Tutorial 5



Untargeted analysis of DIA datasets

In this tutorial, we will perform an untargeted analysis of a data-independent acquisition (DIA) dataset using the FragPipe computational tool collection. We will analyse a subset of samples from the published clear cell renal cell carcinoma (ccRCC) studies. that were originally described in the following publication: D. J. Clark et al. "Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma", Cell 2019 179(4):964-983. doi: 10.1016/j.cell.2019.10.007 (https://pubmed.ncbi.nlm.nih.gov/31675502/). Briefly, in the original studies, researchers from the CPTAC (Clinical Proteomic Tumor Analysis Consortium) profiled tumor (T) samples, together with normal adjacent tissue (NAT) samples from each cancer patient, to understand the tumorigenesis of ccRCC. 110 tumor and 83 NAT samples were collected from patients and their proteomes were profiled via mass spectrometry. These samples were originally profiled using: i) tandem mass tag (TMT), and ii) data-independent acquisition (DIA). The DIA set was generated on an Orbitrap Lumos mass spectrometer with a variable window acquisition scheme.

Here, we will use just 10 DIA runs from 5 ccRCC patients, one tumor and one paired NAT sample for each patient. To make the data processing faster, we will use only data in two isolation windows (613 to 650 Th mass range) from each original mzML file.

We will use FragPipe for these analyses, which is a Java Graphical User Interface (GUI) for a suite of computational tools enabling comprehensive analysis of mass spectrometry-based proteomics data. It is powered by MSFragger, an ultrafast proteomic search engine suitable for both conventional and open (wide precursor mass tolerance) peptide identification. FragPipe includes the Philosopher toolkit for downstream statistical post-processing of MSFragger search results (PeptideProphet, iProphet, ProteinProphet), FDR filtering, label-based quantification, and multi-experiment summary report generation. The software is well documented (https://fragpipe.nesvilab.org/) and the original publication is Yu, F et al.. (2023). Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform.*Nature Communications*, *14*(1), 4154.

In this tutorial, we will combine the MSFragger module with DIANN for direct analysis of data independent acquisition (DIA) data. We will first process the data with MSFragger to identify multiple peptides in chimeric spectra, then statistically validate the identification results with Percolator, and finally perform peptide quantification with DIA-NN. Once we get the identification and quantification results from FragPipe, we will load them in FragPipe-PDV to visualize the identifications, and we will perform some downstream analysis using FragPipe-Analyst. Finally, we will learn how to load the raw data in Skyline to see the extracted ion chromatograms for each of the identified peptides.

Parametrization of FragPipe graphical user interfase

In this first part of the tutorial we will set up the graphical user interface of FragPipe and launch a libraryfree search combining MSFragger-DIA and DIA-NN. The end result will be the generation of a collection matrices with the quantification values at the precursor and protein levels, as well as a summary pdf file of the experiment.

Note: This tutorial is based on FragPipe 20.1-build15.

- Go to Tutorial-5\Fragpipe\tools\Fragpipe-20.1-build15\fragpipe\bin
- Click in the fragpipe.exe icon to open the graphical user interface.



Parametrization of the Config section

In this section we need to make sure that all the different tools that are required by FragPipe are installed in the system and provide FragPipe with the path to the corresponding executables.

- Select the Config tab from the graphical user interfase of FragPipe.
- Go to the MSFragger section below and click "Browse". Navigate to Tutorial-5\Fragpipe\tools\ MSFragger-3.8\MSFragger-3.8 and select the MSFragger-3.8.jar executable.

🎢 Select MSFra	ggerjar		×
Look in:	MSFragge	r-3.8 🗸 🦻 🗁 🛄 🗸	
Recent Items	ext MSFragge	:r-3.8	
Desktop			
Documents			
This PC			
I	File name:	MSFragger-3.8.jar	Select
Network	Files of type:	JAR files 🗸	Cancel

• Go to the lonQuant section below and click "Browse". Navigate to Tutorial-5\Fragpipe\tools\ and select the lonQuant-1.9.8.jar executable.

🃸 Select IonQua	antjar				×
Look in:	l tools			 Image: state of the state of th	-
Recent Items	Name FragPipe-20.1 MSFragger-3.8 IonQuant-1.9.8	Size 15.9 MB	Type File folder File folder JAR File	Modified 10/11/2023 3:08 10/11/2023 3:09 10/3/2023 7:14 PM	
Desktop					
Documents					
This PC					
) Network	File name: IonQu Files of type: JAR f	uant-1.9.8.jar files		~	Select Cancel



• Go to the Philosopher section below and click "Browse". Navigate to Tutorial-5\Fragpipe\tools\ and select the philosopher-v5.1.0-RC7.exe executable.

📸 Select Philoso	opher binary				\times
Look in:	l tools		~	🤌 📂 🛄 -	
Recent Items	FragPipe- MSFragge lonQuant jdk-11.0.1	20.1-build15 :r-3.8 -1.9.8 9_windows-x64_bin			
Desktop	python-3.	9.13-amd64			
Documents					
This PC					
I	File name:	philosopher-v5.1.0-RC7.	exe		Select
Network	Files of type:	All Files		\sim	Cancel

• Go to the DIA-NN section below and click "Browse". Navigate to Tutorial-5\Fragpipe\tools\ FragPipe-20.1-build15\fragpipe\tools\diann\1.8.2_beta_8\win and select the DIA-NN.exe executable.



• Go to the Python section below and click "Browse". Navigate to Tutorial-5\Fragpipe\tools\ and select the python-3.9.13-amd64 installer.



🐮 Select Pythor	n 3.9+ binary				×
Look in:	tools		~	🏚 📂 🛄 •	
Recent Items	FragPipe- MSFragge IonQuant jdk-11.0.1	20.1-build15 er-3.8 -1.9.8 9_windows-x64_bin ner-v5.1.0-RC7 0.12_err_464			
Desktop	Se python-3	.9.13-amd64			
Documents					
This PC					
I	File name:	python-3.9.13-amd64.exe			Select
Network	Files of type:	All Files		~	Cancel

After the installer is finished installing Python, the path should be automatically updated to "C:\Users\[your user]\AppData\Local\Programs\Python\Python39\python.exe". Otherwise, customize the path to python to your local installation.

• Go to the "Spectral Library Generation" section below and click "Install/Upgrade EasyPQP". Wait until the installation of this python module is finished.

Your Config tab should look like this:

TragPipe (v20.1-build15)			- 0	×
👂 Config 👬 Workflow 🙏 Umpire 🔶 Database 🕴 MSFragger 👯 Validation 🥖 PTMs 🍸 Glyco 🖗 Quant (MS1) 🚇 Quant (Isobaric) 🏦 Spec Lib 🛱 Quant (DIA) 🌘	Run 🔏	ownstream		
System OS: Windows 10, Architecture: AMD64 About Open cache location Clear cache and close Java Inic: 11.0.16, 1, Java HolSpot(TM) 64-Bit Server VM, Oracle Corporation NET Core Infr: NT Core Infr: Architecture: AMD64 FragPipe: v20.1-build15		1	R / G P I P E	
More info and docs: DIA-Umpire, Crystal-C, MSBooster, Percolator, PTMProphet, PTM-Shepherd, Q-Pair, TMT-Integrator, EasyPOP, ErapPipe-PDV, SAINT				
MSFragger				
Z: \Tutorial-5\Fragpipe\tools\MSFragger-3.8\MSFragger-3.8\MSFragger-3.8.jar	Browse	Down	load / Update	
MSFragger version: 3.8 More info and docs: MSFragger				
IonQuant				- 1
2: \Tutorial-5\Fragpipe\tools\JonQuant-1.9.8.jar	Browse	Down	load / Update	
IonQuant version: 1.9.8 More info and docs: IonQuant				
Philosopher				
Z:\Tutorial-5\Fragpipe\tools\philosopher-v5.1.0-RC7.exe	Browse	Down	load / Update	
Philosopher version: N/A More info and docs: <u>Philosopher</u>				
DIA-NN				
Z: \Tutorial-\$\Fragpipe\tools\Fragpipe\tools\FragPipe-20.1-build15\fragpipe\tools\diann\1.8.2_beta_8\win\DiaNN.exe		Browse	Download	
Path to DIA-NN executable file. The file name is DiaNN.exe. Do not use DIA-NN.exe or the installer file. If not customized, use the built-in version 1.8.2 Beta 8				
DIA+IN version: 1.8.2 Beta 8 More info and docs: DIA+IN				
_ Python				
C: Users proteomics AppData Local Programs (Python 39) python.exe		Browse	Download	
Python version: Python 3.9.13				
r Database Spitting				
Database Splitting: Available. Used for searching very large databases by splitting into smaller chunks.				
Spectral Library Generation				
EasyPQP: Available. Version: 0.1.40				
Instal/Upgrade EasyPQP				
Tabs at the top represent processing steps that will be performed sequentially if enabled.				~



Parametrization of the Workflow section

FragPipe supports multiple proteomics workflows which can be customized, saved and shared with other users.

In the Workflow tab:

- Choose the workflow "DIA_SpecLib_Quant" workflows, which corresponds to the DIA spectral library generation and quantification using DIA-NN.
- Press "Load workflow" to load the parameters of the selected workflow.
- In Global settings, set the amount of RAM memory to zero. A RAM setting of 0 will allow FragPipe to automatically detect available memory and allocate a safe amount.
- In the "Parallelism" you can select the number of logical cores to use. Set this to the number of cores that your computer has minus one.

In "Input LC-MS Files" section we will load and annotated all the mzML files that contain the raw data acquired in the aforementioned experiment.

- Check "Regular MS". Note that the option 'IM-MS' is meant only for Bruker timsTOF PASEF data whereas 'Regular MS' is meant for all other data types (including FAIMS).
- Click "Add files" and navigate to Tutorial-5\Fragpipe\mzml and select the 10 mzML files. Click "Select".





Table 1: File names and conditions

File Name	BioReplicate	Condition
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_NAT.mzML	1	NAT
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_T.mzML	1	Т
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00606_NAT.mzML	2	NAT
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00606_T.mzML	2	т
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-01882_NAT.mzML	3	NAT
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-01882_T.mzML	3	Т
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00577_NAT.mzML	4	NAT
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00577_T.mzML	4	Т
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00733_NAT.mzML	5	NAT
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00733_T.mzML	5	Т

Now we need to annotated the Experiment and the Bioreplicate associated to each raw file according to the information provided in Table 1.

- For each file (row) select "Set experiment" and type the condition "NAT" or "T" according to Table 1. Note that one can select multiple rows with the "Control" key and annotate them simultaneously.
- For each file (row) select "Set Bioreplicate" and type the condition 1 to 5 according to Table 1. Note that the two conditions are always paired as they come from the same individual, and therefore, for each pair, we need to set the same bioreplicate number.

Save as manife	est Load manifest	Assign files to Experiments/	Groups (select ro	ws to activate action	buttons):				
Consecutive	By parent directory	By file name Set	experiment	Set replicate	Clear groups		Set DDA	Set DIA	Set GPF-DIA
						s	et DIA-Quant	Set DIA-Lib	Set WWA
Path (can drag 8	& drop from Explorer)				Experiment (can be empty, alphanu	Bioreplicate (can be empty a	nd intege D	ata type (DDA, DIA,	GPF-DIA, DIA-Q
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	3L-00418_NAT.m	zML	NAT		1 DI.	A	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3L-00418_T.mzM	L	т		1 DI.	A	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	3L-00606_NAT.m	zML	NAT		2 DI	A	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3L-00606_T.mzM	L	т	2 DIA			
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	3L-01882_NAT.m	zML	NAT		3 DI.	A	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3L-01882_T.mzM	L	т		3 DI.	Ą	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3N-00577_NAT.m	nzML	NAT		4 DI.	A	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3N-00577_T.mz№	1L	т		4 DI.	A	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3N-00733_NAT.m	nzML	NAT		5 DI.	Ą	
G:\DIA_tutorial\tv	wo_windows\CPTAC_CCRCC	_W_JHU_20190112_LUMOS_C	C3N-00733_T.mzM	1L	т		5 DI.	A	



Parametrization of the Database section

We will skip the Umpire tab as it is not meant to be executed in the selected workflow, and move directly to the Database tab.

• Click "Download" to retrieve a fresh UniProt human database including reviewed sequences only, plus contaminants and decoys.

Downlo	pad options		
2	Select organism / Input proteome ID		
	Homo sapiens (Human) - Uniprot ID: UP000005640		
	O Mus musculus (Mouse) - Uniprot ID: UP000000589		
	○ S. cerevisiae (Yeast) - Uniprot ID: UP000002311		
	O SARS-CoV-2 (COVID-19) - Uniprot ID: UP000464024		
	Other:		
	Options		
	Reviewed sequences only		
	Add decoys		
	Add common contaminants		
	Add isoforms		
	Add iRT sequences		
	Spike-in sequences (do not include decoys)		
	FASTA file path optional Browse		
	OK Cancel		

• Save the fasta in the Tutorial-5 folder.

Note: Alternatively, one can also use the fasta file database provided in the "Tutorial-5\FragPipe" folder with the name 2023-10-03-decoys-reviewed-contam-UP000005640.fas

∬ FragPipe (√20.1-build15)		
😴 Config 🚏 Workflow 📐 Umpire 📌 Database 🕴 MSFragger 👫 Validation 🥒 PTMs 🏋 Glyco 🕸 Quant (MS1) 🔀 Quant (Isobaric) 🏦 Spec Lib	🟠 Quant (DIA) 🔹 R	un 💦 Downstream
FASTA sequence database		
FASTA file path C:\Tutorial-5\Fragpipe\2023-10-03-decoys-reviewed-contam-UP000005640.fasta	'owse Download	Add decoys
Decoy protein prefix rev_ Try auto-detect tag		
"Browse" to select a FASTA file from a previous FragPipe analysis, or "Download" to retrieve a new one from UniProt. Use "Add decoys" to append decoy sequences and/or add common <u>contaminant sequences</u> . IMPORTANT: Sequence headers must follow certain <u>format rules</u> . To download a database containing two or more organisms during 'Download', list all UniProt proteome IDs separated by commas, e.g., UP000005640,UP000464024 to	get a combined human +	COVID-19 database.



Parametrization of the MSFragger section

In the MSFragger tab you can check the search parameters that will be used to interpret the acquired spectra our analysis. The parameters have already been filled with the default values associated to the workflow selected. Let's review them.

- In **Peak Matching** the precursor and fragment mass tolerances are specified. In this case, it is set to 20 ppm, which is the standard in a so-called closed data search using data acquired in a high-resolution mass spectrometry analyzer. Options on whether there is the need for automatic mass calibration, parameter optimization and isotope peak selection correction are also present.
- In **Protein Digestion** we define that an enzymatic digestion was used to prepare the sample, and that trypsin was the enzyme used. The range of peptide length and mass are also specified. All these parameters are important when interpreting the spectra *in silico* because they define the set of potential peptides that can be present in the sample.

Note: Calibration and Optimization is set by default to "Mass Calibration, Parameter Optimization". This option will effectively perform multiple simplified MSFragger searches with different parameters to find the optimal settings for your data. In practice, it results in 5-10% improvement in the number of identified PSMs at the expense of increased analysis time. To save time, you can consider changing this option to "Mass Calibration" or even "None", especially if you already know your data (e.g. from previous searches of the same or similar files) and can adjust the corresponding MSFragger parameters (fragment tolerance, number of peaks used, intensity transformation) manually, if needed.

Run MSFragger	
Save Config Load : Custom MSFragger parameter file from disk $ \lor $	MSSER GGER
⊂ Common Options (Advanced Options are at the end of the page)	
Peak Matching	
Precursor mass tolerance PPM v -20 🜩 Fragment mass tolerance PPM v 20 🜩	
Calibration and Optimization Mass calibration, parameter optimization \lor Isotope error 0/1/2	
Protein Digestion	
Cleavage ENZYMATIC V Clip N-term M	
Enzyme name 1 stricttrypsin Load rules stricttrypsin V Cuts 1 KR No cuts 1 Missed deavages 1 1 📩 Sense 1	C ~
Enzyme name 2 null Load rules null V Cuts 2 No cuts 2 Missed deavages 2 1 😴 Sense 2	C ~
Peptide length 7 - 50 - Peptide mass range 500 - 5,000 - Split database 1 -	

 In Modifications both variable and fixed modifications that can be found in the analysed peptides are specified, as well as the maximum number of allowed modifications per peptide and the maximum number of occurrences per single modification.



Max variable moo	is on a peptide			-	
Enabled	Site (editable)	Mass Delta	(edita Max occurre	nces	
\checkmark	M	15.9949	3	A	
	[^	42.0106	1		
	STY	79.96633	3		
	nQnC	-17.0265	1		
	nE	-18.0106	1		
	site_06	0.0	1		
	site_07	0.0	1		
			-		
ixed modification	s site 08	0.0	1	~	
ixed modification Enabled	site 08	0.0	1 Mass Delta (edita	ble)	
ixed modification Enabled	site 08	0.0	Mass Delta (edita	ble)	
ixed modification Enabled	site 08 Site C-Term N-Term	0.0	Mass Delta (edita	ble)]
ixed modification Enabled	site 08 Site C-Term N-Term C-Term	0.0 Peptide Peptide Protein	Mass Delta (edita 0.0 0.0 0.0	ble)	
ixed modification Enabled	site 08	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0	Mass Delta (edita 0.0 0.0 0.0 0.0	ble)	
ixed modification Enabled	site 08 Site C-Term C-Term C-Term G (glyc	0.0 Peptide Peptide Protein Protein ine)	Mass Delta (edita 0.0 0.0 0.0 0.0 0.0 0.0	ble)	
ixed modification Enabled	site 08 Site C-Term C-Term C-Term C-Term C-Term G G (g)vc A (alar	0.0 Peptide Peptide Protein Protein ine) ine)	Mass Delta (edita 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0	ble)	
ixed modification Enabled	s Site 08	0.0 Peptide Peptide Protein Protein ine) ine) Pe)	Mass Delta (edita 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.	ble)	
ixed modification Enabled	s Site 08	0.0 Peptide Peptide Protein Protein ine) ine) ne) ne) ne)	Mass Delta (edita 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.	ble)	
ixed modification Enabled	s Site 08 C-Term C-Term C-Term C-Term G (glyc G (glyc A (alar S (serir P (prol) V (valir	0.0 Peptide Peptide Protein Protein ine) ine) ine) ne) ne) ne) ne)	Mass Delta (edita 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.	ble)	

• Finally, there is the section of **Advanced Options**, which includes parameters for *Spectral Processing* that determine how many spectral peaks are taken into consideration. It also includes the *Advanced Output Options* and *Advances Peak Matching Options* that define the number of top N peptides to use for quantification, the output format, the fragment ion series, fragment charge range and minimum number of matched fragments to be considered during the search.

Spectral Processing		
Activation Type Filter $$ all $ \lor$	Precursor mass mode selected $$	Check spectral files Require precursor
Min peaks 15 🜩	Use top N peaks 300 🛓	Min ratio 0.00 🗭 🗌 Reuse DIA fragment peaks
Clear m/z range 0 🔷 - 0 🗢	Intensity transform None $$	
Remove precursor peak Only peak with precurso	or charge \checkmark removal m/z range -1.5 🜩 - 1.5	•
Advanced Output Options		
Report top N for DDA 1 🛓 🗸 R	eport alternative proteins Output format PEPXML_P	PIN ~
Report top N for DIA 5 📥 🛛 V	/rite calibrated mzML Group variable None	~
Report top N for GPF-DIA 3	/rite uncalibrated MGF Output max expect 50 🚖	
Report top N for WWA 5		
Advanced Peak Matching Options		
Min frags modeling 2	Min matched frags 5 🜩	Max fragment charge 2 😴
Deisotope Yes	✓ Fragment ion series b, y	Define custom ion series
Deneutralloss Yes 🗸	Precursor true tolerance $~$ PPM $~\sim~$	20 🔹 🗋 Override charge with precursor charge 1 🔹 - 4

You can choose to save a customized parameter file to load for future use, or save the entire workflow (from either the 'Workflow' or the 'Run' tab).



Parametrization of the Validation section

The Validation section will also be executed as part of the selected workflow. The search results obtained from MSFragger will be further analyzed by MSBooster, Percolator and ProteinProphet to get confident peptide identifications.

In this process, <u>MSBooster</u> will first use deep learning to predict additional features of the identified peptides including fragmentation spectra, retention time, and detectability (and ion mobility).

┌Rescoring Using Deep Learning Predicti	on
Run MSBooster	Rescoring using deep learning prediction. Require Run Percolator in PSM validation panel.
Predict RT Predict spectra	Use correlated features

These features will be used to modify the initial identification scoring, and then <u>Percolator</u> will use them to improve its discrimination model to increase the number of confident identifications in the DIA dataset.

□ PSM Validation	
Run PSM Validation	
○ Run PeptideProphet Defaults for: Closed Search ∨ Load	Single combined pepxml file per experiment / group
Cmd line opts:decoyprobsppmaccmassnonparamexpectscore	
Run Percolator	
Cmd line opts:only-psmsno-terminatepost-processing-tdcsubset-max-train 500000	

Finally, based on the identified peptides we will run the Protein Inference together with <u>ProteinProphet</u> to generate a confident list of protein groups identified in the sample at a maximum of 1% false discovery rate.

✓] Run ProteinProphet
Cmd line opts:maxppmdiff 2000000minprob 0.5
Filterpickedprot 0.01
Do not use ProteinProphet file
Generate MSstats files Remove contaminants Print decoys Generate peptide-level summary Generate protein-level summary



Parametrization of the Spec Lib section

Next, we will jump directly to the Spec Lib tab as the other ones (PTMs, Glyco, Quant (MS1), and Quant (isobaric)) are not relevant for the selected workflow and will not be executed. In the Spec Lib section we will generate a spectral library from the search results, containing b and y fragment ions, and we will allow for an automatic selection of the runs that will be used as reference for the retention time.

keep intermediate files		
RT calibration	Automatic select	tion of a run as reference RT $\!$
IM calibration	Automatic select	tion of a run as reference IM $\!$
RT Lowess fraction	0 ≑	Fragment types: 🗋 a 🗹 b 📄 c 🗌 x 🗹 y 🛄 z
UniMod annotation tol (Da)	0.02 🚖	neutral loss
Fragment annotation tol (ppm)	15 🜩	

Parametrization of the Quant (DIA) section

In the Quant (DIA) section we will set the quantification to be performed by DIA-NN with a maximum false discovery rate of 1%. In this section, we will also verify that the "Generate MSstats input" is checked.

Г	DIA Ouantification	
	Ousstify with DIA NN	
ľ	The stand-alone DIA-NN program (with full func	tionality) can be downloaded from the DIA-NN GitHub repository. Reference
ŀ		
	FDR	0.01 🜩
	Apply run-specific protein FDR	
	Quantification strategy	Robust LC (high precision) $$
	Unrelated runs	
	Replace library spectra with predicted	
	Generate MSstats input	
	Spectral library (optional)	
	Cmd line opts:	
L		

Parametrization of the Run section

In this final section, we will indicate the output directory and run the analysis.



- Create a new folder in "Tutorial-5\FragPipe" called output and set is as the output directory of the results.
- Click "RUN". The analysis will now be launched and it will take about 10-20 minutes. Once finished, do not close the FragPipe window. We will need it later to visualize the results.

```
FragPipe (v20.1-build15)

      Config ** Workflow & Umpire * Database * MSFragger ** Validation / PTMs * Glyco @ Quant (MS1) @ Quant (Isobaric) ** Spec Lib ** Quant (DIA) Run ** Downstream

      About

      Output dir:
      Z:\Tutorial-S\FragPipe\output

      Browse
      Open in File Manager

      RUN
      Stop
      Dry Run

      Save Technical SDRF
      Export Log
      Report Errors

      Open FragPipe-PDV viewer
      Mill
      Delete calibrated mzML
      Delete temp files
```

Exploration of the FragPipe main results tables

In this first part of the tutorial we will go through the main results tables generated by FragPipe and some of its intermediate files.

Inspection of the FragPipe main output

- Go to Tutorial-5 and locate the "diann-output" folder that has been generated by FragPipe.
- Inside the "diann-output" folder locate the report.pg_matrix.tsv file and open it in Excel to inspect the protein-level output from the DIA-NN quantification module. You will see columns with the information such as protein group identifiers, gene names, and intensities from DIA runs calculated with the MaxLFQ algorithm embedded in DIA-NN.

1	А	В	С	D	E	F	G	н	- I	J	К	L	м	N	0
1	Protein.Group	Protein.Ids	Protein.Names	Genes	First.Protein.Description	C:\esabid	C:\esabidc								
2	A0A075B6H7	A0A075B6H7		IGKV3-7		1.15E+07		1.67E+07	5.23E+06	1.56E+06	3.31E+06	6.02E+06	582433	1.25E+06	9.46E+06
3	A0A075B6J2	A0A075B6J2		IGLV2-33		992891		852624	116486	129461		486764	108517	429645	127771
4	A0A075B6Q5	A0A075B6Q5		IGHV3-64		3.28E+06	3.00E+06	2.99E+06	1.58E+06	1.26E+06		3.64E+06	1.42E+06		1.01E+06
5	A0A075B6R9	A0A075B6R9		IGKV2D-2	4	7.77E+06	3.33E+06	8.29E+06	3.80E+06	2.57E+06	904343	6.36E+06	1.73E+06	3.85E+06	1.09E+06
6	A0A075B6S2	A0A075B6S2		IGKV2D-2	9	8.75E+06	310027	9.98E+06		375331		182868		379454	999943
7	A0A0A0MS15	A0A0A0MS15		IGHV3-49		1.79E+06	1.16E+06	3.40E+06	712559	1.43E+06	408955	1.35E+06		789723	1.37E+06
8	A0A0B4J1U7	A0A0B4J1U7		IGHV6-1		4.49E+06	217386	4.98E+06	220183	164504		351491	73166.3	244250	666500
9	A0A0B4J1∨0	A0A0B4J1∨0		IGHV3-15		2.93E+06	416617	2.74E+06	364616	289347	145960	697663	116222	285306	463471
10	A0A0B4J1∨6	A0A0B4J1∨6		IGHV3-73		1.73E+06	353681	1.90E+06	480621	450664	124681	644654		452981	950859
11	A0A0B4J1Y8	A0A0B4J1Y8		IGLV9-49		917628		556435	281035	485385					
12	A0A0B4J1Z2	A0A0B4J1Z2		IGKV1D-4	3	319835		245127							
13	A0A0C4DH24	A0A0C4DH24		IGKV6-21				314816						58471.5	392222
14	A0A0C4DH31	A0A0C4DH31		IGHV1-18		969456		1.43E+06							
15	A0A0C4DH36	A0A0C4DH36		IGHV3-38		1.23E+06		1.98E+06		516025	192168		181477	533340	583892
16	A0A1B0GUS4	A0A1B0GUS4		UBE2L5		223969	546748	350964	418114	190428	522785	818209	421558	320091	992621

 In the same folder, now open the msstats.csv file. This file contains the information required for MSstats at the fragment level. Note that the information related to "Experiment" and "Bioreplicate" annotated in FragPipe is provided now to MSstats as the "Condition" column and the "BioReplicate" column, respectively.

Note: Msstats is an R package for statistical inference of proteomics data and can be accessed through the web at <u>www.msstatsshiny.com</u>. Kohler D, et al. MSstats Version 4.0: Statistical Analyses of Quantitative Mass Spectrometry-Based Proteomic Experiments with Chromatography-Based Quantification at Scale. J Proteome Res. 2023 May 5;22(5):1466-1482. doi: 10.1021/acs.jproteome.2c00834.



	А	В	с	D	E	F	G	н	1	J
1	ProteinName	PeptideSequence	PrecursorCharge	Fragmention	ProductCharge	IsotopeLabelType	Condition	BioReplicate	Run	Intensity
2	Q86U42	(UniMod:1)AAAAAA	2	b4	1	L	NAT	4	CPTAC_	1821522
3	Q86U42	(UniMod:1)AAAAAA	2	y10	1	L	NAT	4	CPTAC	856179
4	Q86U42	(UniMod:1)AAAAAA	2	у9	1	L	NAT	4	CPTAC_	741283.7
5	Q86U42	(UniMod:1)AAAAAA	2	у8	1	L	NAT	4	CPTAC_	876211.6
6	Q86U42	(UniMod:1)AAAAAA	2	y11	1	L	NAT	4	CPTAC_	816950.1
7	Q86U42	(UniMod:1)AAAAAA	2	у6	1	L	NAT	4	CPTAC_	654794.6
8	Q86U42	(UniMod:1)AAAAAA	2	у7	1	L	NAT	4	CPTAC_	420455.8
9	Q86U42	(UniMod:1)AAAAAA	2	b5	1	L	NAT	4	CPTAC_	1019006
10	Q86U42	(UniMod:1)AAAAAA	2	y12	1	L	NAT	4	CPTAC_	549599
11	Q86U42	(UniMod:1)AAAAAA	2	y13	1	L	NAT	4	CPTAC_	175193.2
12	Q86U42	(UniMod:1)AAAAAA	2	b6	1	L	NAT	4	CPTAC_	802748.4
13	Q86U42	(UniMod:1)AAAAAA	2	b4	1	L	NAT	1	CPTAC_	1975461
14	Q86U42	(UniMod:1)AAAAAA	2	y10	1	L	NAT	1	CPTAC_	1463101
15	Q86U42	(UniMod:1)AAAAAA	2	у9	1	L	NAT	1	CPTAC_	1288745
16	Q86U42	(UniMod:1)AAAAAA	2	у8	1	L	NAT	1	CPTAC	1028672
17	Q86U42	(UniMod:1)AAAAAA	2	y11	1	L	NAT	1	CPTAC	1430375

Inspection of intermediate FragPipe output files

If you are curious, you can explore FragPipe output files to get a better understanding of various FragPipe modules.

- Open the psm.tsv file with Excel and inspect the information that you have for each peptide-spectrum match. You will see the "SpectralSim" column, which indicates how well each PSM's experimental fragment intensities match predicted intensities from the spectral prediction model; this is the 'spectral entropy' score, a value between 0 and 1, with 1 being a perfect match. You will also see the "RTScore", which shows how much the experimental retention time of each PSM deviates from what is expected based on retention time predictions; the lower the value, the better.
- Open now the .png files inside *MSBooster_RTplots*. Each of these files corresponds to a different mzML file and shows the calibration curve fit between the predicted iRT scale and the experimental RT from this experiment's chromatography setup.
- Open the log file with a text editor or your Notepad. In this file you will find all the commands that have been executed by the FragPipe workflow. Note the mass correction values printed at the mass calibration step of MSFragger. Inspect also the Percolator weights. The greater the magnitude of the weights, the more influence that variable has in Percolator rescoring.
- Open the library.tsv file with Excel. This is the library files built using EasyPQP from PSM.tsv and mzML files, and contains peptide ions passing 1% protein-level, peptide-level, and PSM-level FDR. For precursors identified from multiple PSMs (in the same or different runs), fragment ion intensities and retention time (after alignment to the reference run) of the best scoring PSM are used. This file is used as input to DIA-NN for extracting quantification.

Visualization of the FragPipe main results

Visualization of identification results in FragPipe-PDV

In this section we will visualize the identification results from FragPipe at different levels, including experiments, proteins, peptides and PSM information.

• Go to the Run tab in the graphical user interfase of FragPipe and open FragPipe-PDV by clicking on '**FragPipe-PDV viewer**' to open the results.





A FragPipe-PDV viewer will open with five main panels including the information about the "Experiments", "Protein coverage", "Spectrum viewer with annotations", "Table of identified proteins", and the "Table with Peptide-Spectrum Matches (PSMs)".

There are several functions embedded in FragPipe-PDV that we will explore. For example, one can look for certain peptide sequences or protein of interest.

• Search the protein "CTNA1", using the searching function located on the top right corner. How many PSMs are associated to this protein? How many different unique peptide sequences have been identified for this protein? What are their PeptideProphetProbabilities

				-		\times
Protein (String) ~					C	2
Peptide (String) Spectrum (String)						<u> </u>
Protein (String)	Selected	~	Ź↓	₹₹	Ч	ч

In FragPipe-PDV you can also see the annotated spectra in which peptides were identified. FragPipe-PDV has several options to configure the settings for peptide spectra visualization.

• Go to the "Tools" menu below the spectrum, click and select "Show Predicted" to show the predicted spectrum in a mirror spectra format.





Why do you think spectra are populated with so many different unmatched peaks? Are they good identifications?

• To clean the spectra, click "Show Matched Peaks" in the "Settings" menu to remove background peaks



Do the identifications look better now? Do you think that they are more credible?

Now we will see how different peptides can be identified in the same single MS2 spectrum. For this example, we will use one peptide SMEDSVDVSAPK from sp|Q8IVF2|AHNK2_HUMAN Protein. This is one of the ccRCC cancer biomarkers (overexpressed in tumor samples) that we will also use as an example later in this tutorial.

- Search for protein Q8IVF2 and find all peptides identified for this protein.
- Click on the peptide SMEDSVDVSAPK listed above to view its spectrum.
- Go to the Tutorial-5 folder and open the PSM.tsv file with Excel and look for the SMEDSVDVSAPK peptide. Note that the peptide has been identified twice, in two different spectra from two different raw files.
- Check the spectrum number for each of the identifications looking at the value of the first column.

	А		В	с	
1	Spectrum	•	Spectri 💌	Peptide	Τ,
4528	CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_T.01248.01248.0		C:\esabid	SMEDSVDVSAPK	
27639	CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00577_T.01224.01224.0		C:\esabid	SMEDSVDVSAPK	

 Now filter the PSM.tsv file by the spectrum column containing the spectrum "CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_T.01248.01248.0". Notice that there is another peptide, TDYM[+16]VGSYGPR, that was identified in that same MS2 scan.

	A		В	С	D
1	Spectrum 🤤	T 9	Spectri 💌	Peptide 📃 💌	Modified Peptide 🛛 💌
4528	CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_T.01248.01248.0]0	C:\esabid	SMEDSVDVSAPK	
4529	CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_T.01248.01248.0	0	C:\esabidi	TDYMVGSYGPR	TDYM[147]VGSYGPR
8091					

You can use the PDV viewer to visualize both peptides at the same time.

• Select "Check Peptide" function in the "Tools" menu to check different peptide matches on the current spectra.





• Click "Add" and type the sequence of the peptide that you want to add into the visualization. Enter the peptide sequence without modifications.

PDV - Che	ck Peptide	Input												×
Peptide Sequ	ence							tdymv	gsygpr					
	NH-	т	D	Y	М	v	G	s	Y	G	Ρ	R	-COOH	
lons	Other	Loss	I Ch	arge	De Novo	I S	ettings	Export	t j To	ools j	Add	_		Add < Wh

• Click amino acid to add a modification on it.



Note: Remember than in "Settings" you can select to see either all peaks or only the matched peaks.





Downstream analysis of FragPipe main results

Downstream analysis using FragPipe-Analyst

In this section we will do a downstream analysis and visualization of the quantitative results of the obtained results with FragPipe-Analyst and we will perform a principal components analysis (PCA) and a statistical assessment of protein abundance changes.

 Go to the "Run" tab of the graphical user interface of FragPipe and click in the FragPipe-Analyst button. Alternatively, you can also access FragPipe-Analyst directly in your browser at <u>http://fragpipe-analyst.nesvilab.org/</u>

Open FragPipe-Analyst	
-----------------------	--

FragPipe-Analyst	
🕈 Home	Getting Started
🗸 Analysis 🔨 🔇	
?Documentation	Quick Start
	 Choose the type of experiment you performed. Currently, DDA-based LFQ (MS1-based or spectral count), TMT, and DIA are supported. For DDA LFQ: Upload combined_protein.tsv generated by lonQuant in FragPipe Upload experiment_annotation.tsv file. Edit the template file generated by FragPipe; Check here for details. Select quantification method (MS1-based Intensity or MaxLFQ Intensity, or spectral counts). For TMT: Upload gene-level report [abundance/ratio]_gene_[normalization].tsv generated by TMT-Integrator (TMT-I) in FragPipe (we recommend abundance_gene_MD.tsv file). If peptide option is enabled in the server, TMT-I peptide report could also be used alternatively. Upload experiment_annotation.tsv file. Edit the template file generated by FragPipe; Check here for details. For DIA: Upload protein group (PG) matrix (diann-output.pg_matrix.tsv) generated by DIA-NN in FragPipe Upload protein group (PG) matrix (file. Check here for details. Optional: Adjust the p-value cut-off, the log2 fold change cut-off, missing value imputation, FDR correction method in the Advanced Options . Note that the missing value imputation method is att by default to Tops up to Do acod to the server and DA acod to the server of DA acod to TMT.
	Press 'Run' Hint: Check the Documentation tab for a detailed explanation of inputs advanced options and
	Proteomics & Integrative Bioinformatics Lab at the University of Michigan (PJ. Alexey Nesvizhskii) and the Monash Proteomics & Metabolomics Facility, Monash University (PJ. Ralf Schittenhelm).

- Choose the Analysis option from the left-hand side menu and in the "Data type" dropdown menu, choose "DIA".
- Follow the instruction on the left-hand side menu to upload report.pg_matrix.tsv and experiment annotation.tsv.



FragPi	pe-Analyst
📌 Home	
📕 Analysis	~
Data type:	
DIA	-
Upload prote matrix *.tsv	in group (PG)
Upload prote matrix *.tsv Browse	in group (PG) report.pg_mati
Upload prote matrix *.tsv Browse Uplo	in group (PG) report.pg_matu ad complete
Upload prote matrix *.tsv Browse Uplo	in group (PG) report.pg_matu ad complete le annotation
Upload prote matrix *.tsv Browse Uplo Upload samp Browse	in group (PG) report.pg_matu ad complete le annotation experiment_an

- Inspect the "Advanced Options" and review the values for the different sections.
 - o Set the "Min percentage of non-missing values globally" to zero.
 - o Set the "Min percentage of non-missing values in at least one condition" to zero.
 - Set the "Adjusted p-value cutoff" to 0.01.
 - Set the "Log2 fold change cutoff" to 1.
 - Set the "Normalization type" to "No normalization".
 - Set the "Imputation type" to Perseus-type
 - Set the "Type of FDR correction" to Benjamini Hochberg.

Note: In the FragPipe-Analyst, a Perseus-like imputation is used by default and the imputed matrix will be used to perform differential expression analysis via Limma. In this type of imputation, missing values are replaced by random numbers drawn from a normal distribution with a width of 0.3 and down shift of 1.8.

Advanced Options ~ Min percentage of non- missing values globally 0	Normalization type No normalization Variance stabilizing normalization (LFQ and DIA only)
Min percentage of non- missing values in at least one condition	Imputation type No imputation
0	 Perseus-type MLE
Adjusted p-value cutoff	 knn min ≥ zero
Log2 fold change cutoff	Type of FDR correction O Benjamini Hochberg O Local and tail area-based



• Click the "Run" button located at the bottom of the page to start the downstream analysis. You should see your result shortly in the web interface.

Examples to show	_															
FragPipe-Analyst	=															
A Hame	Qua	ntification	Absence/Pri	esence												
Analysis 👻																
Data type: DIA –		DE	e a dataset to: results(.csv)	save		118		di	SIGNIFICANT FEATURES 236 out of 3361 7.02% of features differential	ly expressed scross :	all conditions		<u>à</u> . Dawnie	and Report		
Upload protein group (PG) matrix *.tsv								_	-							
Browse report.pg_msti	Rest	ilts Table														
Unioard sample apoptation	Show	10 v en	dries					Sea	rche	Vel	sano plot	Heatmap Featur	re Plot			Result Plots
Browse experiment_sn		Protein ID	Gene Name	NAT_vs_T_log2 fold change	NAT_vs_T_paval (NAT_vs_T_p.adj (significant (NAT_vs_T_significant (insputed ()	Co	mparison		•	Fontsize	Display names	
Upload complete	1	A3KMH1	VWAB	2.85	0.0000464	0.0017	true	true	true						Adjusted p values	
®o-Advanced Options →	2	ARNDB9	PALM3	3.31	0.000019	0.000941	true	true	true	-					Show gene names Don't color pertides from t	the same runtain
Min percentage of non-	з	000217	NDUFSB	3.31	0.000037	0.00145	true	true	false						0	
nsissing values globally	4	000489	PLOD2	-2.21	0.00058	0.00871	true	true	true	54	lact faaturne f	ion Deculte Table to high	light them on the slat			
0	δ	000615	LAD1	1.81	0.000282	0.00679	true	true	true			rom nazona racia co mgi	ing it chain on the plot			
Min percentage of non-		014581	NDUFAB1	4.69	0.000233	0.005	true	true	true					NAT_vs_T		
nsissing values in at least one condition	7	014638	ENPP3	-3.21	0.000121	0.0033	true	true	true						SPTBN2 NDHD	KRT7
0 0	8	014832	PHITH	4.48	0.000505	0.00786	true	true	true			PFKP			NDUFS4 ALDH1A2 SLC12A1 SF FECH	RP1
A Room An and an arrest	9	015020	SPTBN2	4.91	3.37 a- 8	0.0000409	true	true	true	•		NNM	PYGI.	UQCRFS1 A	P1A1 / GPD1L CDH16 IR4 MCCC1 CKMT2 ATP6V1B1 AB4	v
0.01	10	015244	SLC22A2	4.98	0.0000278	0.00118	true	true	true				PLOD3	COX58-P COX	CBNDUFS2_TJP3 TACSTD2 SENDUFS1_DXNAD1EPCAM_CKMT1ABL 1 FREM3CALMACAD8	DH1 DEED
	Shaw	ine 1 to 10 of	3.34Lentries				Previous	1 2 3 4	5 337 Navt	- 3			ERGIC1 SERPINH	1 ETFDH PFXEB2	SIGURE ATP1B1 VIL1 SUOX SPP2 GGH PALM3 CYP17A1 SLC22A2 ADH1	2
1 Coge total change cutom	Ref	resh Table								p-valu		SCGN	GBP5 LOXL2 ALDOC CSPG4 ERD1A ALDOC	MRPL13	UGL2 / OAT SLC23A1 CYP4401 HPD WAR DCXR CYP4401 HPD ACADOB DDXAD30 ALDH6A1	ÅLD OB
										justed			ENPP3 PSMB9 PFN4 CAHH LTN1 NDRG1 DV HPCAL1 PPAT	APH2	TINAG NDUFABI	DX
Normalization type										PV 2			PLOD2 PRDX4 T	GLN2 MRPL22	BMND1 AGXT	
 Varianzestabilizing 										Bq -			17.			
normalization (LRQ and DIA only)																
Insulation type										1.						
 No imputation 													1.74			
Perseus-type													1.			
● 9ILE										0	-					

The results include:

• "Results Table" with the statistical assessment of between the different conditions indicated in the experimental design described in the experiment_annotation.tsv file.

Res	ults Table				
Sho	w 10 v er	ntries	Search:		
	Protein ID	Gene Name	NAT_vs_T_log2 fold change	÷	NAT_vs_T_p.val
1	A3KMH1	VWA8		2.85	0.000046
2	A6NDB9	PALM3		3.31	0.00001
з	000217	NDUF58		3.31	0.00003
4	C00469	PLOD2		-2.21	0.0005
5	000515	LAD1		1.81	0.00025
6	O14561	NDUFAB1		4.69	0.00023
7	O14638	ENPP3		-3.21	0.00012
8	014832	РНҮН		4.48	0.00050
9	015020	SPTBN2		4.91	3.37e-
10	O15244	SLC22A2		4.98	0.000027
<					>
Sha	wing 1 to 10 o	f3,361 entries			
	Previ	aus 1	2 3	4	5 337
					Next
R	efresh Table				



• "Volcano Plot" displaying the protein fold-changes and adjusted p-values for the different comparisons.



• "Heatmap" of protein abundances in each individual sample with hierarchical clustering of the different conditions.



• "Feature Plot", displaying boxplots or violin plots of your favourite (selected) proteins from the Results Table.





• "PCA Plot" showing the result of the principal components analysis.



• "Sample correlation plot" showing the Pearson correlation among the protein abundances matrix of the different samples.





• "Sample CVs" shows the distribution of the coefficient of variation of the different proteins among the replicates of each biological condition.



• "Feature Numbers" shows the number of proteins used for the quantification in each sample.





• "Sample coverage" shows the number of proteins that were found with valid quantitation values in all 10 samples, or only in 9 samples, 8, 7, etc.





• "Missing values - Heatmap" the distribution of missing values per protein and samples in the dataset.



• "Density plot" shows the distribution of protein abundances in the original data, in the filtered data, and in the set of proteins in which imputation was performed. In this case, the original data and the filtered data look the same as we did not force any filtering of the data in the "Advanced Options" of the Analysis section.





Normally, one can start with explorative analyses such as Principal Component Analysis (PCA) to see if the protein data exhibit tumor/normal difference. Following that, one can look for showing differential abundances in tumor samples compared to normal samples. Since the comparison is done for many proteins, multiple test correction is needed to control the false discovery rate.

- Take your time now to explore the results, data and plots in FragPipe-Analyst.
- Once you have explored the different sections, answer the following questions:
 - Do you think the proteome data exhibits any difference between tumor and normal? Inspect the PCA plot. What PC (principal component) captures the major differences between the protein expression profiles in tumor vs normal? How much of the total variance can be attributed to the difference between tumor and normal?
 - Inspect quality control (QC) plots to see if there are any issues with the data (e.g., too few proteins identified in one of the runs, consistent differences in protein abundance distributions between samples, etc.). e.g. inspect "Sample Correlation" plot to see if there are any outliers. Inspect "Density Plot" to see protein abundance distributions before imputation, and after missing value imputation.
 - Find and select a known cancer suppressor of kidney cancer, sFRP1. Find it on the volcano plot. Check its abundance levels across the tumor and normal samples in this dataset using "Feature Plot". Select the protein in the "Results Table" to see the "Feature Plot", and make sure to check the "Show imputed values" option.

Note: When viewing volcano plots or doing enrichment analysis, pay special attention to which side represents which condition to correctly interpret the plots.

- Once the statistical analysis is conducted we get a list of proteins with abundance values altered in between tumor and healthy tisses. However, it is often difficult to make sense of individual genes, especially when there are many of them. Enrichment analysis enables us to aggregate the evidence to biological pathway (Pathway enrichment) or processes (Gene Ontology) to gain a higher-level insight of tumor features.
 - Go to the section "Pathway enrichment". Select the "Hallmark" pathway database, check the "Up" direction, and click "Run Enrichment". What are the most enrich pathways among the proteins in the "Up" direction?





• Repeat the analysis using direction "Down". What are the most enrich pathways among the proteins in the "Down" direction?



Note: Currently the pathway enrichment can be done also with other databases like KEGG and Reactome, in addition to the Hallmark. However, at the moment the pathway enrichment analysis in FragPipe-Analysis only supports human data. In case you want to use any other external tool, you can download the results of the differential expression from FragPipe-Analyst by clicking on the "Save" button on the upper left corner and use in as input for other tools.

Note: To get better agreement with the published results, you can change the "Adjusted p-value cutoff" in the Analysis "Advanced Options" section to 0.05 (from 0.01).

Finally, take into account that we are only using a very small fraction of the global proteome data from the original paper. Therefore, it is likely that you will see discrepancies between the analysis in this tutorial and the final results of the publication, including the number of proteins quantified, the proteins identified as showing a significant change in abundance, and the enriched pathways. However, it is noteworthy to highlight that even this very small dataset is capable of recovering many of the observations in the paper.



Visualization of raw data in Skyline

In this section we will use Skyline to visualize the results in terms of extracted chromatograms generated by FragPipe during the library-free analysis of our dataset. We will use the built-in import wizard of Skyline to import the results and we will dedicate some time to review the raw data. Finally, we will define the experimental groups and perform the statistical inference for the group comparison.

Note This part of the tutorial is based in Skyline-daily 23.1.1.268

- Open a "Blank document" in Skyline
- Go to Settings \rightarrow Default.

Save the document as dia-fragpipe.sky to the folder "Tutorial-5\FragPipe\output".

Note: In order to avoid memory problems, please save your Skyline sessions in the computer's C drive and not in any external drive or USB stick.

• First go to Settings → Peptide Settings and set the parameters as indicated

Peptide Settings	\times
Digestion Prediction Filter Library Modifications Quantification	
Min length: Max length: 6 50	
Exclude N-terminal AAs:	
Exclude potential ragged ends	
Exclude peptides containing:	
Auto-select all matching peptides	
Cancer	



• Go to File \rightarrow Import \rightarrow Peptide Search...

丸 S	kyline-daily				_	×
File	Edit Refine View Settings Tools Help					
1	Start					×
1	New Ctrl+N					
1	Open Ctrl+O					
	Open containing folder					
	Save Ctrl+S					
	Save As					
	Share					
<u>**</u>	Upload to Panorama					
	Import	•	Results]		
	Export	•	Peak Boundaries			
	1 2022MK006-6-samples-francisco.sky		Peptide Search			
	2 Z:\users\Alena Gros\Anna Yuste\2022MT003\reported\2022MT003_013_reported.sky		FASTA			
	3 Z:\users\Alena Gros\Anna Yuste\2022MT003\reported\2022MT003_020_reported.sky		Assay Library			
	4 Z:\users\Alena Gros\Anna Yuste\2022MT003\reported\2022MT003_022_reported.sky		Transition List			
	Exit		Document			
			Annotations			
				,		

- In the Import Peptide Search window, click in "Add Files...".
- Go to the Tutorial-5\Fragpipe\output and select all the interact-*.pep.xml files. Click "Open".
- Set the "Score Threshold" to 0.8 which is the threshold corresponding to 1% FDR when FragPipe performs the search.
- Set Workflow to "DIA". Then, click 'Next'.

-					
Spe	ctral Library				
• B	uild 🔿 Use existing				
Start f	rom:				
Searc	h results (build library directly)	\sim			
Result	t files:				
	File 🔺	Score Type	Score Threshold	^	Add Files
•	interact-CPTAC_CCRCC_W_JHU_2019	PeptideProp	0.8		Remove Files
	interact-CPTAC_CCRCC_W_JHU_2019	PeptideProp	0.8		
	interact-CRTAC_CCRCC_W_IHU_2019	PentideProp	0.0	× .	
ют			-110		
iRT st None	andard peptides:				
iRT st None ☐ Inc	andard peptides:				
iRT st None Inc	andard peptides: V slude ambiguous matches				
iRT st None Inc	andard peptides: v				
iRT st None] Inc	andard peptides: v				
iRT st None Inc	andard peptides: v				
iRT st None] Inc	andard peptides: vilude ambiguous matches				
iRT st None Inc Vorkflo	andard peptides: ude ambiguous matches w A with MS1 filtering				
iRT st None Inc Vorkflo DD/	andard peptides: clude ambiguous matches w A with MS1 filtering				
iRT st None Inc Vorkflo DD/ DD/ DIA	andard peptides: Slude ambiguous matches Sw A with MS1 filtering M				



• Click "Browse..." and go to Tutorial-5\Fragpipe\mzml and select all mzML files as "Result files", and click 'Next'.

🛐 Import Peptide Search		×
Extract Chromatograms		
Results files		
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_NAT		
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00418_T CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00606_NAT		
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-00606_T		
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-01882_NAT CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3L-01882_T		
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00577_NAT CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00577_T		
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3N-00733_NAT		
CPTAC_CCRCC_W_JH0_20190112_L0MOS_C3N-00733_1		
Browse Remove		
Files to import the American		
Hies to import simultaneously:		
INTERPORT MILLING		
	Next >	Cancel

• Check the option to remove the Common prefix, and click "OK".

Import Results	\times
The files you have chosen have a common prefix. Would you like to remove some or all of this prefix to shorten the names used in Skyline?	
◯ Do not remove	
Remove	
Common prefix:	
CPTAC_CCRCC_W_JHU_20190112_LUMOS_C3	
Replicate names: L-00418_NAT L-00418_T L-00606_NAT L-00606_T L-01882_NAT L-01882_T N-00577_NAT N-00577_T N-00733_NAT N-00733_T	
OK Cancel	



• Check the modifications corresponding to "Acetyl (N-term) = [42]" and "Oxidation (M) = M[16]", and click "Next".

🚋 Import Peptide Search	\times
Add Modifications	
The imported search results appear to contain the modifications listed below Please select the ones you wish to add to the document: Acetyl (N-term) = [42] Oxidation (M) = M[16]	•
Add modification	
< Back Next > Cancel	

• Adjust the Transition Settings as depicted in the screenshot below, and click "Next".

丸 Import Peptide Search		\times
Configure Transitio	n Settings	
Precursor charges: 10 2, 3, 4, 5, 6	on charges: Ion types: y, b, p	
Product ions from: ion 3 ~	Product ions to: last ion ~	
Min m/z: 50 m/z	Max m/z: 2000 m/z	
Use DIA precursor windo	ow for exclusion	
Ion match tolerance:	Pick: 6 product ions	
Ion match tolerance unit:	6 min product ions	
	< Back Next > Cance	1



• Adjust the Full-scan settings as in the screenshot below. Note that we set a retention time tolerance of only 0.4 minutes. Then click "Next".

MS1 filtering	
Isotope peaks included:	Precursor mass analyzer:
Count ~	Centroided \sim
Peaks:	Mass Accuracy:
3	10 ppm
MS/MS filtering	
Acquisition method:	Product mass analyzer:
DIA ~	Centroided \sim
Isolation scheme:	Mass Accuracy:
Results only \sim	10 ppm
Use high-selectivity extra	ction
Retention time filtering	
Use only scans within	5 minutes of MS/MS IDs
O Use only scans within	5 minutes of predicted RT
O Include all matching sc	ans

• Finally, in the "Import FASTA (required)" section, set the "Max missed cleavages to 1", then click "Browse..." and go to Tutorial-5\Fragpipe and select the fasta file that you downloaded at the beginning of the tutorial when configuring the FragPipe analysis (2023-10-03-decoys-reviewed-contam-UP000005640.fas). Click "Finish".

🏦 Import Peptide Search	×						
Import FASTA (required)							
Enzyme: Max missed cleavages: Trypsin [KR P] I FASTA records begin with '>' and have the protein name followed by the optional protein description.							
C:\esabido\Tutorial-5-Thursday-Untargeted-DIA-Fragpipe\Fragpipe\2023-10-03-decoys-rev Browse	<u>}</u>						
Decoy generation method: Decoys per target: Shuffle Sequence 1 Automatically train mProphet model							
< Back Finish Cano	el :						



• Check that the Associate Proteins panel looks like the screenshot, and click "OK".

Associate Proteins				?	\times			
Organize all document peptides into associated proteins or protein groups.								
Protein parsimony op Create protein gro Shared peptides are	tions pups ² 2							
Duplicated between	proteins		~					
Find minimal prote	ein list that explains all roteins ² tein: sults:	peptides ²						
	Mapped	Unmapped	Targets					
Proteins	1970	38,948	1970					
Peptides	3045	0	3477					
Shared Peptides	483		483					
1971 proteins, 6954 peptides, 6954 precursors, 62,586 transitions OK Cancel								

After clicking 'OK' in the next dialog window, Skyline will start to load the data.





Finally, we will explore the extracted chromatogram in Skyline to inspect the intensity of the peptide SMEDSVDVSAPK from protein sp|Q8IVF2|AHNK2_HUMAN, a known tumor biomarker of kidney cancer.

- Go to View C Transitions C Products to only show fragments
- Go to Edit C Find... and search for the peptide SMEDSVDVSAPK
- In the 'View' menu, select 'Peak Areas' and then 'Replicate Comparison' to visualize the intensities among all samples, to confirm the upregulation of this peptide in tumor samples. Make sure that the option "Normalized To" is set to "Default (None). You can right click in the Peak Areas plot to check it.

How does the data look like? In which samples do you have an associated identification? Do the signals in these samples look better?



Now look for protein sFRP1 and inspect the different peptides that have been identified for this protein. How to they look like? Are the signals clearer in the healthy samples compared to the tumor samples?

