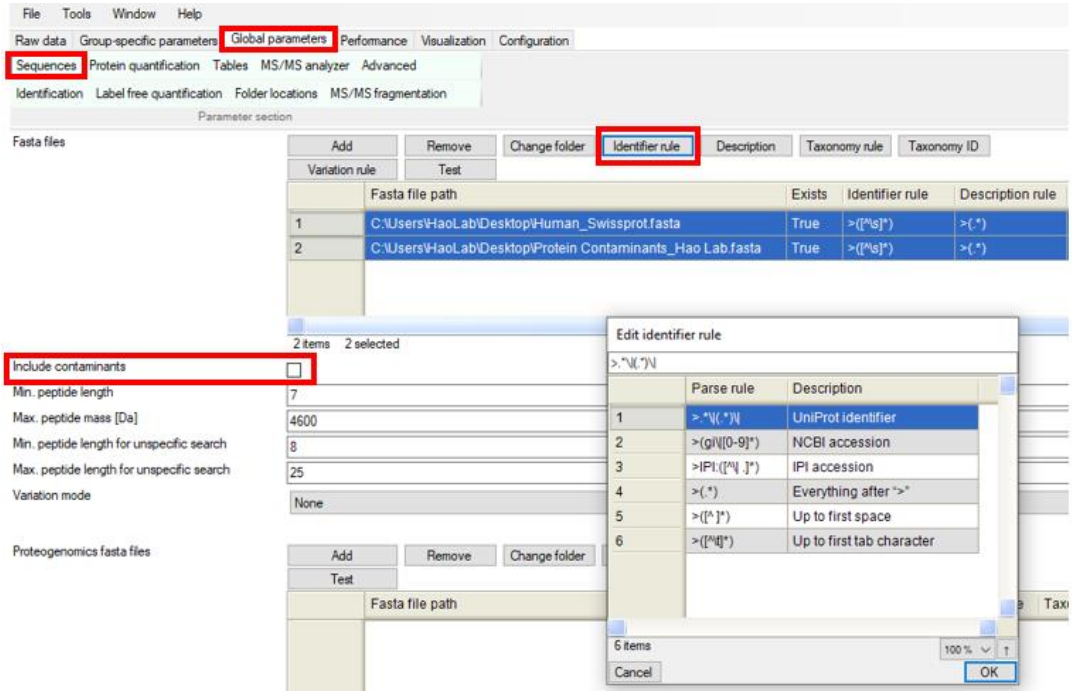
3.2. MaxQuant for DDA

3.2.1 Launch MaxQuant. Load *.raw* files. Click the "Global parameters" tab and then select "Sequences".

3.2.2 Select the "Protein Contaminants_Hao Lab.Fasta" and then click on "Identifier rule". Select "UniProt Identifier".

3.2.3 Unselect "Include contaminants".

NOTE: Including the existing MaxQuant contaminant database will not affect results. However, contaminant proteins from the new FASTA will not be marked in the contaminant column in the MaxQuant results file, which may lead to confusion. Contaminant proteins are marked in the UniProt ID column with the prefix "Cont_" as described on page 2.



4. Building Contaminant Protein Spectral Libraries

To establish comprehensive contaminant protein spectral libraries for DIA proteomics, we created a series of contaminant-only samples using various proteolytic enzymes, affinity purification beads and fetal bovine serum (FBS) that are commonly used for cell culture medium. Contaminant Protein Spectral Library is available to download in Github and ProteomeXchange (#PXD031139). For proteomics software that allows the input of multiple spectral libraries, our contaminant spectral library and custom proteomics data can be included together. For software that only allows one spectral library input, an integrated spectral library can be built using our contaminant-only raw data and custom proteomics data. We have tested that the integrated spectral library performs similarly to two separate libraries. Either method is better compared to the results analyzed without the contaminant library.

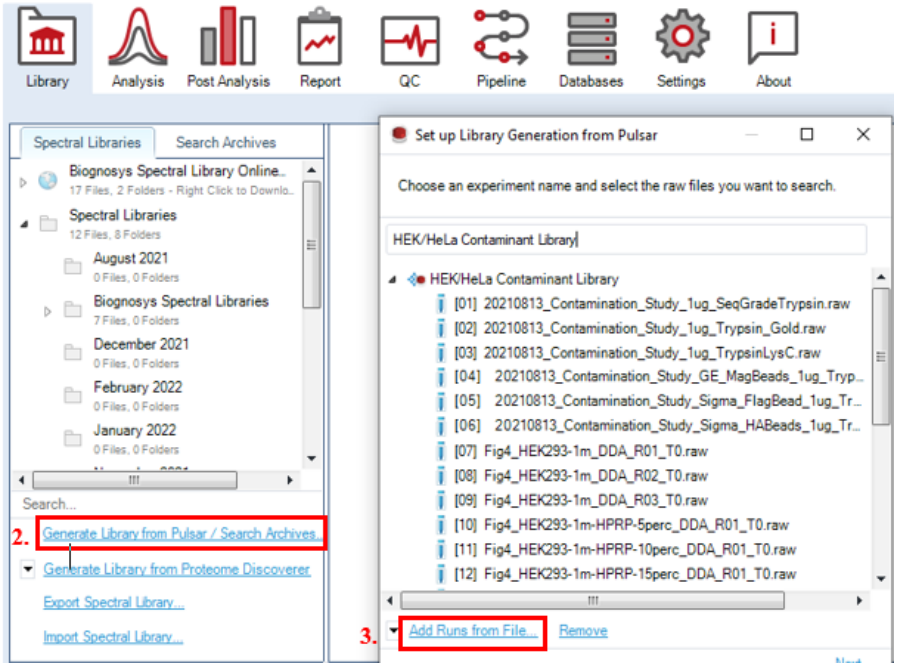
4.1. Building a Spectral Library in Spectronaut

4.1.1. Launch Biognosys Spectronaut and select the “Databases” tab. Import the “Protein Contaminants_Hao Lab.Fasta”.

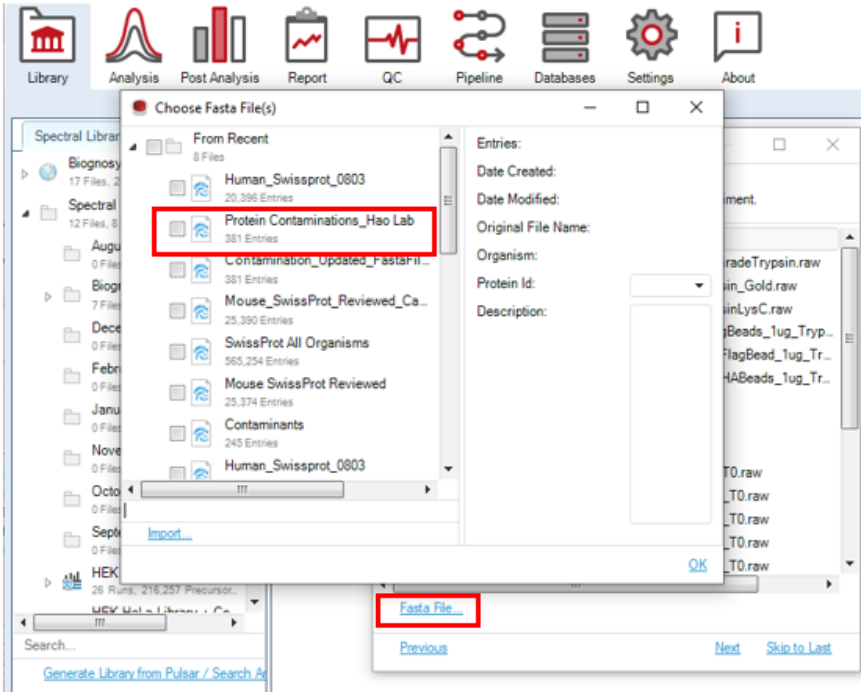
4.1.2. Select the “Library” tab. Click “Generate Library from Pulsar/Search Archives”.

4.1.3. Select “Add Runs from File” to add *.raw* files.

Note: The *.raw* files from our universal contaminant-only experiment can be included to ensure the accurate detection and inclusion of contaminant spectra within the library.



4.1.4. Click “Next” and then “Fasta File.” Select the “Protein Contaminants_Hao Lab.Fasta”. Select the remaining settings to build the desired library.



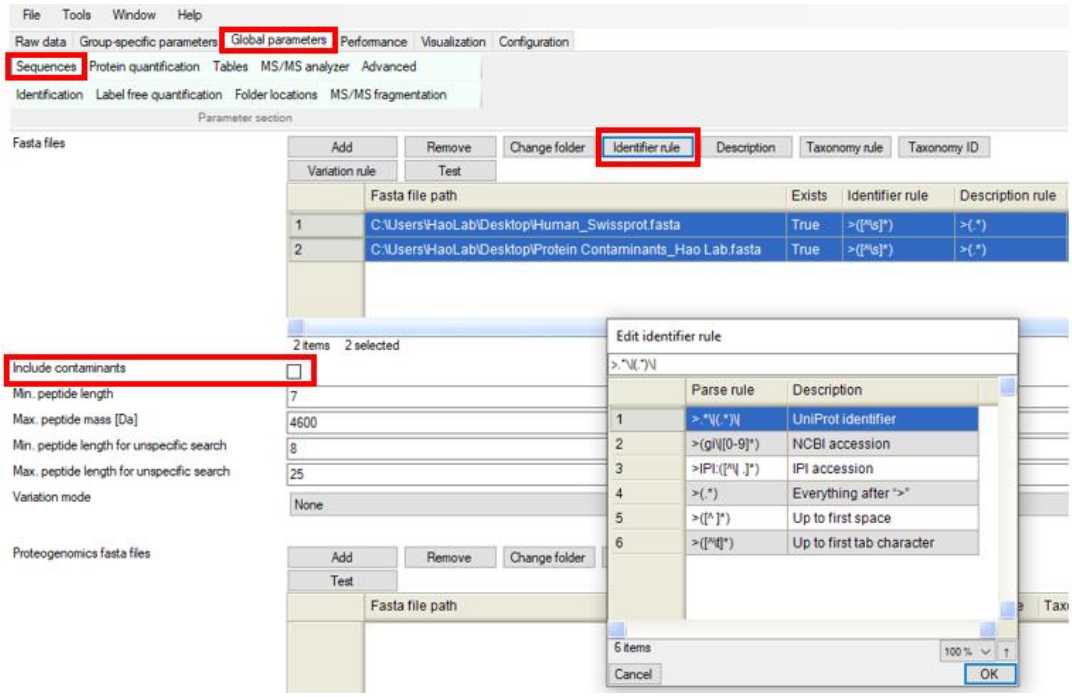
4.2. Building a Spectral Library in MaxQuant

4.2.1 Launch MaxQuant. Load *.raw* files. Click the "Global parameters" tab and then select "Sequences".

4.2.2 Select the "Protein Contaminants_Hao Lab.Fasta" and then click on "Identifier rule". Select "UniProt Identifier".

4.2.3 Unselect "Include contaminants".

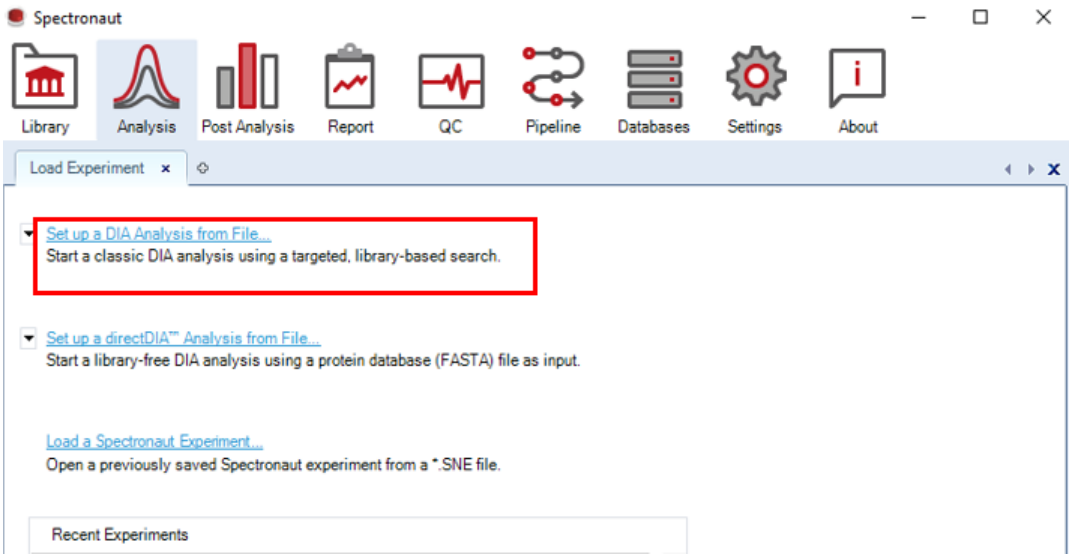
NOTE: Including the existing MaxQuant contaminant database will not affect results. However, contaminant proteins from the new FASTA will not be marked in the contaminant column in the MaxQuant results file, which may lead to confusion. Contaminant proteins are marked in the UniProt ID column with the prefix "Cont_" as described on page 2.



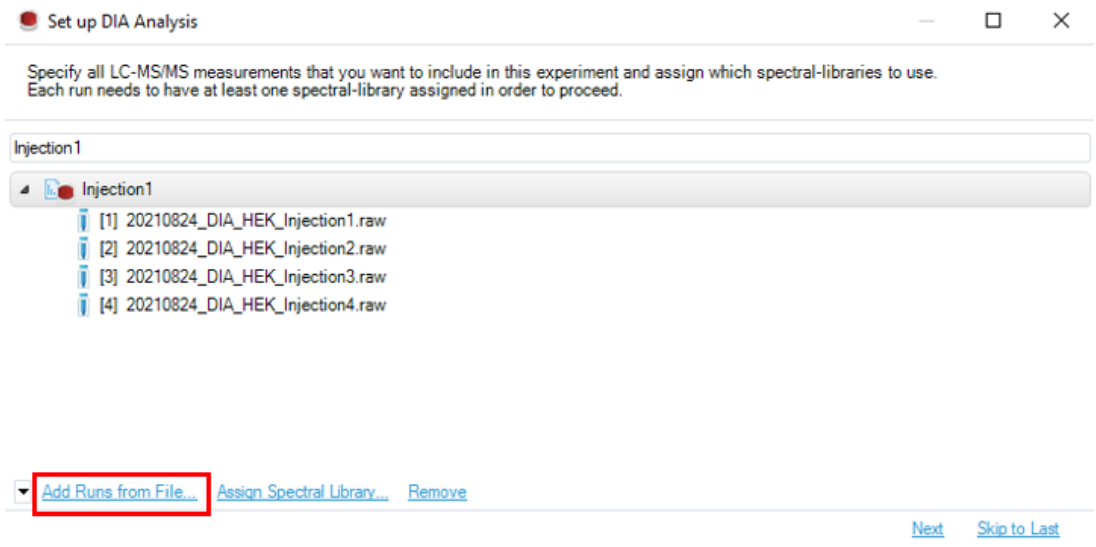
5. Using Contaminant FASTA and Spectral Libraries in Library-based DIA

5.1. Spectronaut for Library-based DIA.

5.1.1. Launch Biognosys Spectronaut. Select “Set up DIA Analysis from File”.



5.1.2. Load the *.raw* files for the study.

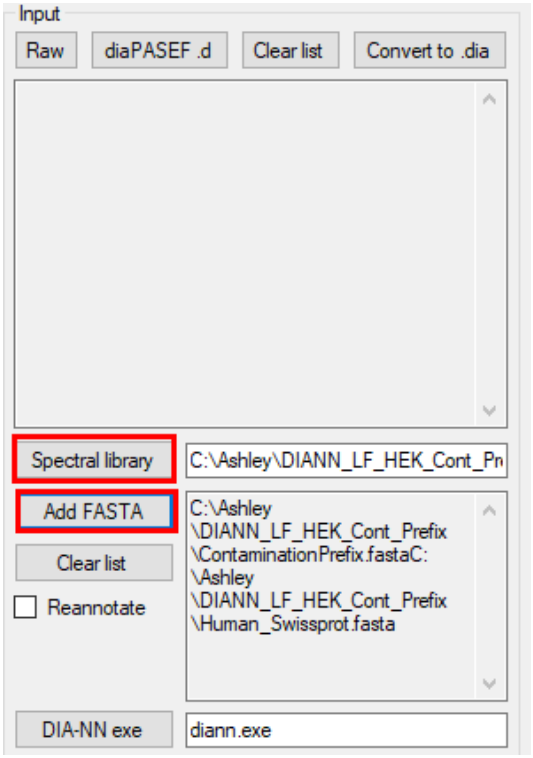


5.1.3. Select the contaminant-containing spectral library and contaminant FASTA used during library creation.

5.2. DIA-NN for Library-based DIA.

5.2.1. Launch DIA-NN. Click “Spectral library” and add the contaminant FASTA that was built using Spectronaut.

5.2.2. Load the *.raw* files. Under “Add FASTA” select the appropriate FASTA libraries to build the spectral library.

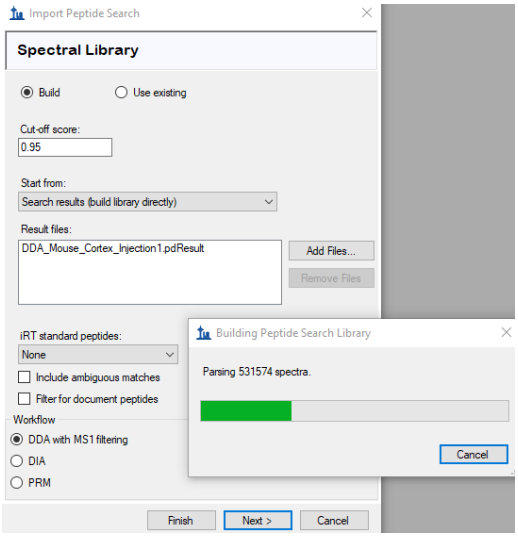


5.3. Skyline for Library-based DIA.

5.3.1. Launch Skyline (version 21.2) and open a “Blank Document”.

5.3.2. A spectral library can be built by selecting “File”, “Import” and then “Peptide Search.”

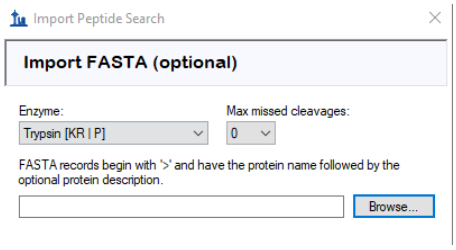
5.3.3. Import the *.pdResult* file from Proteome Discoverer or *msms.text* file from MaxQuant. Select “Next” to build the peptide search library.



5.3.4. Select the appropriate .raw files and click “Next”.

5.3.5. Select the FASTA File and then “Finish”.

NOTE: Only a single FASTA library can be imported. The contaminant FASTA file will need to be combined with the organism FASTA.



5.3.6. Library-based DIA analysis can be conducted using established Skyline workflows. However, the conjoined FASTA file used to build the library should be included during data analysis.

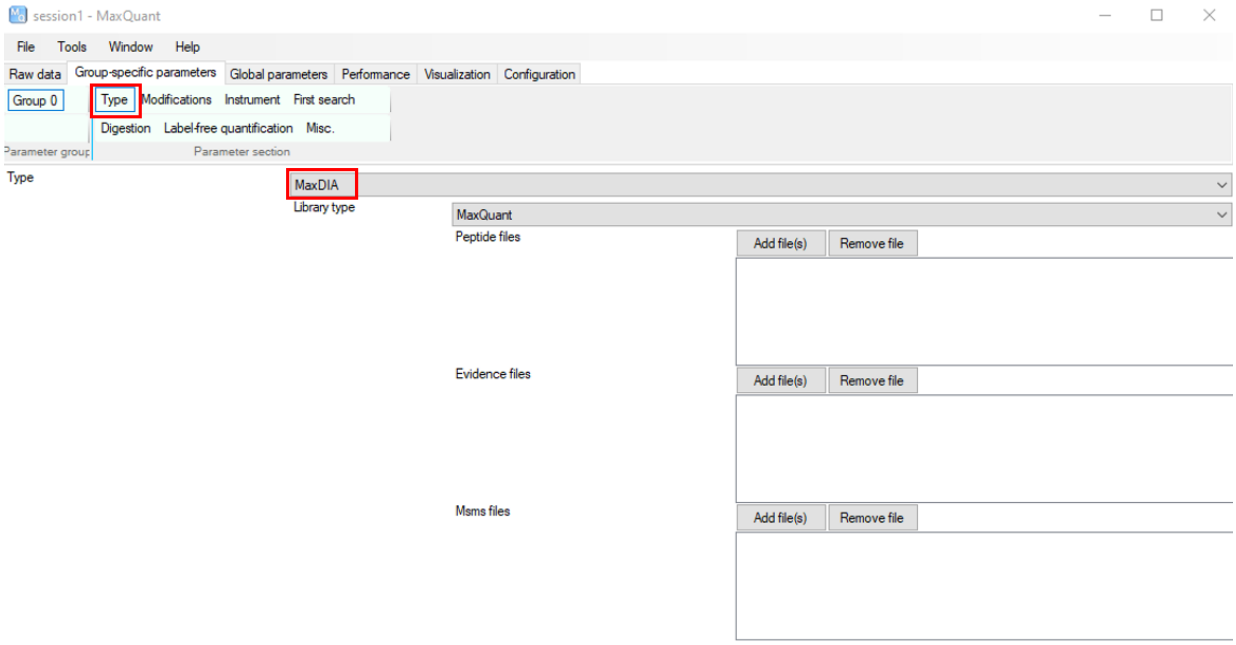
5.4. MaxDIA for Library-based DIA.

5.4.1. Launch MaxQuant. Load .raw files. Click the “Global parameters” tab and then select “Sequences”.

5.4.2. For library-based DIA proteomics, you must include the same contaminant and organism specific FASTA files used to generate the spectral library. Select the “Protein Contaminants_Hao Lab.Fasta” and the organism specific UniProt FASTA file. Click on “Identifier rule” and select “UniProt Identifier”.

5.4.3. Select the “Group-Specific Parameters” tab. Click “Type” and select “MaxDIA” from the drop-down menu.

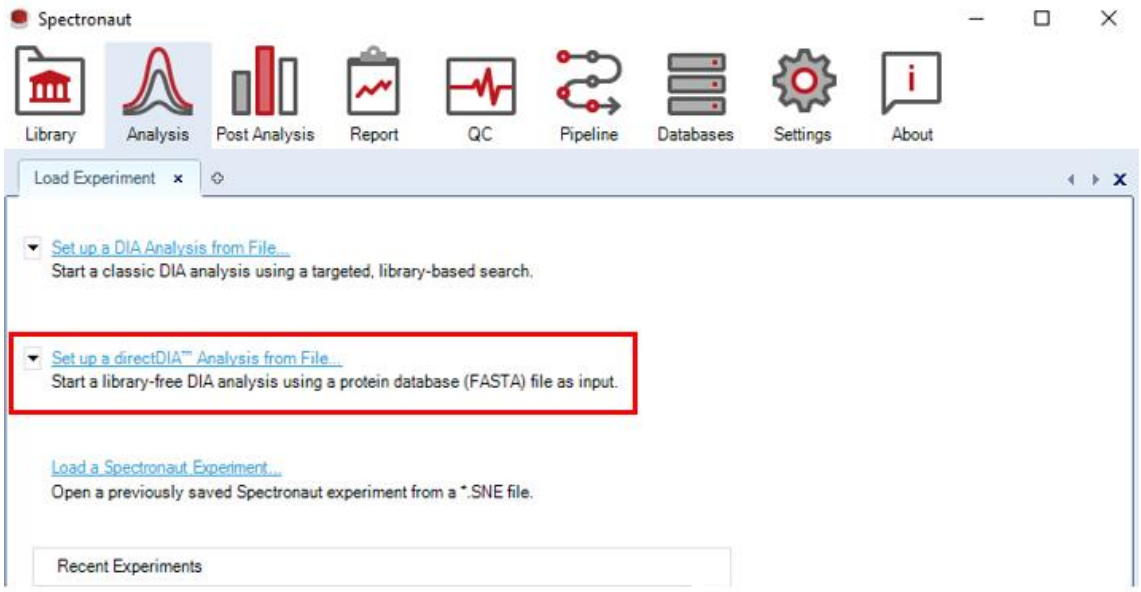
5.4.4. Import the peptide, evidence and msms.*txt* file for library-based DIA.



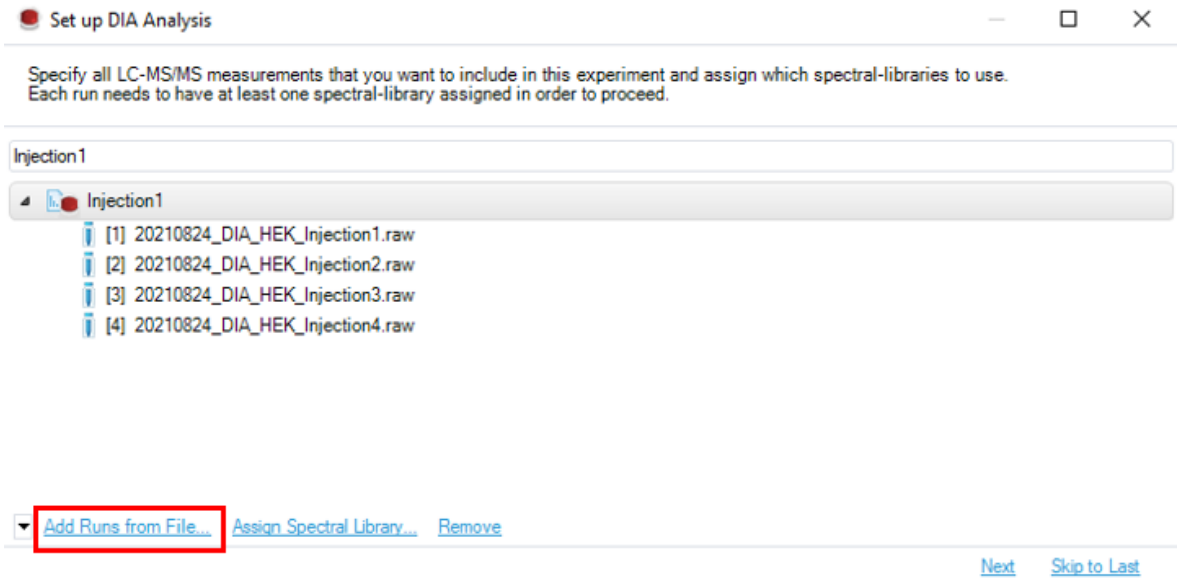
6. Using Contaminant FASTA in Library-free DIA

6.1. Spectronaut for library-free DIA

6.1.1. Launch Spectronaut. Click “Set up a directDIA Analysis from File”.



6.1.2. Load the *.raw* files for the study.

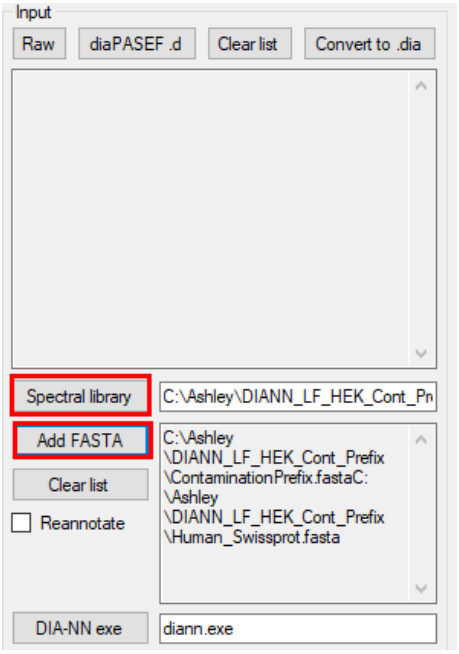


6.1.3. Select the contaminant FASTA and organism FASTA.

6.2. DIA-NN for library-free DIA

6.2.1. Launch DIA-NN. Click “spectral library” and add the contaminant library that was built using Spectronaut.

6.2.2. Under “Add FASTA” select the appropriate FASTA libraries to build the spectral library.



6.3. PECAN for library-free DIA

6.3.1. Launch EncylopeDIA (version 1.12.31). Select the Walnut tab.

6.3.2. Import the contaminant FASTA library to the “Background” and “Target” sections.

NOTE: Only a single FASTA library can be imported into the workflow. The Hao Lab Contaminant library must be combined with your organism FASTA database.