PIONEER: Pipeline for Generating High-Quality Spectral Libraries for DIA-MS Data

Srikanth S. Manda,[1] Zainab Noor,[1] Peter G. Hains,[1] and Qing Zhong[1,2]

[1] ProCan®, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, New South Wales, Australia
[2] Corresponding author: *qzhong@cmri.org.au*

Data-independent-acquisition mass spectrometry (DIA-MS) is a state-of-the-art proteomic technique for high-throughput identification and quantification of peptides and proteins. Interpretation of DIA-MS data relies on the use of a spectral library, which is optimally created from data acquired from the same samples in data-dependent acquisition (DDA) mode. As DIA-MS quantification relies on the spectral libraries, having a high-quality, non-redundant, and comprehensive spectral library is essential. This article describes the major steps for creating a high-quality spectral library using a combination of multiple complementary search engines. We discuss appropriate strategies to control the false discovery rate for the final spectral library as a result of merging multiple searches. © 2021 The Authors Current Protocols © 2021 Wiley Periodicals LLC.

Basic Protocol 1: Searching DDA-MS files with multiple search engines
Basic Protocol 2: Merging results from multiple search engines
Basic Protocol 3: Creating spectral libraries from merged results
Alternate Protocol: Using CLI for automating tasks
Support Protocol: Creating concatenated FASTA files

Keywords: DIA • mass spectrometry • proteomics • spectral library • SWATH

How to cite this article:
Manda, S. S., Noor, Z., Hains, P. G., & Zhong, Q. (2021). PIONEER: Pipeline for generating high-quality spectral libraries for DIA-MS data. *Current Protocols*, *1,* e69. doi: 10.1002/cpz1.cpz69

INTRODUCTION

There has been an exponential increase in the use of mass spectrometry (MS)−based proteomics techniques in the last two decades. Data-dependent acquisition (DDA) and data-independent acquisition (DIA) are the most common MS data acquisition techniques. The DDA mode is generally used in discovery studies with the aim of identifying the maximal number of proteins from a limited number of complex biological samples. By contrast, DIA is more frequently used to quantify proteins by combining the merits of both DDA and targeted acquisition methods such as selective reaction monitoring (SRM; Ludwig et al., 2018; Peterson, Russell, Bailey, Westphall, & Coon, 2012), enabling large-scale and consistent protein quantification. Sequential windowed acquisition of all theoretical fragment ion spectra (SWATH)-MS operates in DIA mode, which can accurately quantify thousands of proteins in a reproducible manner (Collins et al., 2017; Poulos et al., 2020). Peptide identification in DIA-MS data requires a spectral library, which is a

Current Protocols e69, Volume 1
Published in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpz1.69

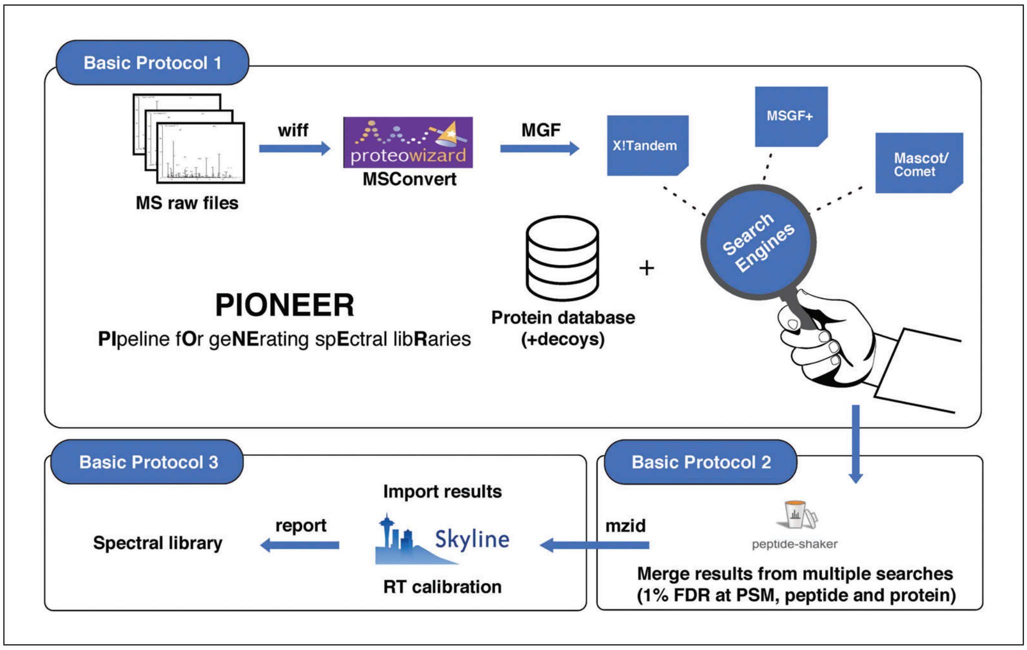


Figure 1 Overview of the pipeline for generating spectral libraries (PIONEER).

curated, searchable, and non-redundant collection of peptide tandem mass spectra. These spectra are usually generated by pooling and fractionating cohort samples running in DDA mode. The acquired spectra are searched against the theoretical spectra that are generated by in silico digestion of a protein database. The spectral library thus serves as a template, providing information about the underlying protein, peptide sequences, mass-to-charge ratios (*m*/*z*) of precursor and fragment ions, precursor and fragment charges, fragment ion types, relative fragment ion intensities, and normalized retention time. By comparing the tandem mass spectra generated in DIA mode with the information in the spectral library, peptides can be reliably identified and accurately quantified (Ludwig et al., 2018).

The protocols in this article provide a step-by-step guide to the generation of a high-quality spectral library using a combination of search engines to increase the protein coverage and to control the false discovery rates (FDR) in order to minimize incorrect identifications (Fig. 1). Basic Protocol 1 describes how to perform searches using three complementary open-source search engines, namely X!Tandem (Craig & Beavis, 2004), Comet (Eng, Jahan, & Hoopmann, 2013), and MSGF+ (Kim & Pevzner, 2014). Basic Protocol 2 illustrates how to merge the results from different search engines using PeptideShaker (Vaudel et al., 2015). Basic Protocol 3 presents the final step of spectral library generation, which uses Skyline (MacLean et al., 2010) to create the final library from the merged results. The Alternate Protocol depicts a command-line version for Basic Protocols 1 and 2, which can be used to automate large-scale jobs consisting of multiple fractionated samples. Also, a Support Protocol demonstrates the creation of a concatenated FASTA database containing decoy sequences, and the merging of multiple spectral libraries with retention time differences using iSwathX (Noor et al., 2019). Basic Protocols 1 and 2 and the Support Protocol can be implemented in either Windows or Linux environments, and Basic Protocol 3 and the Alternate Protocol require Windows 10 or later. The final library is compatible with OpenSWATH (Rost et al., 2014) and other common DIA-MS analysis tools such as Peakview® , Skyline (MacLean et al., 2010), Spectronaut (Bruderer et al., 2015), and DIA-NN (Demichev, Messner, Vernardis, Lilley, & Ralser, 2020) when formatted accordingly. All files described in these protocols can be downloaded from the link provided in Internet Resources.

STRATEGIC PLANNING

Peptide identification is the most time-consuming step in Basic Protocol 1 and Alternate Protocol. There is a range of search engines available for this task, and each one has its advantages and disadvantages. Many studies have reported increased identifications with the use of multiple search engines (Cho et al., 2015; Matthiesen, Prieto, & Beck, 2020; Paulo, 2013; Shteynberg, Nesvizhskii, Moritz, & Deutsch, 2013). While Basic Protocol 1 describes the use of three search engines, namely X!Tandem, Comet, and MSGF+, Basic Protocol 2 shows the merged results of a different combination of three search engines consisting of X!Tandem, Mascot (Perkins, Pappin, Creasy, & Cottrell, 1999), and MSGF+. These two different sets of search engines were used to illustrate the versatility of the protocols. The four search engines used in the two sets were chosen based on complementarity, compute resource requirements, and run time, weighted by the requirement for a commercial license for Mascot. Researchers without a commercial license for Mascot can utilize the other three search engines, which are open source. If computer resources are limited, researchers are encouraged to use either X!Tandem or Comet only for faster computation, whereas stand-alone MSGF+ can be used for more thorough searches. It is advised to use an odd number of search engines, which allows consensus identifications by majority voting. The protein databases should be in FASTA format, and the decoy sequences (preferably reverse sequences) should have a suffix of `_Reversed` appended to the FASTA header. Also, retention time (RT) peptides (Searle et al., 2018) should be added to the same database before initializing any search.

SEARCHING DDA-MS FILES WITH MULTIPLE SEARCH ENGINES

BASIC PROTOCOL 1

Raw data are first converted to the Mascot generic format (MGF), which can be converted from proprietary instrument files of various MS vendors such as SCIEX, ThermoFisher, Bruker, and Agilent. The resulting MGF file will be searched against the respective protein database of interest. Sample data (Supp.Data) are provided from HEK293 cell line fractions acquired in DDA mode on a SCIEX TripleTOF 6600 instrument with a 90-min high performance liquid chromatography (LC) gradient. The data files are in SCIEX *wiff* format. In this protocol, we use the SearchGUI (Barsnes & Vaudel, 2018) tool, which provides an easy-to-use graphical user interface (GUI) for searching using multiple search engines. It supports the following search engines: X!Tandem, MyriMatch (Tabb, Fernando, & Chambers, 2007), MS Amanda (Dorfer et al., 2014), MS-GF+ (Kim & Pevzner, 2014), Comet, Tide (Diament & Noble, 2011), and Andromeda (Cox et al., 2011). Here, X!Tandem, MS-GF+ and Comet are used as the three default search engines, and others can be selected if required. The protein database used in this study consists of Uniprot (UniProt, 2019) canonical protein sequences appended with decoys and RT peptides.

Necessary Resources

Hardware

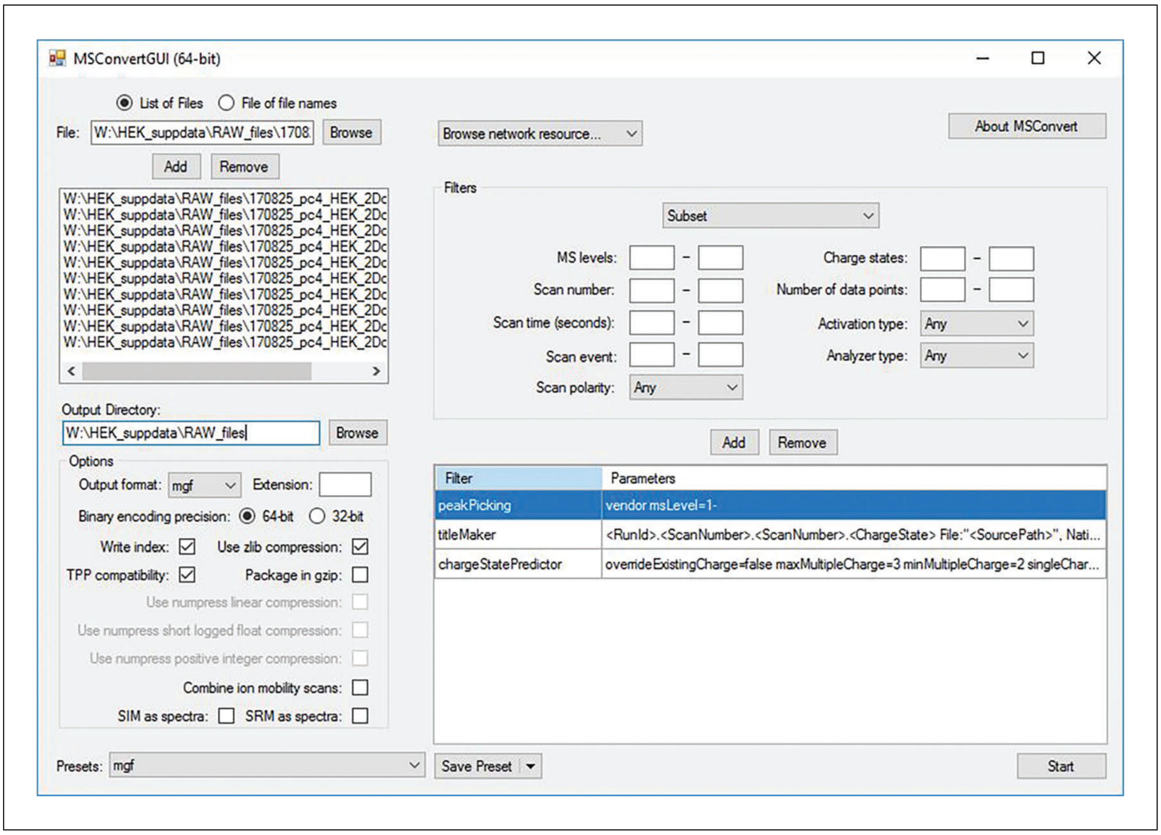
A computer with Windows 10 or later, or Ubuntu, preferably a workstation
A minimum of 16 GB RAM

Software (download the latest versions from the links provided in Internet Resources)

MSConvert (Proteowizard)
SearchGUI
Java version 8.0 or higher

Input files

Spectrum raw files such as *wiff*, *raw*, etc.
Protein database in FASTA format (with decoys appended)
Parameter file (*par*)



● List of Files ○ File of file names

File: W:\HEK_suppdata\RAW_files\1708

W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc
W:\HEK_suppdata\RAW_files\170825_pc4_HEK_2Dc

Output Directory: W:\HEK_suppdata\RAW_files

Options

Output format: mgf Extension:

Binary encoding precision: ● 64-bit ○ 32-bit

Write index: ☑ Use zlib compression: ☑

TPP compatibility: ☑ Package in gzip: □

Use numpress linear compression: □

Use numpress short logged float compression: □

Use numpress positive integer compression: □

Combine ion mobility scans: □

SIM as spectra: □ SRM as spectra: □

Presets: mgf

Filters

Subset

MS levels: - Charge states: -

Scan number: - Number of data points: -

Scan time (seconds): - Activation type: Any

Scan event: - Analyzer type: Any

Scan polarity: Any

Filter	Parameters
peakPicking	vendor msLevel=1-
titleMaker	<RunId>.<ScanNumber>.<ScanNumber>.<ChargeState> File:"<SourcePath>", Nati...
chargeStatePredictor	overrideExistingCharge=false maxMultipleCharge=3 minMultipleCharge=2 singleChar...

Figure 2 MSConvert main interface to add the input wiff files, set the parameters and convert to MGF format.

Converting raw files to MGF format

1. Open the MSConvertGUI and change the default settings to vendor-specific as displayed in Figure 2. Browse and locate the folder with the 10 HEK *wiff* files (`../HEK_suppdata/RAW_files/`). The default output directory will be the same directory. Change if you want a different location.

   The MGF format is a generic format accepted by almost all search engines. Because it is a time-consuming step, users are advised to convert all files beforehand.

2. Click “Start” to obtain the 10 MGF files in the output folder.

3. Click “Save Preset” to save the settings for any future experiment.

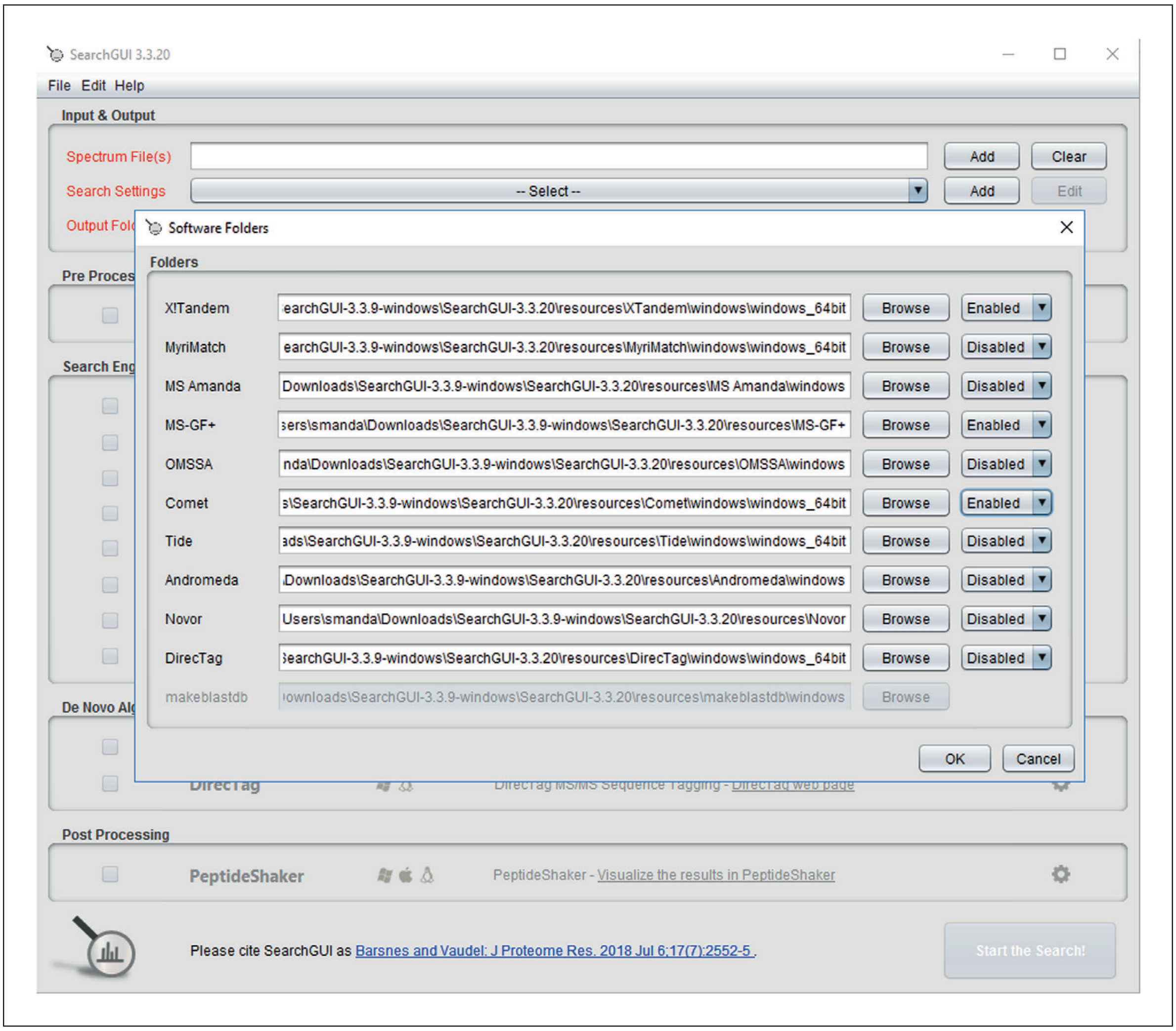
Searching using SearchGUI

4. Configure the desired search engines in SearchGUI. Open SearchGUI and navigate to “Edit” > “Software Locations.” SearchGUI comes with prebuilt executables for all of the aforementioned search engines. To use a single search engine or a combination of search engines, choose “Enabled” in “Software Folders” (Fig. 3) and click “OK.” Here, X!Tandem, Comet, and MSGF+, are selected.

   If the current version of a supplied search engine is not up to date, users can download it manually from the respective source and “Browse” to the local folder in the “Software Folders.” Users are encouraged to try other search engines to find the best combination that suits their requirements.

5. Close “Software Folders” and click “Add” to include the 10 MGF files (step 2) as “Spectrum File(s)” in the main interface of SearchGUI.

6. Click “Add” in “Search Settings,” choose the “Import from File” option below, and select the parameter file (`../HEK_suppdata/PeptideShakerResults/threesearchengine_50ppm.par`). This will populate the desired settings.



X!Tandem	earchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\XTandem\windows\windows_64bit	Browse	Enabled
MyriMatch	earchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\MyriMatch\windows\windows_64bit	Browse	Disabled
MS Amanda	Downloads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\MS Amanda\windows	Browse	Disabled
MS-GF+	sers\smanda\Downloads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\MS-GF+	Browse	Enabled
OMSSA	nda\Downloads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\OMSSA\windows	Browse	Disabled
Comet	s\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\Comet\windows\windows_64bit	Browse	Enabled
Tide	ads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\Tide\windows\windows_64bit	Browse	Disabled
Andromeda	Downloads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\Andromeda\windows	Browse	Disabled
Novor	Users\smanda\Downloads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\Novor	Browse	Disabled
DirecTag	earchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\DirecTag\windows\windows_64bit	Browse	Disabled
makeblastdb	ownloads\SearchGUI-3.3.9-windows\SearchGUI-3.3.20\resources\makeblastdb\windows	Browse	

DirecTag MS/MS Sequence Tagging - DirecTag web page

PeptideShaker - Visualize the results in PeptideShaker

Please cite SearchGUI as Barsnes and Vaudel: J Proteome Res. 2018 Jul 6;17(7):2552-5 .

Figure 3 Enabling search engines in SearchGUI. This module allows selecting different search engines to use for identification along with the search parameters.

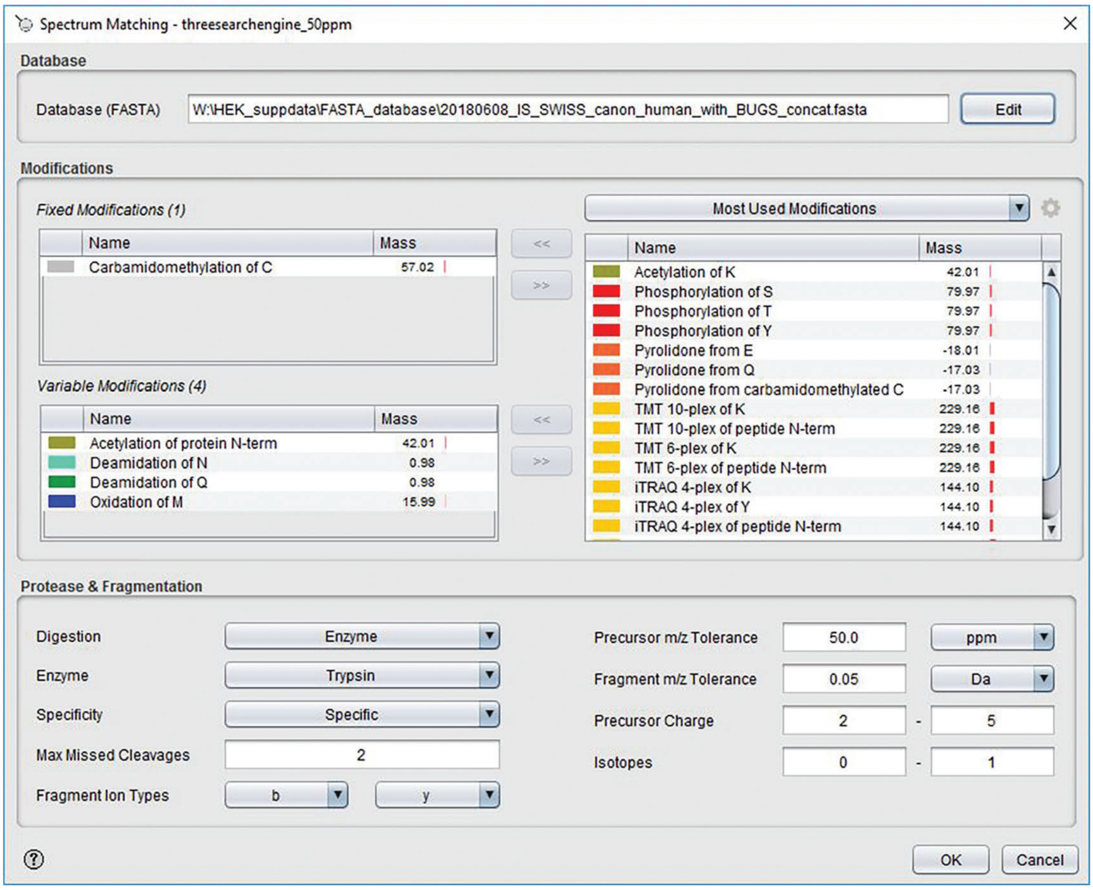
Click on the "Spectrum Matching" tab to verify that the settings and location of the *FASTA* database (`../HEK_suppdata/FASTA_database/..`) is correct (Fig. 4). Click "Ok" to return to the main screen.

The parameter file (`threesearchengine_50ppm.par`) is a preset file with settings pertaining to the dataset. For individual experiments, select the appropriate enzyme, modifications, and tolerance levels. For all three search engines (step 4), the following parameters are used. The precursor tolerance is set to 50 ppm and fragment tolerance to 0.05 Da. Carbamidomethylation at cysteine is used as a "Fixed Modification," while Oxidation at methionine, Deamidation at N and Q, and Acetylation at N-term are used as "Variable Modifications." A total of two missed cleavages are allowed in the search with fully tryptic peptides. The FASTA database used here is a UniProt human protein database appended with decoys and RT peptides. Also, in "Edit" > "Advanced Settings," select "No Zipping" in "Group Identification Files."

7. Choose an output folder for the result files and click on "Start the Search."

   The search should be completed in about an hour or more depending on the system's memory and available cores.

8. The output folder will contain the 10 result files from each search (`../HEK_suppdata/SearchEngineResults/..`).



Spectrum Matching - threesearchengine_50ppm

Database

Database (FASTA): W:\HEK_suppdata\FASTA_database\20180608_IS_SWISS_canon_human_with_BUGS_concat.fasta

Modifications

Fixed Modifications (1)

Name	Mass
Carbamidomethylation of C	57.02

Variable Modifications (4)

Name	Mass
Acetylation of protein N-term	42.01
Deamidation of N	0.98
Deamidation of Q	0.98
Oxidation of M	15.99

Most Used Modifications

Name	Mass
Acetylation of K	42.01
Phosphorylation of S	79.97
Phosphorylation of T	79.97
Phosphorylation of Y	79.97
Pyrolidone from E	-18.01
Pyrolidone from Q	-17.03
Pyrolidone from carbamidomethylated C	-17.03
TMT 10-plex of K	229.16
TMT 10-plex of peptide N-term	229.16
TMT 6-plex of K	229.16
TMT 6-plex of peptide N-term	229.16
iTRAQ 4-plex of K	144.10
iTRAQ 4-plex of Y	144.10
iTRAQ 4-plex of peptide N-term	144.10

Protease & Fragmentation

Digestion: Enzyme

Enzyme: Trypsin

Specificity: Specific

Max Missed Cleavages: 2

Fragment Ion Types: b y

Precursor m/z Tolerance: 50.0 ppm

Fragment m/z Tolerance: 0.05 Da

Precursor Charge: 2 - 5

Isotopes: 0 - 1

Figure 4 SearchGUI and PeptideShaker "Spectrum Matching" module to provide a search database and search parameters including peptide modifications and fragmentation settings.

Mascot outputs a dat format, X!Tandem outputs XML, Comet outputs pepXML, and MSGF+ outputs mzIdentML (mzid). Three search engines yield 30 output files.

BASIC PROTOCOL 2

MERGING RESULTS FROM MULTIPLE SEARCH ENGINES

Here, we describe the merging of results from different searches into a single *mzid* file. This can be achieved by using PeptideShaker software (Vaudel et al., 2015). PeptideShaker reanalyzes the results and converts scores of different search engines to posterior error probability values. These values are used to combine different libraries internally. It also handles the FDR at various levels of interest using the target-decoy approach. In the current approach, we combine the results from a different set of three search engines, namely X!Tandem, Mascot, and MSGF+, after applying 1% FDR at peptide-spectrum match (PSM), peptide and protein levels for the final results. Although the tool is available both for the GUI and command-line interface (CLI), we describe only GUI here. Procedures for CLI can be found in the Alternate Protocol.

Necessary Resources

Hardware

A computer with Windows 10 or later or Ubuntu, preferably a workstation
A minimum of 16 GB RAM

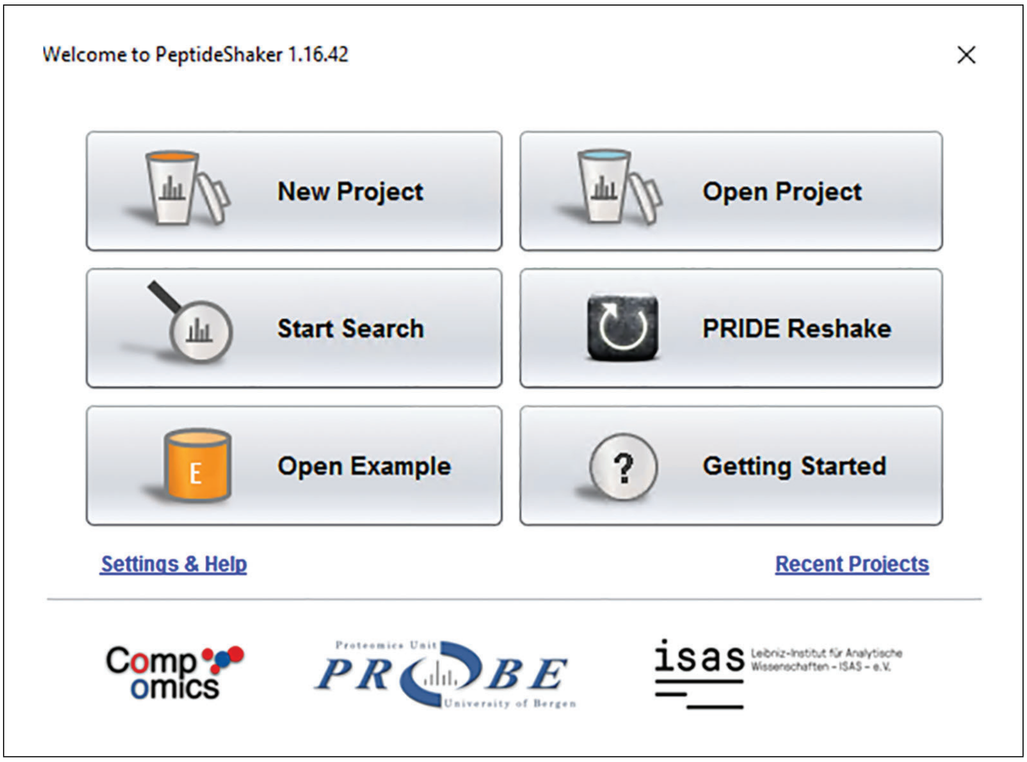


Figure 5 PeptideShaker main interface to start a new project, open a saved project or run an example project.

Software

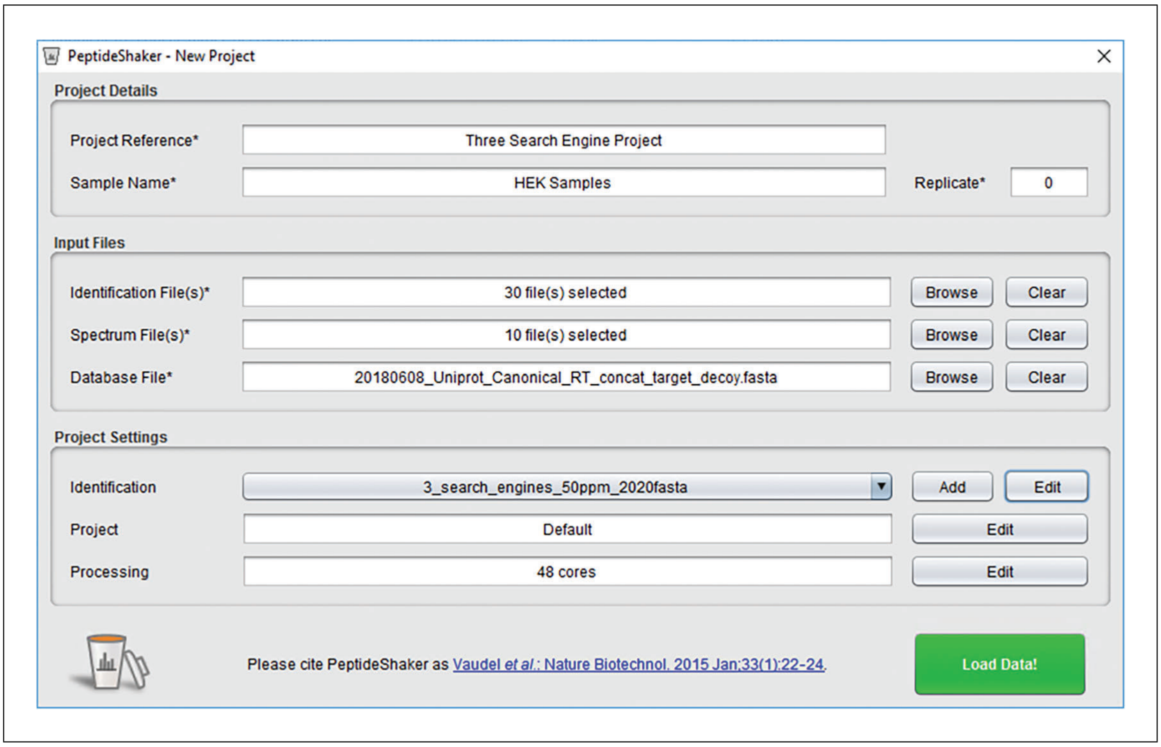
PeptideShaker (download the latest version from the link provided in Internet Resources)

Input files

MGF files and location
Search results and location
Parameter file (*par*)
Concatenated database (FASTA format)

1. Double click the `PeptideShaker.jar` file to start the GUI. Select "New Project" in the main interface of the PeptideShaker (Fig. 5).
2. Fill the required fields in the "PeptideShaker–New Project" module of the software (Fig. 6).
3. Specify a project name in the line marked "Project Reference" (Fig. 6).
4. Specify a sample name in the line marked "Sample Name" (Fig. 6).
5. Under the "Input Files" box, browse and locate the folder with the identification files (`../HEK_suppdata/SearchEngineResults/..`) (Fig. 6) in the space line marked "Identification File(s)." These are the search result files from the different search engines. In this case, we have 30 result files from three search engines.

   The folder contains results from each of the search engines. This step uses three search results, i.e., Mascot, X!Tandem, and MSGF+. Users are encouraged to try different combinations of search engines to observe differences in identifications.



PeptideShaker - New Project

Project Details

Project Reference*: Three Search Engine Project

Sample Name*: HEK Samples

Replicate*: 0

Input Files

Identification File(s)*: 30 file(s) selected

Spectrum File(s)*: 10 file(s) selected

Database File*: 20180608_Uniprot_Canonical_RT_concat_target_decoy.fasta

Project Settings

Identification: 3_search_engines_50ppm_2020fasta

Project: Default

Processing: 48 cores

Please cite PeptideShaker as Vaudel et al.: Nature Biotechnol. 2015 Jan;33(1):22-24.

Figure 6 PeptideShaker “New Project” module to provide project details, input files, and search parameters.

6. Browse and locate the folder with the 10 spectrum files (`../HEK_suppdata/RAW_files/`) (Fig. 6). In the line marked “Spectrum Files(s).” These are the 10 MGF files from (Basic Protocol 1, step 2).

7. Click on Browse in the line marked “Database File” to set the identification parameters used during the searching/identification (Basic Protocol 1, step 6). This will lead to another module, “Identification Settings.” The parameters can be saved as *par* format for future use.

8. Under the “Project Settings” box, click on “Add” to specify the “Identification” parameters. The parameter file (`../HEK_suppdata/PeptideShaker Results/threesearchengine_50ppm.par`) can be imported by clicking the “Import from File” in “Identification Settings” module (Fig. 7).

   To change or set any of the individual parameters, follow the points 9-13. The settings are the same as described earlier in Basic Protocol 1 for SearchGUI (Fig. 4).

9. In “Identification Settings,” name the settings, e.g., “Three Search Engine.”

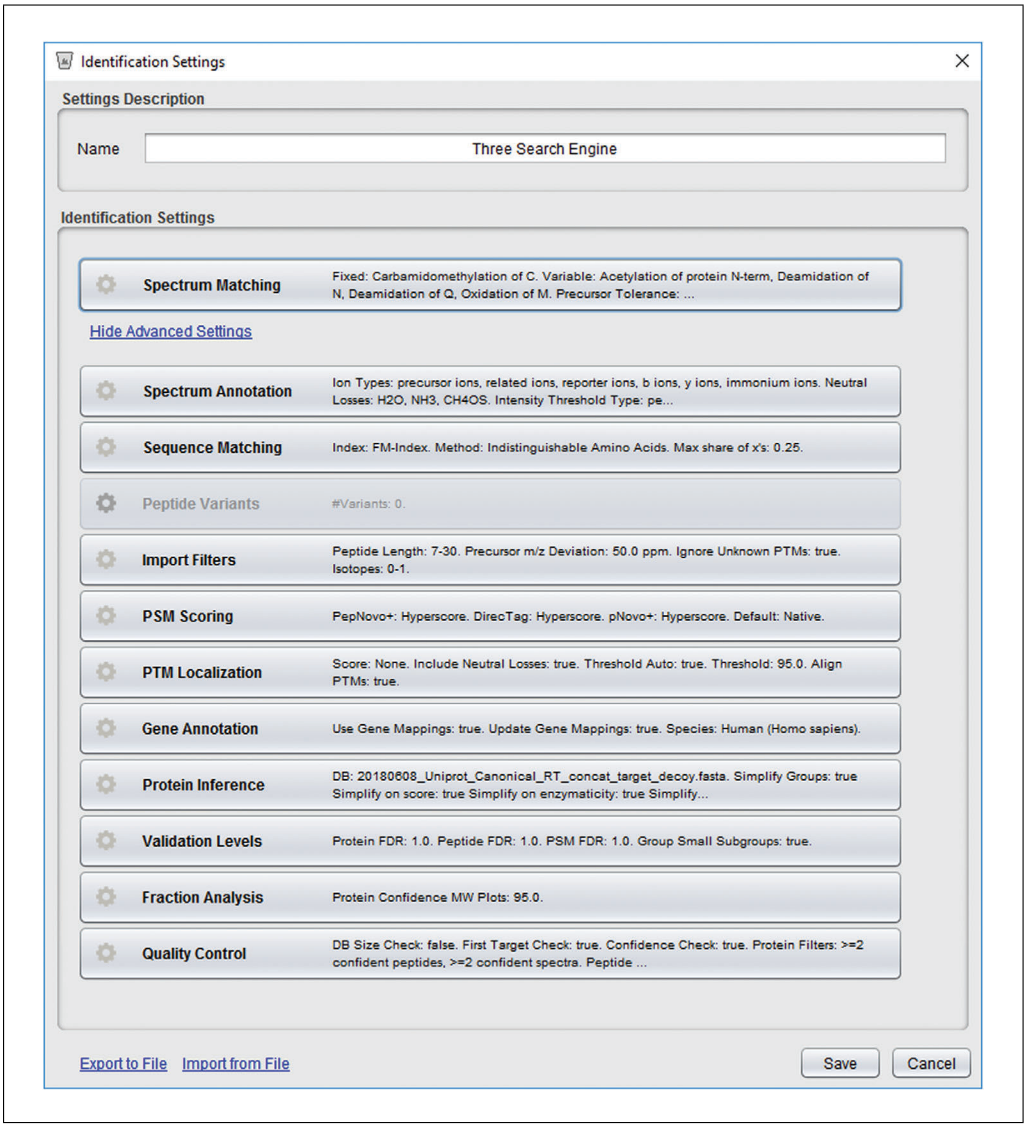
10. Click on “Spectrum Matching” to set up the settings (Fig. 7).

    Database: Select the same FASTA database as in Basic Protocol 1.
    Modifications: Select the same “Fixed Modifications” and “Variable Modifications” as in Basic Protocol 1, step 6.
    Protease & Fragmentation: Select the same as in Basic Protocol 1, step 6.

11. In addition to the “Spectrum Matching” settings, click “Show Advanced Settings” to further specify spectrum and precursor/fragment settings (Fig. 7).

12. In “Import Filter” settings, set “Peptide Length” to a minimum of 7 amino acids (AA) and maximum 30 AA peptide length (Fig. 8A).

13. In “Validation Levels,” set the FDR to 1% at PSM, peptide, and protein levels (Fig. 8B).



Identification Settings

Settings Description

Name: Three Search Engine

Identification Settings

Spectrum Matching: Fixed: Carbamidomethylation of C. Variable: Acetylation of protein N-term, Deamidation of N, Deamidation of Q, Oxidation of M. Precursor Tolerance: ...

Hide Advanced Settings

Spectrum Annotation: Ion Types: precursor ions, related ions, reporter ions, b ions, y ions, immonium ions. Neutral Losses: H2O, NH3, CH4OS. Intensity Threshold Type: pe...

Sequence Matching: Index: FM-Index. Method: Indistinguishable Amino Acids. Max share of x's: 0.25.

Peptide Variants: #Variants: 0.

Import Filters: Peptide Length: 7-30. Precursor m/z Deviation: 50.0 ppm. Ignore Unknown PTMs: true. Isotopes: 0-1.

PSM Scoring: PepNovo+: Hyperscore. DirecTag: Hyperscore. pNovo+: Hyperscore. Default: Native.

PTM Localization: Score: None. Include Neutral Losses: true. Threshold Auto: true. Threshold: 95.0. Align PTMs: true.

Gene Annotation: Use Gene Mappings: true. Update Gene Mappings: true. Species: Human (Homo sapiens).

Protein Inference: DB: 20180608_Uniprot_Canonical_RT_concat_target_decoy.fasta. Simplify Groups: true Simplify on score: true Simplify on enzymaticity: true Simplify...

Validation Levels: Protein FDR: 1.0. Peptide FDR: 1.0. PSM FDR: 1.0. Group Small Subgroups: true.

Fraction Analysis: Protein Confidence MW Plots: 95.0.

Quality Control: DB Size Check: false. First Target Check: true. Confidence Check: true. Protein Filters: >=2 confident peptides, >=2 confident spectra. Peptide ...

Figure 7 PeptideShaker “Identification Settings” module to provide “Spectrum Matching” parameters and “Advanced Settings.” The parameter file can be imported by “Import from File.”

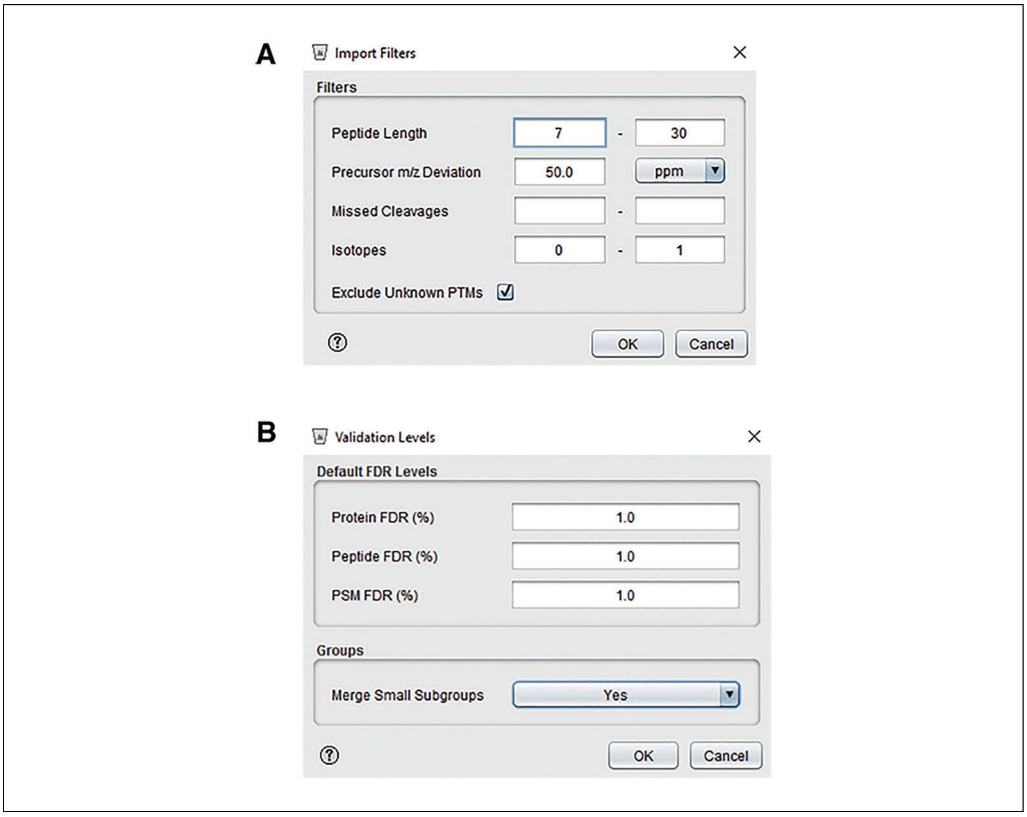
14. Go back to “PeptideShaker–New Project” module and click “Load Data!”. This will load the input files, start the data processing, and perform the merging and filtering of the data.

Output

15. Once completed, PeptideShaker will show the list of identified proteins and peptides, their precursor and fragment level spectra, and other spectral information in different tabs in the GUI (Fig. 9).

16. Save the PeptideShaker project in Compomics Peptide Shaker Format (*cpsx*). The already saved results can be found at (`../HEK_suppdata/PeptideShakerResults/`). These files can be reloaded in the PeptideShaker to visualize the results anytime later (Fig. 10).

17. In the PeptideShaker, click “Export Project” to export the results in *mzid* format (`../HEK_suppdata/PeptideShakerResults/..`) (Fig. 10). This file will be used as an input in Basic Protocol 3 to generate the final spectral library.



Import Filters

Filters

Peptide Length	7	-	30
Precursor m/z Deviation	50.0		ppm ▼
Missed Cleavages		-	
Isotopes	0	-	1
Exclude Unknown PTMs ☑			

Validation Levels

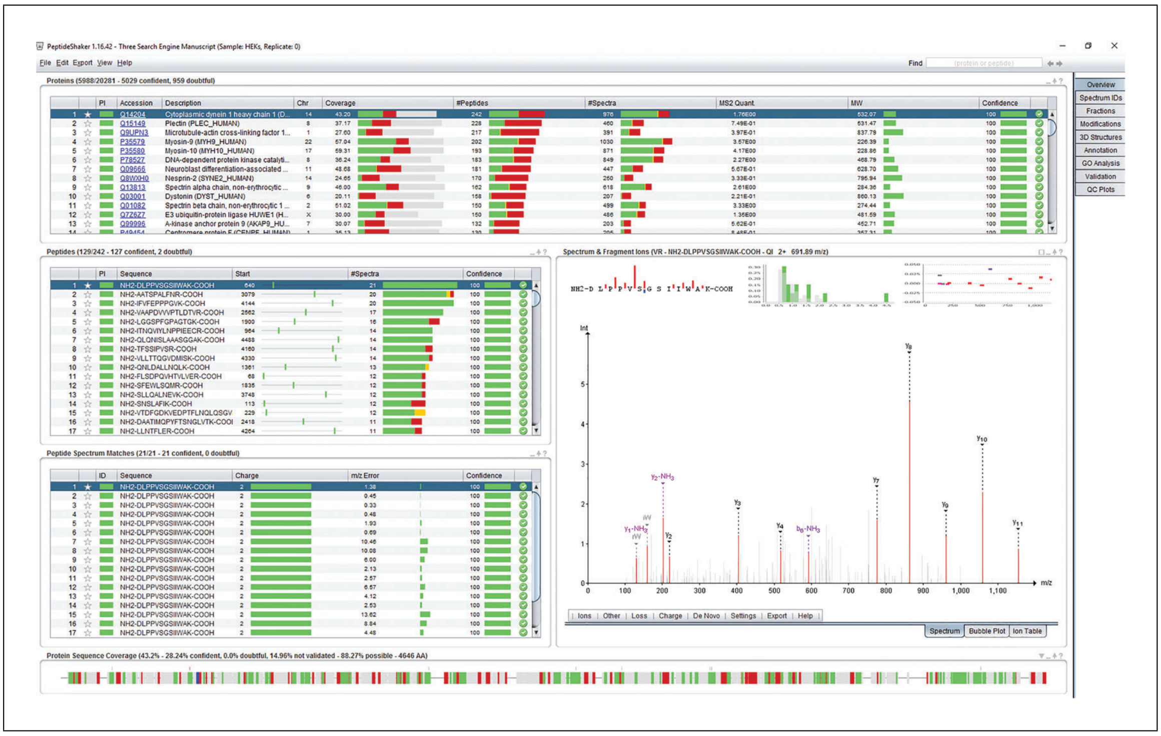
Default FDR Levels

Protein FDR (%)	1.0
Peptide FDR (%)	1.0
PSM FDR (%)	1.0

Groups

Merge Small Subgroups	Yes ▼

Figure 8 PeptideShaker "Advanced Settings" in "Identification Settings" to set (**A**) "Import Filters," which allow setting the minimum and maximum peptide length, missed cleavages, and isotopes of the peptide to include in the library, and (**B**) "Validation Levels," which allow setting the False Discovery Rate (FDR) at all protein, peptide, and PSM levels.


PeptideShaker 1.16.42 - Three Search Engine Manuscript (Sample: HEKs, Replicate: 0)

Proteins (5988/20281 - 5029 confident, 959 doubtful)

		PI	Accession	Description	Chr	Coverage	#Peptides	#Spectra	MS2 Quant.	MW	Confidence
1	★	🟩	Q14204	Cytoplasmic dynein 1 heavy chain 1 (D...	14	43.20	242	976	1.76E00	532.07	100
2	☆	🟩	Q15149	Plectin (PLEC_HUMAN)	8	37.17	228	460	7.49E-01	531.47	100
3	☆	🟩	Q9UPN3	Microtubule-actin cross-linking factor 1...	1	27.60	217	391	3.97E-01	837.79	100
4	☆	🟩	P35579	Myosin-9 (MYH9_HUMAN)	22	57.04	202	1030	3.57E00	226.39	100
5	☆	🟩	P35580	Myosin-10 (MYH10_HUMAN)	17	59.31	193	871	4.17E00	228.86	100
6	☆	🟩	P78527	DNA-dependent protein kinase catalyti...	8	36.24	183	849	2.27E00	468.79	100
7	☆	🟩	Q09666	Neuroblast differentiation-associated ...	11	48.68	181	447	5.67E-01	628.70	100
8	☆	🟩	Q8WXH0	Nesprin-2 (SYNE2_HUMAN)	14	24.65	170	250	3.33E-01	795.94	100
9	☆	🟩	Q13813	Spectrin alpha chain, non-erythrocytic ...	9	46.00	162	618	2.61E00	284.36	100
10	☆	🟩	Q03001	Dystonin (DYST_HUMAN)	6	20.11	158	207	2.21E-01	860.13	100
11	☆	🟩	Q01082	Spectrin beta chain, non-erythrocytic 1 ...	2	51.02	150	499	3.33E00	274.44	100
12	☆	🟩	Q7Z6Z7	E3 ubiquitin-protein ligase HUWE1 (H...	X	30.00	150	486	1.35E00	481.59	100
13	☆	🟩	Q99996	A-kinase anchor protein 9 (AKAP9_HU...	7	30.07	132	203	5.62E-01	452.71	100
14	☆	🟩	P49454	Centromere protein F (CENPF_HUMAN)	1	35.13	130	205	8.48E-01	357.31	100

Peptides (129/242 - 127 confident, 2 doubtful)

		PI	Sequence	Start	#Spectra	Confidence
1	★	🟩	NH2-DLPPVSGSIIWAK-COOH	640	21	100
2	☆	🟩	NH2-AATSPALFNR-COOH	3079	20	100
3	☆	🟩	NH2-IFVFEPPPGVK-COOH	4144	20	100
4	☆	🟩	NH2-VAAPDVVVPTLDTVR-COOH	2562	17	100
5	☆	🟩	NH2-LGGSPFGPAGTGK-COOH	1900	16	100
6	☆	🟩	NH2-ITNQVIYLNPPIEECR-COOH	964	14	100
7	☆	🟩	NH2-QLQNISLAAASGGAK-COOH	4488	14	100
8	☆	🟩	NH2-TFSSIPVSR-COOH	4160	14	100
9	☆	🟩	NH2-VLLTTQGVDMISK-COOH	4330	14	100
10	☆	🟩	NH2-QNLDALLNQLK-COOH	1361	13	100
11	☆	🟩	NH2-FLSDPQVHTVLVER-COOH	68	12	100
12	☆	🟩	NH2-SFEWLSQMR-COOH	1835	12	100
13	☆	🟩	NH2-SLLQALNEVK-COOH	3748	12	100
14	☆	🟩	NH2-SNSLAFIK-COOH	113	12	100
15	☆	🟩	NH2-VTDFGDKVEDPTFLNQLQSGV	229	12	100
16	☆	🟩	NH2-DAATIMQPYFTSNGLVTK-COOI	2418	11	100
17	☆	🟩	NH2-LLNTFLER-COOH	4264	11	100

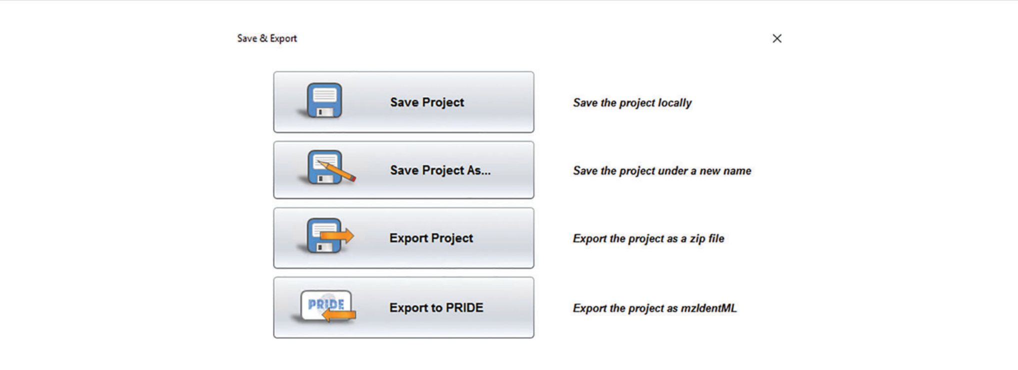
Peptide Spectrum Matches (21/21 - 21 confident, 0 doubtful)

		ID	Sequence	Charge	m/z Error	Confidence
1	★	🟩	NH2-DLPPVSGSIIWAK-COOH	2	1.38	100
2	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	0.45	100
3	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	0.33	100
4	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	0.48	100
5	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	1.93	100
6	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	0.69	100
7	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	10.46	100
8	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	10.08	100
9	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	6.00	100
10	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	2.13	100
11	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	2.57	100
12	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	6.57	100
13	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	4.12	100
14	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	2.53	100
15	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	13.62	100
16	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	8.84	100
17	☆	🟩	NH2-DLPPVSGSIIWAK-COOH	2	4.48	100

Spectrum & Fragment Ions (VR - NH2-DLPPVSGSIIWAK-COOH - QI 2+ 691.89 m/z)

Protein Sequence Coverage (43.2% - 28.24% confident, 0.0% doubtful, 14.96% not validated - 88.27% possible - 4646 AA)

Figure 9 PeptideShaker results interface showing detailed results at protein, peptide, and PSM level. Protein coverage, peptide confidence, and fragment level spectra can be visualized in the main interface.



Save & Export

Save Project — Save the project locally

Save Project As... — Save the project under a new name

Export Project — Export the project as a zip file

Export to PRIDE — Export the project as mzIdentML

Figure 10 PeptideShaker interface to save and export the results. This module allows saving the project in *cpsx* and zipped format. The merged results can be exported in *mzid* format, which is compatible with the PRIDE repository.

CREATING SPECTRAL LIBRARIES FROM MERGED RESULTS

BASIC PROTOCOL 3

The final step of the procedure consists of generating the spectral library from the merged *mzid* file from Basic Protocol 2. Here, we use the Skyline (MacLean et al., 2010) interface to generate the library from the output of PeptideShaker. Skyline provides a detailed set of parameters for precursor and fragment ions, along with their charges and modifications being included in the library. Moreover, it has a module to calibrate the retention time using standard RT peptides, either pre-defined or set by the user. The final spectral library with the calibrated retention time can be exported from Skyline to different formats, and can be directly incorporated into a range of DIA-MS data analysis tools.

Necessary Resources

Hardware

A computer with Windows 10 or later, preferably a workstation
A minimum of 16 GB RAM

Software

Skyline (download the latest version from the link provided in Internet Resources)

Input files

MGF files and location
Merged result file from PeptideShaker (Basic Protocol 2)

1. Open Skyline and create a new document from the “File” menu.

2. Before importing and building the library from the *mzid* file from Basic Protocol 2, “Peptide” and “Transition” settings need to be set using the “Settings” menu, which defines what precursor and fragment ions should be included in the library.

   The peptide settings specified below are specific to the example provided in this study. Based on these settings, researchers are advised to adjust the settings for their projects accordingly.

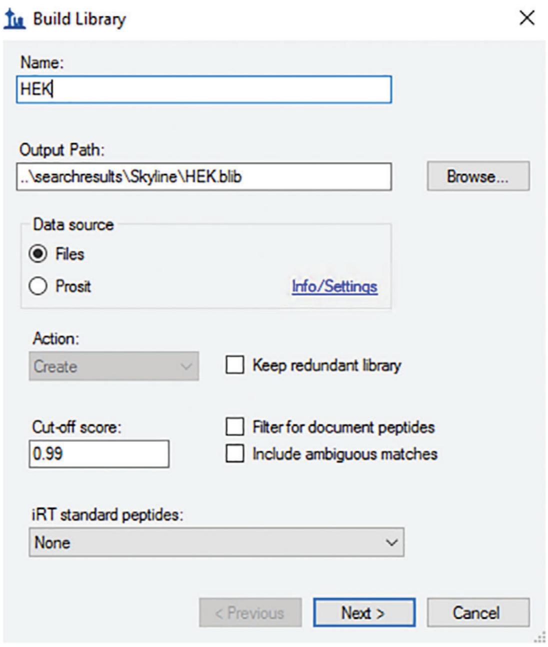
3. To enter parameters for peptides and precursors (Table 1), select “Settings” > “Peptide Settings.”

4. To enter parameters for fragment ions, select “Settings” > “Transition Settings” (see Table 2).

   The transition settings specified below are specific to acquisition settings for the SCIEX instrument in this study. Researchers are advised to adjust the settings according to the acquisition method in their experiment.

Table 1 List of Peptide Settings in Skyline

Peptide settings		
Digestion	Enzyme	Trypsin [KR I P]
	Max missed cleavages	2
	Background proteome	Same as Basic Protocol 1
Filter	Min length	7
	Max length	40
	Exclude N terminal AAs	0
	Exclude potential ragged ends	Unchecked
	Exclude peptides containing	(Blank)
	Auto-select all matching peptides	Checked
Modifications	Structural modifications	Carbamidomethyl (C), Oxidation (M)
	Max variable mods	3
	Max neutral losses	1
	Isotope label type	Light
	Isotope modifications	(Blank)
	Internal standard type	None



Build Library

Name: HEK

Output Path: ..\searchresults\Skyline\HEK.blib

Data source

● Files

○ Prosit

Action: Create

□ Keep redundant library

Cut-off score: 0.99

□ Filter for document peptides

□ Include ambiguous matches

iRT standard peptides: None

Figure 11 Skyline module for building the library from PeptideShaker results. The confidence-score cut-off and standard RT peptides can be selected in this module.

5. The next step is to build the library. For this, go to the “Settings” > “Peptide Settings” > “Build Library” module (Fig. 11) and fill it as follows:

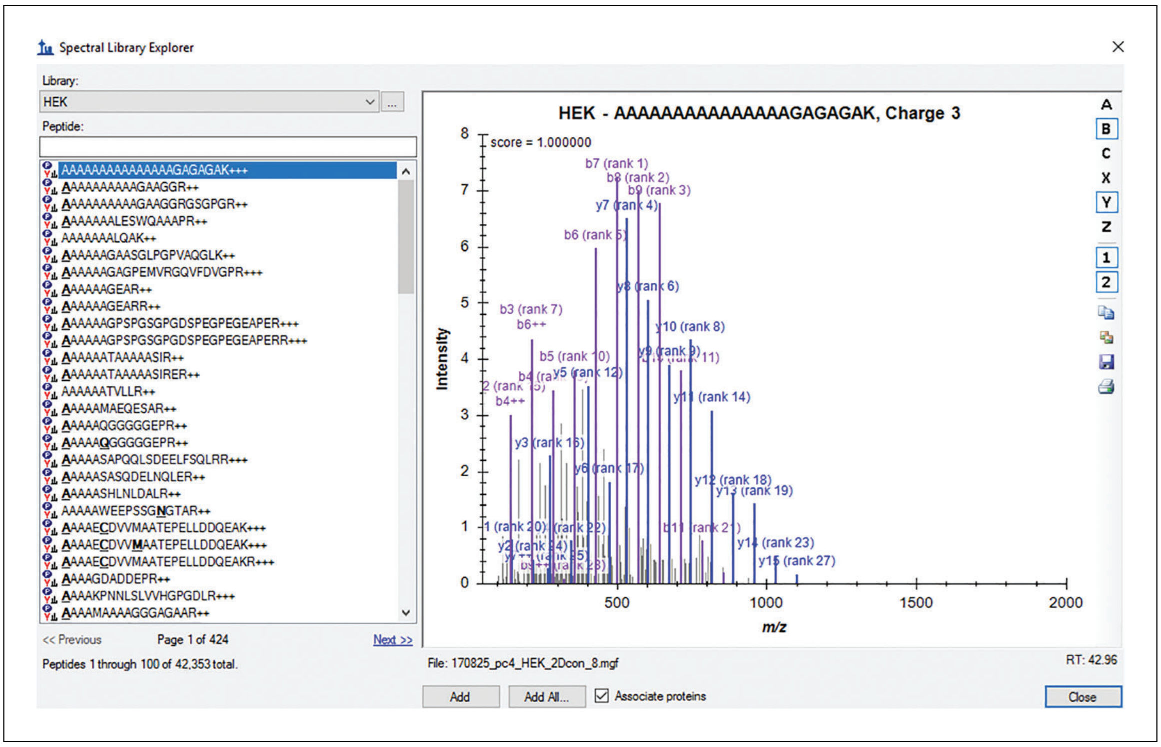
   “Name”: Provide the name for the library.
   “Output Path”: Set the output path where the library files are saved.
   “Data source”: Set the data source as “Files.”
   “Cut-off score”: Set the cut-off score as 0.99.
   Uncheck “Keep redundant library.”

Table 2 List of Transition Settings in Skyline

Transition settings		
Filter	Precursor charges	2, 3, 4, 5
	Ion charges	1, 2
	Ion types	y, b
	Product ions from	*m/z* > precursor
	Product ions to	Last ion -1
	Product ions–Special ions	(Blank)
	Use DIA precursor window for exclusion	(Blank)
	Auto select all matching transitions	Checked
Library	Ion match tolerance	0.05
	If a library spectrum is available, pick its most intense ions	Checked
	Pick XX product ions	12
	Pick XX minimum product ions	(Blank)
	From filtered ions charges and types	Checked
	From filtered ions charges and types plus filtered product ions	Unchecked
	From filtered product ions	Unchecked
Instrument	Min *m/z*	200
	Max *m/z*	2000
	Dynamic min product *m/z*	Unchecked
	Method match tolerance *m/z*	0.055
	Firmware	(Blank)
	Firmware	(Blank)
	Min time	(Blank)
	Max time	(Blank)

6. After filling these settings, press “Next.” In the “Build Library” module, click “Add Files” and import the *mzid* file from the PeptideShaker output files (`../HEK_suppdata/PeptideShakerResults/..`). Then, click “Finish.”
7. Skyline will start reading and importing the *mzid* file, and the status can be seen at the bottom left of the Skyline interface.
8. After it finishes reading the file, the “Spectral Library Explorer” module will appear. This module can also be accessed from “View” > “Spectral Libraries.” “Spectral Library Explorer” also shows the list of those modifications that are found in the peptides in addition to those already defined in Table 1 using a separate module called “Add Modifications.” In “Spectral Library Explorer,” each peptide and its corresponding spectrum can be visualized (Fig. 12). It will generate the spectral library in *blib* format (`../HEK_suppdata/Skyline/..`).

   For simplicity, we have selected unmodified peptides only [excluding Carbamidomethylation (C)]. Researchers can select the modifications of their interest based on the experimental design and biological question of interest.



Spectral Library Explorer

Library: HEK

Peptide:

AAAAAAAAAAAAAAAGAGAGAK+++
AAAAAAAAAAGAAGGR++
AAAAAAAAAAGAAGGRGSGPGR++
AAAAAAALESWQAAAPR++
AAAAAAALQAK++
AAAAAGAASGLPGPVAQGLK++
AAAAAGAGPEMVRGQVFDVGPR+++
AAAAAGEAR++
AAAAAGEARR++
AAAAAGPSPGSGPGDSPEGPEGEAPER+++
AAAAAGPSPGSGPGDSPEGPEGEAPERR+++
AAAAATAAAAASIR++
AAAAATAAAAASIRER++
AAAAATVLLR++
AAAAMAEQESAR++
AAAAQGGGGGEPR++
AAAAQGGGGGEPR++
AAAASAPQQLSDEELFSQLRR+++
AAAASASQDELNQLER++
AAAASHLNLDALR++
AAAAWEEPSSGNGTAR++
AAAECDVVMAATEPELLDDQEAK+++
AAAECDVVMAATEPELLDDQEAK+++
AAAECDVVMAATEPELLDDQEAKR+++
AAAGDADDEPR++
AAAKPNNLSLVVHGPGDLR+++
AAAMAAAAGGGAGAAR++

<< Previous Page 1 of 424 Next >>

Peptides 1 through 100 of 42,353 total.

File: 170825_pc4_HEK_2Dcon_8.mgf

RT: 42.96

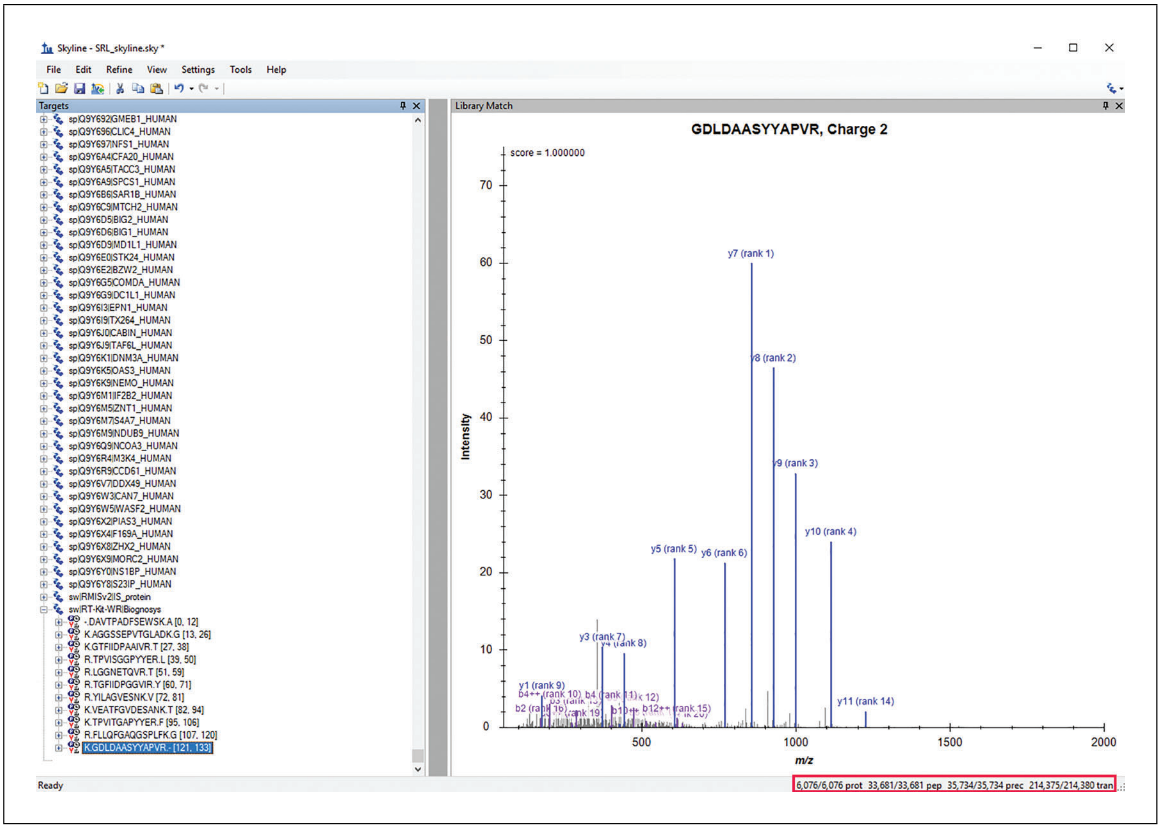
Figure 12 Skyline module for spectral library explorer. Using this explorer, fragment spectra for each peptide in each library can be visualized.

9. To export this library from Skyline, these peptides would need to be added to the target list in Skyline. To add the peptides to the target list, click “Associate Proteins” and “Add All” in “Spectral Library Explorer” module. During the process of adding peptides to the target list, Skyline will notify if any peptides belong to more than one protein. Click “Do not Add” and click “OK” to add all the unique peptides and associated proteins in the target list. The number of proteins, peptides, precursors, and transitions can be visualized at the bottom right of the Skyline interface (Fig. 13).

10. To perform the retention time calibration and generate indexed retention time (iRT) peptides, go to the “Settings” > “Peptide Settings” > “Prediction” tab. In the “Retention Time Predictor” module, click on the small calculator symbol and click “Add” to add a new calculator to predict the retention time of the peptides (Fig. 14A). In the “Edit iRT Calculator” module, fill as follows:

    Name: Provide the name for the calculator.
    iRT database: Provide the path where the iRT database with predicted retention time values is saved.
    iRT standards: Either select from the given set of standard peptides or click on “Add” to add the list of your standard peptides.
    Other iRT values: Add all the library peptides in the calculator to retrieve their iRTs (indexed retention times) by clicking “Add” and “Add Spectral Library.”

    After adding all the peptides, a dialog box will appear. It shows that the peptides have been added successfully along with the regression model. A plot of actual and predicted retention times can be visualized by clicking “Success” in this dialog box. Click “Ok” to finish creating the calculator. The resulted calculator in *irtdb* format will also be saved in the same folder (`../HEK_suppdata/Skyline/..`).



sp|Q9Y692|GMEB1_HUMAN
sp|Q9Y696|CLIC4_HUMAN
sp|Q9Y697|NFS1_HUMAN
sp|Q9Y6A4|CFA20_HUMAN
sp|Q9Y6A5|TACC3_HUMAN
sp|Q9Y6A9|SPCS1_HUMAN
sp|Q9Y6B6|SAR1B_HUMAN
sp|Q9Y6C9|MTCH2_HUMAN
sp|Q9Y6D5|BIG2_HUMAN
sp|Q9Y6D6|BIG1_HUMAN
sp|Q9Y6D9|MD1L1_HUMAN
sp|Q9Y6E0|STK24_HUMAN
sp|Q9Y6E2|BZW2_HUMAN
sp|Q9Y6G5|COMDA_HUMAN
sp|Q9Y6G9|DC1L1_HUMAN
sp|Q9Y6I3|EPN1_HUMAN
sp|Q9Y6I9|TX264_HUMAN
sp|Q9Y6J0|CABIN_HUMAN
sp|Q9Y6J9|TAF6L_HUMAN
sp|Q9Y6K1|DNM3A_HUMAN
sp|Q9Y6K5|OAS3_HUMAN
sp|Q9Y6K9|NEMO_HUMAN
sp|Q9Y6M1|IF2B2_HUMAN
sp|Q9Y6M5|ZNT1_HUMAN
sp|Q9Y6M7|S4A7_HUMAN
sp|Q9Y6M9|NDUB9_HUMAN
sp|Q9Y6Q9|NCOA3_HUMAN
sp|Q9Y6R4|M3K4_HUMAN
sp|Q9Y6R9|CCD61_HUMAN
sp|Q9Y6V7|DDX49_HUMAN
sp|Q9Y6W3|CAN7_HUMAN
sp|Q9Y6W5|WASF2_HUMAN
sp|Q9Y6X2|PIAS3_HUMAN
sp|Q9Y6X4|F169A_HUMAN
sp|Q9Y6X8|ZHX2_HUMAN
sp|Q9Y6X9|MORC2_HUMAN
sp|Q9Y6Y0|NS1BP_HUMAN
sp|Q9Y6Y8|S23IP_HUMAN
sw|RMISv2|IS_protein
sw|RT-Kit-WR|Biognosys
-.DAVTPADFSEWSK.A [0, 12]
K.AGGSSEPVTGLADK.G [13, 26]
K.GTFIIDPAAIVR.T [27, 38]
R.TPVISGGPYYER.L [39, 50]
R.LGGNETQVR.T [51, 59]
R.TGFIIDPGGVIR.Y [60, 71]
R.YILAGVESNK.V [72, 81]
K.VEATFGVDESANK.T [82, 94]
K.TPVITGAPYYER.F [95, 106]
R.FLLQFGAQGSPLFK.G [107, 120]
K.GDLDAASYYAPVR.- [121, 133]

6,076/6,076 prot 33,681/33,681 pep 35,734/35,734 prec 214,375/214,380 tran

Figure 13 Skyline module for the target list. Proteins, peptides, and transitions added from the library can be visualized here along with the spectra.

11. To retrieve these iRTs, the last step is to add a predictor. For this, go to “Settings” > “Peptide Settings” > “Prediction” tab. In “Retention Time Predictor” module (Fig. 14B), click on the drop-down menu and click “Add” and fill as follows:

    Name: Provide the name for the predictor.
    Calculator: Select the calculator created in the previous step.

Output

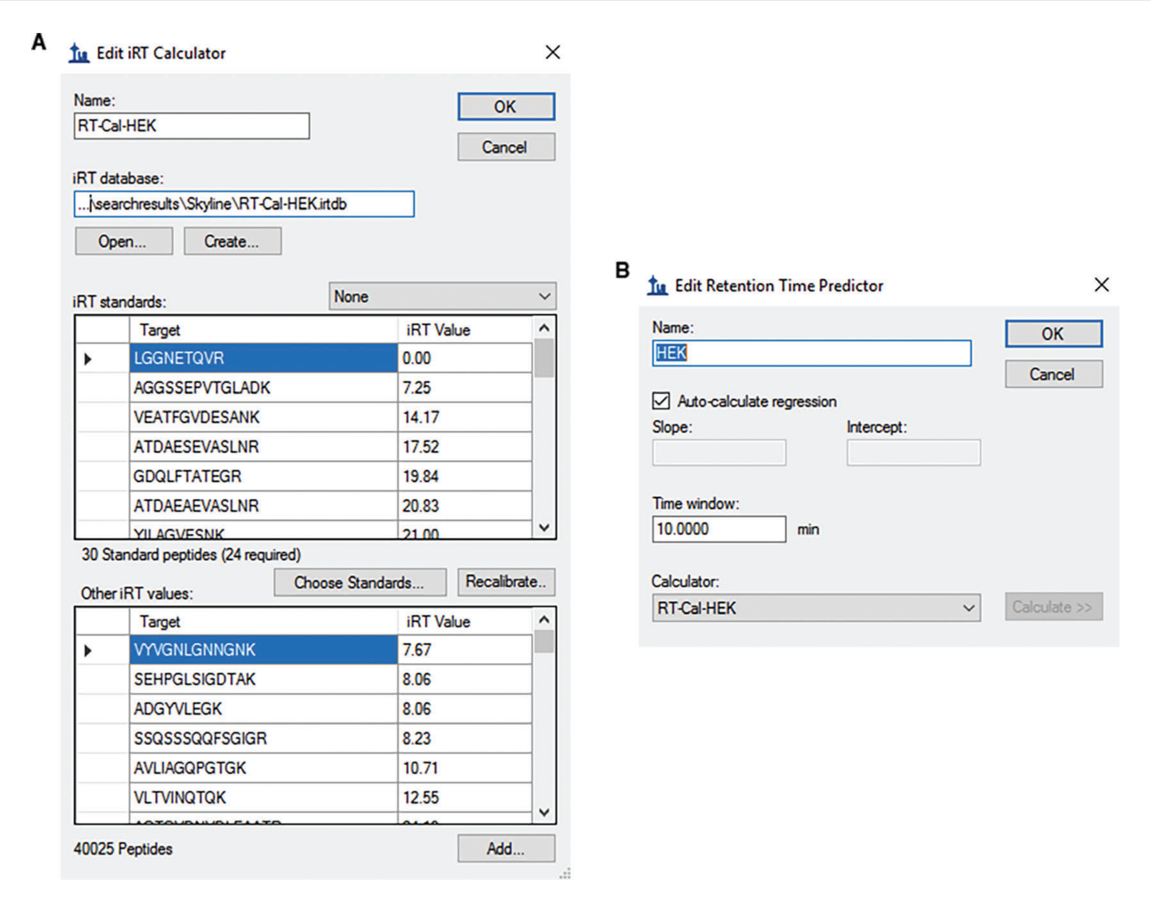
12. To export the library from Skyline, go to “File” > “Export” > “Report.” Skyline has many report templates, which can be modified and downloaded, e.g., the Spectronaut version can be downloaded from *https://biognosys.com/media.ashx/spectronautlibrary.skyr*. Here, we have used the OpenSWATH template (`OpenSWATH.skyr`) downloaded from *http://openswath.org/en/latest/docs/skyline.html* (`../HEK_supp/Skyline/..`) (Fig. 15).

13. This library can now be used with OpenSWATH or converted and used in any DDA-MS library-based DIA-MS data analysis software, including but not limited to PeakView, Spectronaut, Skyline, and DIA-NN (Demichev et al., 2020).

USING COMMAND-LINE INTERFACE (CLI) FOR AUTOMATING TASKS

ALTERNATE PROTOCOL

Users can perform all of the steps in Basic Protocols 1 and 2 using the CLI of MSConvert, SearchGUI, and PeptideShaker, which requires some basic knowledge of shell scripting. This protocol is recommended when researchers have a large number of DDA-MS raw files and aim to automate the process of library generation. Once all the parameters and settings are optimized by the users, they can be used in the CLI.



Name: RT-Cal-HEK

iRT database: ...\searchresults\Skyline\RT-Cal-HEK.irtdb

iRT standards: None

Target	iRT Value
LGGNETQVR	0.00
AGGSSEPVTGLADK	7.25
VEATFGVDESANK	14.17
ATDAESEVASLNR	17.52
GDQLFTATEGR	19.84
ATDAEAEVASLNR	20.83
YILAGVESNK	21.00

30 Standard peptides (24 required)

Other iRT values:

Target	iRT Value
VYVGNLGNNGNK	7.67
SEHPGLSIGDTAK	8.06
ADGYVLEGK	8.06
SSQSSSQQFSGIGR	8.23
AVLIAGQPGTGK	10.71
VLTVINQTQK	12.55

40025 Peptides

Name: HEK

☑ Auto-calculate regression

Slope:

Intercept:

Time window: 10.0000 min

Calculator: RT-Cal-HEK

Figure 14 Skyline module to create the (**A**) retention time calculator, which provides both the already defined sets of standard iRT peptides and the option to set the user-defined peptides to use in the calculator, and (**B**) retention time predictor based on a retention time calculator.

Necessary Resources

Hardware

Same as in Basic Protocols 1 and 2

Software (Converting)

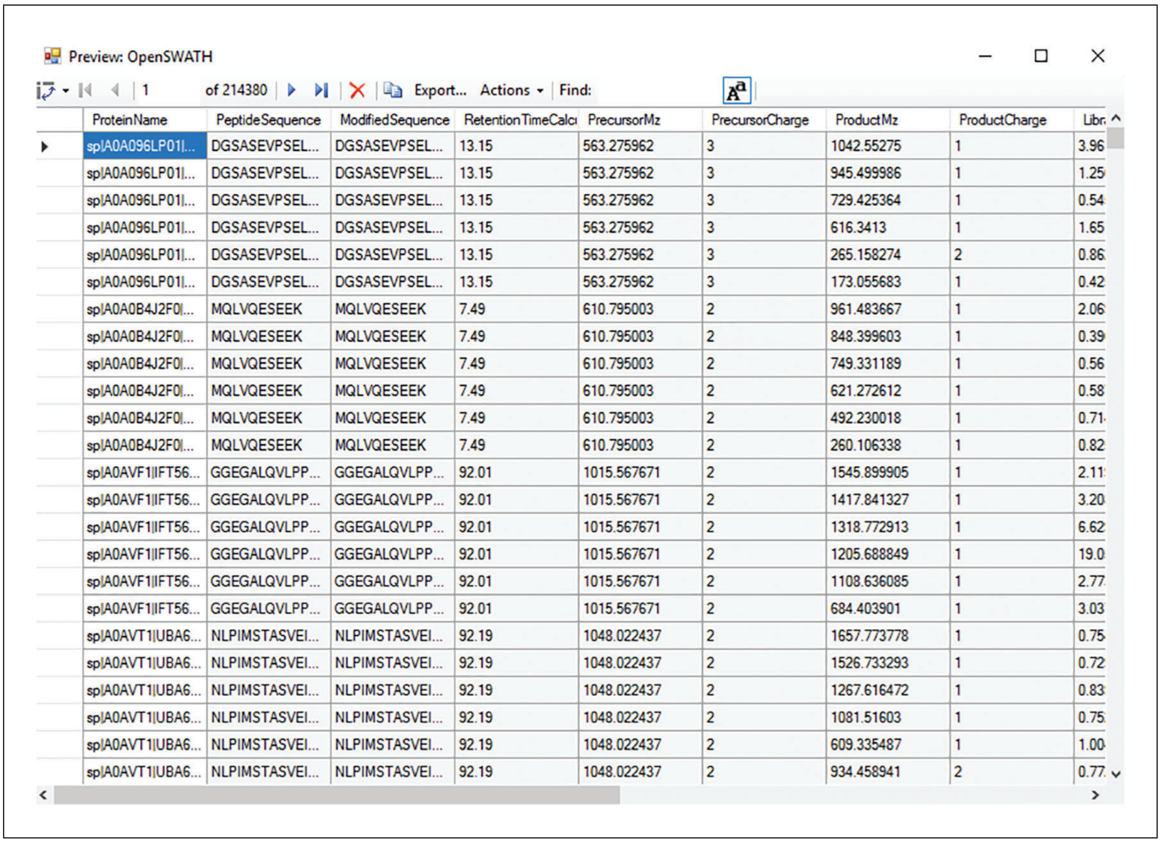
MSConvert
Access to CLI on Windows

Software (Searching)

SearchGUI
Java version 8 or higher
Spectrum files (MGF format)
Parameter file (*par*)
Access to CLI on Windows or Linux

Software (Merging and Export)

Peptideshaker
Java version 8 or higher
MGF files and location
Search results and location
Parameter file (*par*)
Concatenated database (FASTA format)



	ProteinName	PeptideSequence	ModifiedSequence	RetentionTimeCalc	PrecursorMz	PrecursorCharge	ProductMz	ProductCharge	Libr
▶	sp|A0A096LP01|...	DGSASEVPSEL...	DGSASEVPSEL...	13.15	563.275962	3	1042.55275	1	3.96
	sp|A0A096LP01|...	DGSASEVPSEL...	DGSASEVPSEL...	13.15	563.275962	3	945.499986	1	1.25
	sp|A0A096LP01|...	DGSASEVPSEL...	DGSASEVPSEL...	13.15	563.275962	3	729.425364	1	0.54
	sp|A0A096LP01|...	DGSASEVPSEL...	DGSASEVPSEL...	13.15	563.275962	3	616.3413	1	1.65
	sp|A0A096LP01|...	DGSASEVPSEL...	DGSASEVPSEL...	13.15	563.275962	3	265.158274	2	0.86
	sp|A0A096LP01|...	DGSASEVPSEL...	DGSASEVPSEL...	13.15	563.275962	3	173.055683	1	0.42
	sp|A0A0B4J2F0|...	MQLVQESEEK	MQLVQESEEK	7.49	610.795003	2	961.483667	1	2.06
	sp|A0A0B4J2F0|...	MQLVQESEEK	MQLVQESEEK	7.49	610.795003	2	848.399603	1	0.39
	sp|A0A0B4J2F0|...	MQLVQESEEK	MQLVQESEEK	7.49	610.795003	2	749.331189	1	0.56
	sp|A0A0B4J2F0|...	MQLVQESEEK	MQLVQESEEK	7.49	610.795003	2	621.272612	1	0.58
	sp|A0A0B4J2F0|...	MQLVQESEEK	MQLVQESEEK	7.49	610.795003	2	492.230018	1	0.71
	sp|A0A0B4J2F0|...	MQLVQESEEK	MQLVQESEEK	7.49	610.795003	2	260.106338	1	0.82
	sp|A0AVF1|IFT56...	GGEGALQVLPP...	GGEGALQVLPP...	92.01	1015.567671	2	1545.899905	1	2.11
	sp|A0AVF1|IFT56...	GGEGALQVLPP...	GGEGALQVLPP...	92.01	1015.567671	2	1417.841327	1	3.20
	sp|A0AVF1|IFT56...	GGEGALQVLPP...	GGEGALQVLPP...	92.01	1015.567671	2	1318.772913	1	6.62
	sp|A0AVF1|IFT56...	GGEGALQVLPP...	GGEGALQVLPP...	92.01	1015.567671	2	1205.688849	1	19.0
	sp|A0AVF1|IFT56...	GGEGALQVLPP...	GGEGALQVLPP...	92.01	1015.567671	2	1108.636085	1	2.77
	sp|A0AVF1|IFT56...	GGEGALQVLPP...	GGEGALQVLPP...	92.01	1015.567671	2	684.403901	1	3.03
	sp|A0AVT1|UBA6...	NLPIMSTASVEI...	NLPIMSTASVEI...	92.19	1048.022437	2	1657.773778	1	0.75
	sp|A0AVT1|UBA6...	NLPIMSTASVEI...	NLPIMSTASVEI...	92.19	1048.022437	2	1526.733293	1	0.72
	sp|A0AVT1|UBA6...	NLPIMSTASVEI...	NLPIMSTASVEI...	92.19	1048.022437	2	1267.616472	1	0.83
	sp|A0AVT1|UBA6...	NLPIMSTASVEI...	NLPIMSTASVEI...	92.19	1048.022437	2	1081.51603	1	0.75
	sp|A0AVT1|UBA6...	NLPIMSTASVEI...	NLPIMSTASVEI...	92.19	1048.022437	2	609.335487	1	1.00
	sp|A0AVT1|UBA6...	NLPIMSTASVEI...	NLPIMSTASVEI...	92.19	1048.022437	2	934.458941	2	0.77

Figure 15 Preview of the columns to be included in the library file while exporting from Skyline using the “Export Report” module of the software.

Access to CLI on Windows or Linux
PeptideShaker project file (cpsx)

Converting raw files to MGF format

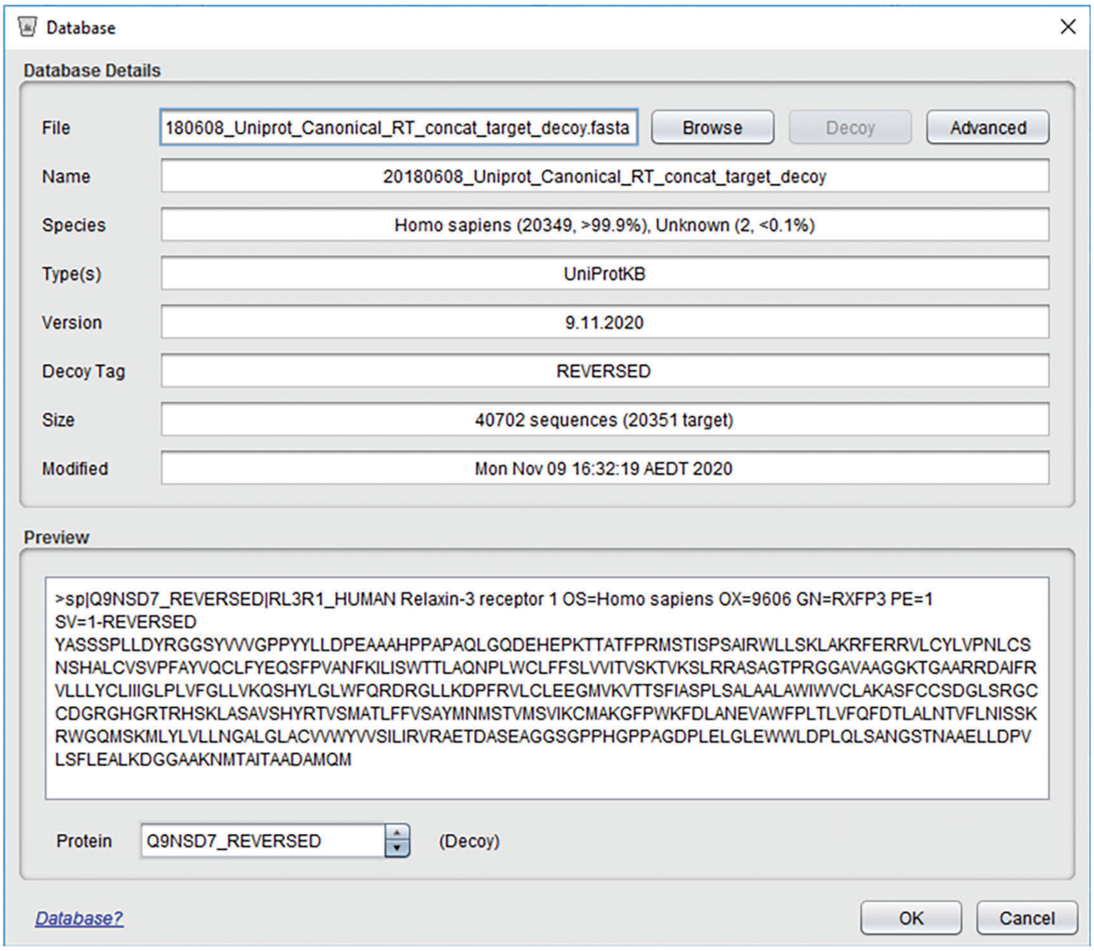
1. On a Windows machine, open “Windows” > “Command Prompt” and navigate to the folder containing the installation of MSConvert, usually at `C:\Program Files (x86)\ProteoWizard 3.0.18351 64-bit\` on a 64-bit machine.

   ```
   $ cd C: \Program Files (x86)\ProteoWizard 3.0.18351
     64-bit\
   ```

2. Run the following command in the folder (on a Windows machine), assuming the *wiff* files are in the location `HEK_suppdata/RAW_files/` and the desired output folder is `HEK_suppdata/RAW_files/`.

   ```
   $ msconvert.exe HEK_suppdata/RAW_files/170825_pc4_
     HEK_2Dcon_10.wiff --mgf --mz64 -z -e .mgf -o
     /home/files/ --filter "peakPicking true 1-" --filter=
     "titleMaker <RunId>.<ScanNumber>.<ScanNumber>.
     <ChargeState> File:<SourcePath>, NativeID:<Id>"
     --filter="chargeStatePredictor maxMultipleCharge=5
     minMultipleCharge=2 singleChargeFractionTIC=0.9"
   ```

   The above command can only be executed in Windows, as it requires proprietary format conversion. The vendor dll files can only be accessible on a Windows machine. The above command will generate a MGF file with the same name as the wiff file in the desired location. This process can be executed for all the files in the folder using a simple loop.



Database Details

File: 180608_Uniprot_Canonical_RT_concat_target_decoy.fasta

Name: 20180608_Uniprot_Canonical_RT_concat_target_decoy

Species: Homo sapiens (20349, >99.9%), Unknown (2, <0.1%)

Type(s): UniProtKB

Version: 9.11.2020

Decoy Tag: REVERSED

Size: 40702 sequences (20351 target)

Modified: Mon Nov 09 16:32:19 AEDT 2020

Preview

```
>sp|Q9NSD7_REVERSED|RL3R1_HUMAN Relaxin-3 receptor 1 OS=Homo sapiens OX=9606 GN=RXFP3 PE=1
SV=1-REVERSED
YASSSPLLDYRGGSYVVVGPPYYLLDPEAAAHPPAPAQLGQDEHEPKTTATFPRMSTISPSAIRWLLSKLAKRFERRVLCYLVPNLCS
NSHALCVSVPFAYVQCLFYEQSFPVANFKILISWTTLAQNPLWCLFFSLVVITVSKTVKSLRRASAGTPRGGAVAAGGKTGAARRDAIFR
VLLLYCLIIIGLPLVFGLLVKQSHYLGLWFQRDRGLLKDPFRVLCLEEGMVKVTTSFIASPLSALAALAWIWVCLAKASFCCSDGLSRGC
CDGRGHGRTRHSKLASAVSHYRTVSMATLFFVSAYMNMSTVMSVIKCMAKGFPWKFDLANEVAWFPLTLVFQFDTLALNTVFLNISSK
RWGQMSKMLYLVLLNGALGLACVVWYVVSILIRVRAETDASEAGGSGPPHGPPAGDPLELGLEWWLDPLQLSANGSTNAAELLDPV
LSFLEALKDGGAAKNMTAITAADAMQM
```

Protein: Q9NSD7_REVERSED (Decoy)

Figure 16 Preview of the concatenated target/decoy database used for performing searches, displaying the type and version of database and number of sequences stored in the database.

Searching using SearchGUI

3. SearchGUI contains an built-in CLI called SearchCLI. To run SearchCLI, navigate to the folder containing the SearchGUI installation. The following commands assume that all converted MGF files are in the location `HEK_suppdata/RAW_files/`. The parameter file is the same as used in Basic Protocols 1 and 2.

   ```
   $ cd C: \Program Files (x86) \SearchGUI-3.3.20\
   ```

   ```
   $ java -cp SearchGUI-3.3.20.jar eu.isas.searchgui.
     cmd.SearchCLI -spectrum_files HEK_suppdata/RAW_
     files/ -output_folder /home/files/ -id_params
     HEK_suppdata/threesearchengine_50ppm.par -xtandem
     1 -msgf 1 -comet 1 -output_option 3
   ```

   Run the command without any arguments to access additional parameters or to choose different search engines. These commands can be run on either a Windows or Linux installation of SearchGUI.

Merging search results using PeptideShakerCLI

4. Assuming the MGF files are located in `HEK_suppdata/RAW_files/` and search results from an earlier step are in `HEK_suppdata/PeptideShakerResults/`, navigate to the folder containing the PeptideShaker installation and type the following:

```
$ java -cp PeptideShaker-X.Y.Z.jar eu.isas.
  peptideshaker.cmd.PeptideShakerCLI -experiment HEK
  -sample HEK_samples -replicate 0 -identification_
  files HEK_suppdata/PeptideShakerResults/ -spectrum_
  files HEK_suppdata/RAW_files/ -out HEK_suppdata/
  PeptideShakerResults/ ThreeSearchEnginePepShaker.
  cpsx -log HEK_suppdata/srllog -db HEK_suppdata/
  FASTA_database /20180608_Uniprot_Canonical_RT_
  concat_target_decoy.Fasta -id_params HEK_suppdata/
  threesearchengine_50ppm.par
```

The `-experiment`, `-sample`, and `-replicate` are free text information, which can be added according to the user's experiment. `-log <dest>` is recommended, as it generates a log file of all commands executed and helps in debugging. This command generates a `peptideshaker` project file with a `cpsx` extension. This can be loaded into the GUI for any further analysis. These commands can be run on either a Windows or Linux installation of SearchGUI.

Exporting as mzid

5. The *cpsx* project file can be used with `peptideshakerMzidCLI` to export the results as *mzid*, which will contain results from all the searches conducted. This file can be further used to create the final spectral library. Navigate to the folder with the peptideshaker installation and run the following command:

```
$ java -cp PeptideShaker-X.Y.Z.jar eu.isas.
  peptideshaker.cmd.MzidCLI -in HEK_suppdata/
  PeptideShakerResults/ ThreeSearchEnginePepShaker.
  cpsx -output_file HEK_suppdata/PeptideShaker
  Results/ ThreeSearchEnginePepShaker.mzid
  -contact_first_name YourFirstName -contact_last_
  name YourLastName -contact_email yourname@
  university.edu -contact_address "Your address"
  -organization_name OrganizationName -organization_
  email "yourname@university.edu"
  -organization_address "Your Address"
```

This will generate a single mzid file from all the search results. Replace the personal details with ones pertaining to your experiment. The mzid file can then be used to follow Basic Protocol 3 as described earlier.

CREATING CONCATENATED FASTA FILES

SUPPORT PROTOCOL

SearchGUI provides a quick way to create a concatenated target/decoy database using the GUI or command line using FastaCLI.

Necessary Resources

Hardware

Same as in Basic Protocols 1 and 2

Software (Converting)

SearchGUI
Database (FASTA format)

1. To create a combined FASTA file using the GUI, open the SearchGUI "Spectrum Matching" settings as explained in Basic Protocol 1 (Fig. 4).

2. Click "Edit" on the "Database (FASTA)" and select the database file FASTA of interest. A prompt will appear `The selected FASTA file does not seem to contain decoy sequences. Add decoys?`. Click "Yes"

3. Decoy will be appended and the information about database details such as "Name," "Species," "Type(s)," "Version" are displayed as shown in Figure 16.
   To create the combined FASTA in CLI, navigate to the SearchGUI installation folder and execute:

   ```
   $ java -cp SearchGUI-X.Y.Z.jar eu.isas.searchgui.
     cmd.FastaCLI -in NameofFastafile -decoy
   ```

   The output will be a `NameofFastafile_concatenated_target_decoy.fasta`. This can be further used for all the analysis in all Basic Protocols and Alternate Protocol.

COMMENTARY

Background Information

DDA is a method where a fixed number of precursor ions are selected on the basis of abundance and analyzed by tandem MS, while DIA is an alternative approach that continuously acquires fragment-ion spectra in an unbiased fashion. SWATH-MS is a state-of-the-art DIA method, which allows fast mass spectrometric conversion of small amounts of tissue into a single, permanent digital file representing the quantitative proteome of a biological sample (Guo et al., 2015; Ludwig et al., 2018). This technique, uses peptide-centric scoring for large-scale identification and quantification of peptides and proteins on the basis of robustness, quantitative characteristics, and a high degree of reproducibility (Gillet et al., 2012; Ludwig et al., 2018).

A variety of strategies have been developed to analyze the SWATH-MS data, which include both spectrum-centric and peptide-centric methods (Ting et al., 2015). The peptide-centric approach relies on a high-quality spectral library. The spectral library contains the information on the *m/z* and LC retention times for all representative peptide features in the samples (Ludwig et al., 2018). The generation of these libraries usually requires acquisition of MS data in DDA mode under the same LC conditions as in DIA mode. The libraries are generated from pooled and fractionated samples or synthetic analogs of peptides of interest. A wide range of chromatographic chemistries are available for fractionation of pooled samples (Yeung et al., 2020). Spectral libraries can be sample-specific or generated using publicly available resources. The SWATHAtlas (*http://www.swathatlas.org*) is a publicly available resource, which contains published spectral libraries on several species including human, *E. coli*, and yeast. Since the instrument and LC conditions differ, the ideal practice is to generate a sample-specific library (Ludwig et al., 2018). A drawback of the spectral library approach is that peptides can be only identified when they are present in the library. As alternatives, spectral library—free or spectrum-centric approaches have been developed, which can generate libraries from the DIA-MS data without the need for any sample-specific libraries (Demichev et al., 2020; Tiwary et al., 2019; Tsou et al., 2015; Yang et al., 2020).

A typical procedure of spectral library generation can be broadly classified into four main steps: (1) searching the raw spectra against a database of interest, (2) merging of results from different searches and statistical validation, (3) retrieving confidently identified spectra and creating a consensus library, and (4) further quality filtering on the library. For large-scale studies, libraries generated on different mass spectrometers under different LC conditions, and from different biological samples, tend to have differences in their retention times. Such libraries can be merged by iSwathX (Noor, Mohamedali, & Ranganathan, 2020), which creates a single unified library with the retention time alignment. Over the past years, several software tools have been developed for the generation of spectral libraries, such as SpectraST (Lam et al., 2007), X!Hunter (Craig, Cortens, Fenyo, & Beavis, 2006), Bibliospec (Frewen, Merrihew, Wu, Noble, & MacCoss, 2006), Pepitome (Dasari et al., 2012), and MSPepSearch (*https://chemdata.nist.gov/*). These are mostly built for DDA-MS data analysis. Similar tools for DIA-MS were lacking before Schubert et al. (2015) provided detailed steps to generate spectral libraries using open-source tools

Table 3 Potential Errors and Possible Solutions

Problem	Possible cause	Solutions
SearchGUI stops/crashes while loading input files	Hardware issues/low computer RAM	A minimum of 16 GB RAM. Open “Edit” >”Java Settings,” increase the available memory, and restart.
PeptideShaker stops/crashes while loading input files		
While importing PeptideShaker results into the Skyline, “No MGF file available” error	Spectrum files in MGF format and PeptideShaker combined search file in *mzid* format are not present in the same folder	Place the three search engine result files and PeptideShaker merged file in the same location while importing in Skyline

such Trans-Proteomic Pipeline (Deutsch et al., 2010), ProteoWizard (Chambers et al., 2012), and OpenMS (Sturm et al., 2008). Some of these tools and steps are tedious and hard to execute without programming knowledge. Although commercial tools are user-friendly, they provide fewer controls over the workflow and require licenses. ProteinPilot-PeakView (Shilov et al., 2007) and Pulsar-Spectronaut (Bruderer et al., 2015) are two popular commercial products. We present here an easy-to-use procedure consisting of three protocols, which generates a high-quality spectral library from multiple searches using a variety of open-source software packages. These protocols have been fully automated at the ACRF International Centre for the Proteome of Human Cancer (ProCan®), which is capable of processing 10,000 tumor samples per year with six mass spectrometers operating in concert (Poulos et al., 2020; Tully et al., 2019) to generate sample-specific high-quality spectral libraries.

Critical Parameters

Parameters that have a significant impact on the overall performance and run time include the selection of candidate search engines, search parameters, and the number of fractions used for search. The selection of search engines is crucial because each search engine provides its own unique set of identification features. Studies in the past have compared different search engines (Cho et al., 2015; Matthiesen et al., 2020; Searle et al., 2018; Shao & Lam, 2017) to identify the optimal combinations. Our experience and published studies suggest that adding one more search engine leads to an increase in identifications (5%-10%), albeit with an increased accumulation of false positives (Barkovits et al., 2020; Jones, Siepen, Hubbard, & Paton, 2009; Tu et al., 2015). Researchers are thus advised to use multiple search engines with caution. The search parameters are important because they specify parent and fragment ion mass tolerance, enzymes, number of missed cleavages, and the size of the protein database. A larger number of fractions facilitates a deeper coverage of the proteome (Mertins et al., 2018; Yeung et al., 2020).

Troubleshooting

See Table 3 for problems that may arise with these protocols, along with the possible causes and solutions.

Time Considerations

The most time-consuming step is the conversion from raw files to the MGF format, which can take around 12-15 min per file. The search time depends on different parameters such as modifications selected, size of the database, and size of the acquired data file. The search speed ranks from the fastest to the slowest are Comet, X!Tandem, Mascot, and MSGF+. A typical search of a raw file against the human proteome database of about 20,000 proteins with decoys takes about 10 min. The merging step of Basic Protocol 2 usually takes about 20 min to 1 hr depending on the number of files. The final spectral library generation takes around 30 min.

Acknowledgments

This work was supported by Cancer Council NSW (IG-18-01), Cancer Institute New South Wales (NSW) (2017/TPG001, REG 171150), National Breast Cancer Foundation (IIRS-18-164), and the Medical Research

Future Fund (MRFF-PD). The authors would like to acknowledge Roger Reddel, Phillip J. Robinson, Rosemary Balleine, and Edith Hurt for comments on the manuscript.

Author Contributions

Srikanth S. Manda: Conceptualization; Formal analysis; Methodology; Software; Validation; Visualization; Writing-original draft. **Zainab Noor:** Resources; Visualization; Writing-original draft. **Peter G. Hains:** Methodology; Writing-review & editing. **Qing Zhong:** Conceptualization; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Visualization; Writing-review & editing.

Literature Cited

Barkovits, K., Pacharra, S., Pfeiffer, K., Steinbach, S., Eisenacher, M., Marcus, K., & Uszkoreit, J. (2020). Reproducibility, specificity and accuracy of relative quantification using spectral library-based data-independent acquisition. *Molecular and Cellular Proteomics*, *19*(1), 181–197. doi: 10.1074/mcp.RA119.001714

Barsnes, H., & Vaudel, M. (2018). SearchGUI: A highly adaptable common interface for proteomics search and de novo engines. *Journal of Proteome Research*, *17*(7), 2552–2555. doi: 10.1021/acs.jproteome.8b00175

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M., Cheng, L. Y., Messner, S., … Reiter, L. (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular and Cellular Proteomics*, *14*(5), 1400–1410. doi: 10.1074/mcp.M114.044305

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., … Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, *30*(10), 918–920. doi: 10.1038/nbt.2377

Cho, J. Y., Lee, H. J., Jeong, S. K., Kim, K. Y., Kwon, K. H., Yoo, J. S., … Paik, Y. K. (2015). Combination of multiple spectral libraries improves the current search methods used to identify missing proteins in the chromosome-centric human proteome project. *Journal of Proteome Research*, *14*(12), 4959–4966. doi: 10.1021/acs.jproteome.5b00578

Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., … Aebersold, R. (2017). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*, *8*(1), 291. doi: 10.1038/s41467-017-00249-5

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794–1805. doi: 10.1021/pr101065j

Craig, R., & Beavis, R. C. (2004). TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics*, *20*(9), 1466–1467. doi: 10.1093/bioinformatics/bth092

Craig, R., Cortens, J. C., Fenyo, D., & Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research*, *5*(8), 1843–1849. doi: 10.1021/pr0602085

Dasari, S., Chambers, M. C., Martinez, M. A., Carpenter, K. L., Ham, A. J., Vega-Montoto, L. J., & Tabb, D. L. (2012). Pepitome: Evaluating improved spectral library search for identification complementarity and quality assessment. *Journal of Proteome Research*, *11*(3), 1686–1695. doi: 10.1021/pr200874e

Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, *17*(1), 41–44. doi: 10.1038/s41592-019-0638-x

Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., … Aebersold, R. (2010). A guided tour of the trans-proteomic pipeline. *Proteomics*, *10*(6), 1150–1159. doi: 10.1002/pmic.200900375

Diament, B. J., & Noble, W. S. (2011). Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, *10*(9), 3871–3879. doi: 10.1021/pr101196n

Dorfer, V., Pichler, P., Stranzl, T., Stadlmann, J., Taus, T., Winkler, S., & Mechtler, K. (2014). MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *Journal of Proteome Research*, *13*(8), 3679–3684. doi: 10.1021/pr500202e

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: An open-source MS/MS sequence database search tool. *Proteomics*, *13*(1), 22–24. doi: 10.1002/pmic.201200439

Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., & MacCoss, M. J. (2006). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry*, *78*(16), 5678–5684. doi: 10.1021/ac060279n

Gillet, L. C., Navarro, P., Tate, S., Rost, H., Selevsek, N., Reiter, L., … Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular and Cellular Proteomics*, *11*(6), O111 016717. doi: 10.1074/mcp.O111.016717

Guo, T., Kouvonen, P., Koh, C. C., Gillet, L. C., Wolski, W. E., Rost, H. L., … Aebersold, R. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. *Nature Medicine*, *21*(4), 407–413. doi: 10.1038/nm.3807

Jones, A. R., Siepen, J. A., Hubbard, S. J., & Paton, N. W. (2009). Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*, *9*(5), 1220–1229. doi: 10.1002/pmic.200800473

Kim, S., & Pevzner, P. A. (2014). MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*, *5*, 5277. doi: 10.1038/ncomms6277

Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K., King, N., Stein, S. E., & Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, *7*(5), 655–667. doi: 10.1002/pmic.200600625

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., & Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: A tutorial. *Molecular Systems Biology*, *14*(8), e8126. doi: 10.15252/msb.20178126

MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., … MacCoss, M. J. (2010). Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, *26*(7), 966–968. doi: 10.1093/bioinformatics/btq054

Matthiesen, R., Prieto, G., & Beck, H. C. (2020). Comparing peptide spectra matches across search engines. *Methods in Molecular Biology*, *2051*, 133–143. doi: 10.1007/978-1-4939-9744-2_5

Mertins, P., Tang, L. C., Krug, K., Clark, D. J., Gritsenko, M. A., Chen, L., … Carr, S. A. (2018). Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. *Nature Protocols*, *13*(7), 1632–1661. doi: 10.1038/s41596-018-0006-9

Noor, Z., Mohamedali, A., & Ranganathan, S. (2020). iSwathX 2.0 for processing DDA spectral libraries for DIA data analysis. *Current Protocols in Bioinformatics*, *70*(1), e101. doi: 10.1002/cpbi.101

Noor, Z., Wu, J. X., Pascovici, D., Mohamedali, A., Molloy, M. P., Baker, M. S., & Ranganathan, S. (2019). iSwathX: An interactive web-based application for extension of DIA peptide reference libraries. *Bioinformatics*, *35*(3), 538–539. doi: 10.1093/bioinformatics/bty660

Paulo, J. A. (2013). Practical and efficient searching in proteomics: A cross engine comparison. *Webmedcentral*, *4*(10), WMCPLS0052. doi: 10.9754/journal.wplus.2013.0052

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18), 3551–3567. doi: 10.1002/(SICI)1522-2683(19991201)20:18⟨3551::AID-ELPS3551⟩3.0.CO;2-2

Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., & Coon, J. J. (2012). Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular and Cellular Proteomics*, *11*(11), 1475–1488. doi: 10.1074/mcp.O112.020131

Poulos, R. C., Hains, P. G., Shah, R., Lucas, N., Xavier, D., Manda, S. S., … Zhong, Q. (2020). Strategies to enable large-scale proteomics for reproducible research. *Nature Communications*, *11*(1), 3793. doi: 10.1038/s41467-020-17641-3

Rost, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S. M., Schubert, O. T., … Aebersold, R. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nature Biotechnology*, *32*(3), 219–223. doi: 10.1038/nbt.2841

Schubert, O. T., Gillet, L. C., Collins, B. C., Navarro, P., Rosenberger, G., Wolski, W. E., … Aebersold, R. (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. *Nature Protocols*, *10*(3), 426–441. doi: 10.1038/nprot.2015.015

Searle, B. C., Pino, L. K., Egertson, J. D., Ting, Y. S., Lawrence, R. T., MacLean, B. X., … MacCoss, M. J. (2018). Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nature Communications*, *9*(1), 5128. doi: 10.1038/s41467-018-07454-w

Shao, W., & Lam, H. (2017). Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass Spectrometry Reviews*, *36*(5), 634–648. doi: 10.1002/mas.21512

Shilov, I. V., Seymour, S. L., Patel, A. A., Loboda, A., Tang, W. H., Keating, S. P., … Schaeffer, D. A. (2007). The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular and Cellular Proteomics*, *6*(9), 1638–1655. doi: 10.1074/mcp.T600050-MCP200

Shteynberg, D., Nesvizhskii, A. I., Moritz, R. L., & Deutsch, E. W. (2013). Combining results of multiple search engines in proteomics. *Molecular and Cellular Proteomics*, *12*(9), 2383–2393. doi: 10.1074/mcp.R113.027797

Sturm, M., Bertsch, A., Gropl, C., Hildebrandt, A., Hussong, R., Lange, E., … Kohlbacher, O. (2008). OpenMS-an open-source software framework for mass spectrometry. *BMC Bioinformatics*, *9*, 163. doi: 10.1186/1471-2105-9-163

Tabb, D. L., Fernando, C. G., & Chambers, M. C. (2007). MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research*, *6*(2), 654–661. doi: 10.1021/pr0604054

Ting, Y. S., Egertson, J. D., Payne, S. H., Kim, S., MacLean, B., Kall, L., … MacCoss, M. J. (2015). Peptide-centric proteome analysis: An alternative strategy for the analysis of tandem mass spectrometry data. *Molecular and Cellular Proteomics*, *14*(9), 2301–2307. doi: 10.1074/mcp.O114.047035

Tiwary, S., Levy, R., Gutenbrunner, P., Salinas Soto, F., Palaniappan, K. K., Deming, L., … Cox, J. (2019). High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. *Nature Methods*, *16*(6), 519–525. doi: 10.1038/s41592-019-0427-6

Tsou, C. C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A. C., & Nesvizhskii, A. I. (2015). DIA-Umpire: Comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, *12*(3), 258–264, 257 p following 264. doi: 10.1038/nmeth.3255

Tu, C., Sheng, Q., Li, J., Ma, D., Shen, X., Wang, X., … Qu, J. (2015). Optimization of search engines and postprocessing approaches to maximize peptide and protein identification for high-resolution mass data. *Journal of Proteome Research*, *14*(11), 4662–4673. doi: 10.1021/acs.jproteome.5b00536

Tully, B., Balleine, R. L., Hains, P. G., Zhong, Q., Reddel, R. R., & Robinson, P. J. (2019). Addressing the challenges of high-throughput cancer tissue proteomics for clinical application: ProCan. *Proteomics*, *19*(21-22), e1900109. doi: 10.1002/pmic.201900109

UniProt, C. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. doi: 10.1093/nar/gky1049

Vaudel, M., Burkhart, J. M., Zahedi, R. P., Oveland, E., Berven, F. S., Sickmann, A., … Barsnes, H. (2015). PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nature Biotechnology*, *33*(1), 22–24. doi: 10.1038/nbt.3109

Yang, Y., Liu, X., Shen, C., Lin, Y., Yang, P., & Qiao, L. (2020). In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nature Communications*, *11*(1), 146. doi: 10.1038/s41467-019-13866-z

Yeung, D., Mizero, B., Gussakovsky, D., Klaassen, N., Lao, Y., Spicer, V., & Krokhin, O. V. (2020). Separation orthogonality in liquid chromatography-mass spectrometry for proteomic applications: Comparison of 16 different two-dimensional combinations. *Analytical Chemistry*, *92*(5), 3904–3912. doi: 10.1021/acs.analchem.9b05407

Internet Resources

Most current version of various software used in the protocols can be downloaded from the locations in Table 4.

Table 4 Internet Resources

Resource	URL
PeptideShaker	*http://compomics.github.io/projects/peptide-shaker*
SearchGUI	*http://compomics.github.io/projects/searchgui*
Proteowizard	*http://proteowizard.sourceforge.net/download.html*
Skyline	*https://skyline.ms/project/home/software/Skyline/begin.view*
SwathXtend	*https://www.bioconductor.org/packages/release/bioc/html/SwathXtend.html*
iSwathX	*https://biolinfo.shinyapps.io/iSwathX/*
Files used in protocols	*https://cloudstor.aarnet.edu.au/plus/s/YLTUEZse2kMpyzl*