Supplementary Note

RNP^{xl} data processing tutorial

1 Installation and preparation of MS data and databases

1.1 Before getting started

Please read the following points carefully and remember them throughout the data analysis:

- **avoid spaces in file and folder names**
  Some programs do not tolerate spaces in file and folder names. Spaces might lead to programs, pipelines or tools to crash with errors that do not indicate spaces as potential error source. Therefore, files and folder names are one of the first things that should be checked when a tool crashes with inexpressive errors.
- **be sure to rename result files after running them through a pipeline**
  When carrying out an entire data analysis workflow, a number of .mzML raw data and .idXML/.csv result files will be created. Renaming each and every one will help to keep track of everything that was done. It might also prevent errors, e.g. picking (centroiding) raw data twice. As described below, some pipelines will also change the file name completely and renaming is essential.

1.2 Installing the required software

OpenMS[1] is an open source framework for LC-MS data analyses. It consists of various algorithms and tools to process mass spectrometry data. Tools can be combined in workflows, allowing powerful and custom tailored data processing. OpenMS employs open data formats from the Proteomic Standard Initiative[2] (PSI). RNP^{xl} has been integrated into the OpenMS framework and is available after OpenMS installation.

The OpenMS installer can be downloaded from www.openms.de/downloads.

The installer contains several programs:

- ProteoWizard[3] (external tool), required for data conversion
- OMSSA[4] (external), database search engine and the makeblastdb tool to generate databases in OMSSA format
- TOPPAS[5], graphical user interface to assemble and run data analysis pipelines
- TOPPView[6], graphical user interface to view mass spectrometry data in mzML format and annotate search results in idXML format

While all tools have been developed platform independently, RNP^{xl} has been extensively tested on Windows 7 64bit and Windows Server 2008 systems as main platform. Additionally, the external msconvert tool requires a windows platform for conversion of the vendor specific raw file format to the open mzML[7] data format used as input for the analysis pipeline. We therefore recommend performing the analysis on a Windows 64 bit platform like Windows 7.

Throughout this tutorial it is assumed that Windows is used. This implies that commands (e.g. for starting a command prompt) or naming of hard discs might differ on other platforms.

A computer with a quad-core processor and 4-8 GB RAM is sufficient to analyze experiments of medium complexity in a reasonable time. The limiting factor is typically the database search. Therefore, searches against small databases (less than 50 proteins) can be performed on less powerful hardware. For searches against entire proteomes, increasing processing power might speed up the searches considerably.

All pipelines described in this script along with a small sample data set are available from:
http://open-ms.de/publications/rnpxl/
On this site and if necessary, updated versions of this tutorial, FAQs and/or trouble shooting guides will also be posted.

1.3 Command prompt basics

Some of the following steps require running programs from the command prompt window. This is found under Programs/Accessories or searching “cmd” in the Windows start menu.

In many cases the desired program has to be started from a certain folder. In order to get into a specific folder, use the change directory commands:

- `cd ..` will change the directory back one level, e.g. `from C:\users\user1` to `C:\users`
- `cd user1` will change the directory from `C:\users` back to `C:\users\user1`
- `cd\` will get to the root of the drive, e.g. from `C:\users\user1` to `C:\`
- to change to a different drive, e.g. D:\, type `D:`
- if the folder name is unique, only the first few letters need to be typed and the full name can be automatically completed by adding `*` or the tab key, both followed by enter

Be sure to provide all extensions described in the corresponding instructions, otherwise the programs might not run correctly and give erroneous results which might not become obvious until much later in the data analysis workflow.

1.4 Preparation of MS data

Use the ProteoWizard tool msconvert from the command line to convert the raw data in vendor file format to the mzML format. Note that the vendor libraries are required for the conversion and are only automatically installed for Windows.

In the command prompt window, access the folder where the raw data is stored. Running msconvert without any argument, e.g. by typing “msconvert” and confirming with enter, gives a list of available options. However, to convert Thermo .raw data into .mzML, msconvert is typically run without any argument but the filename, i.e. simply with:

```
msconvert file.raw
```

If an instrument from a different manufacturer is used, check the list of supported proprietary file formats on http://proteowizard.sourceforge.net/formats/index.html.

Several files can be converted in parallel by opening individual command prompt windows.

Alternatively, the MSConvert graphical user interface can be used. The corresponding executable can be found in the share\OpenMS\THIRDPARTY\pwiz-bin in the OpenMS folder. The input raw files have to be added to the left list, and the output directory have to be specified. The "use zlib compression" box on the left bottom has to be unchecked.

Expert note: The cmd version by default uses 64bit coding for the mass values and 32bit coding for intensities. In the GUI version, both mass values and intensities are either encoded with 64bit or 32bit. Therefore, the resulting files differ considerably in size. However, this does not influence the data analysis. We recommend to use the cmd version or the GUI with 64bit coding.

1.5 Preparation of protein databases

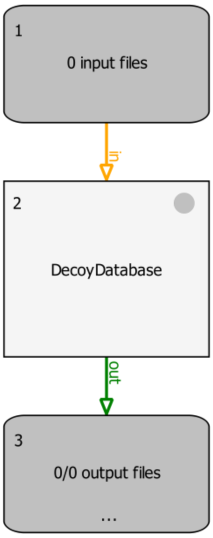
Typically, two databases will be used in the data analysis workflow:

1. A database containing only the protein(s)/proteome of interest for the precursor variant search (step 5 below).
2. A database containing the protein(s)/proteome of interest as well as contaminants with both forward and reverse sequences for the ID filter pipeline (step 3 below). Contaminants are added to remove spectra corresponding to keratins etc. Reverse sequences are necessary to determine the false discovery rate.

The two databases are prepared according to the following steps:

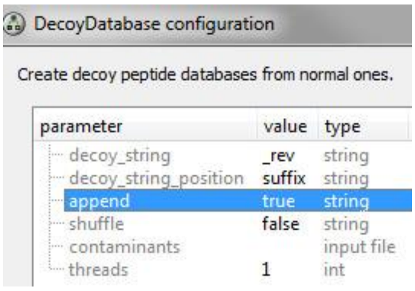
- Download or assemble the desired database containing the protein(s)/proteome of interest and save it in fasta format (e.g. from Uniprot or NCBI).
- Create a second version of the database by adding contaminant sequences, for example the MaxQuant[8] contaminant database (http://maxquant.org/downloads.htm) or cRAP protein sequences (http://www.thegpm.org/crap/index.html).
  Add e.g. _contaminants to the name of the second database to distinguish it from the first.
- Create a target/decoy version of the second database (_contaminants) with the following pipeline (sample pipeline 00_decoy_database; general description of TOPPAS pipelines see below):

Input is the database in `.fasta` format, output is the target/ decoy database which should be renamed in order to distinguish it from the database containing only target sequences.



Supplementary Note Figure 1: Decoy database creation

Important is to set the “append” option to true so that the decoy sequences are inserted after the target ones.
The output file should be renamed (e.g. as _contaminants_tardecoy).



DecoyDatabase configuration

Create decoy peptide databases from normal ones.

parameter	value	type
decoy_string	_rev	string
decoy_string_position	suffix	string
append	true	string
shuffle	false	string
contaminants		input file
threads	1	int

Supplementary Note Figure 2: Parameters of the DecoyDatabase tool

- Convert both protein databases.
  As the search engine OMSSA doesn’t directly read databases in fasta format but uses the BLAST[9] protein sequence file format (.psq) internally, the fasta file must be converted using the BLAST command line application makeblastdb[10]. To this end, enter:

  ```
  makeblastdb -in database.fasta -dbtype prot
  ```

  in the command prompt executed in the folder where the fasta database resides. Be sure to keep all formats of one database (fasta and psq as well as phr and pin) under the same name and in the same folder while carrying out the data analysis.

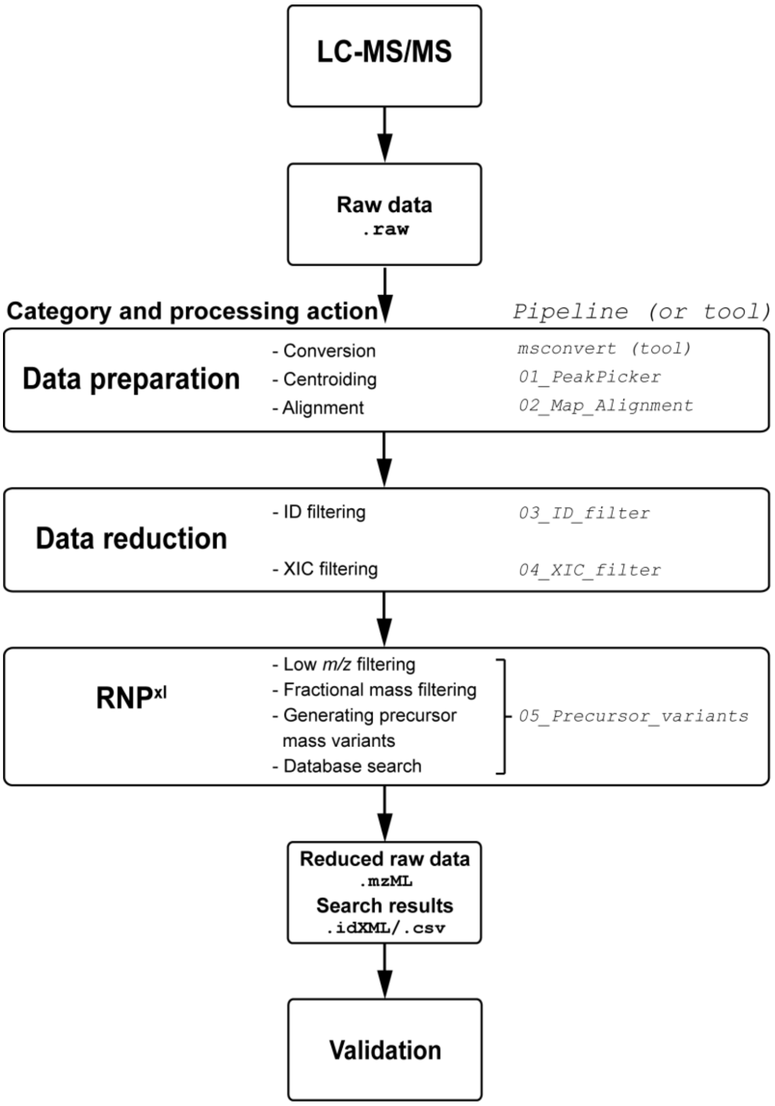
2 Workflow for the analysis of cross-linking data

As mentioned before, we use TOPPAS (The OpenMS Proteomics Pipeline Assistant) as workflow system. It allows conveniently building and executing the necessary processing steps. Built from smaller blocks with simple functionality, it allows easy adaption to different instrumental or experimental setups.

Data analysis pipelines are stored in the .toppas format and can be opened with TOPPAS. Parameters of the different nodes can be viewed and changed by right-clicking on the respective symbol. Input files are chosen accordingly. For more details on TOPPAS, see the tutorial available at: http://open-ms.sourceforge.net/documentation/.

After adaptation of parameters as described in the subsequent steps, each workflow can be executed by selecting `Pipeline->Run` from the main menu and specifying the location where the result files should be stored. We suggest setting the maximum number of jobs to one, i.e. not running more than one pipeline simultaneously.

We compiled a sample dataset and constructed a workflow broken down into five pipelines to explain and demonstrate the different steps of our analysis workflow:



Supplementary Note Figure 3: Workflow and

step	short description	sample pipeline
0	DECOY_database This pipeline will create a target/decoy database (see previous section)	00_DECOY_database
1	data centroiding/peak picking This pipeline will reduce the amount of data (file size) and is required for subsequent steps.	01_PeakPicker_MS1only 01_PeakPicker_MS1andMS2
2	map alignment This pipeline will correct for retention time shifts between control and UV sample.	02_Map_alignment
3	ID filter This pipeline will remove MS² spectra with a good match to a pure peptide from the raw data.	03_ID_filter
4	XIC filter This pipeline will remove MS² spectra of species which also appear in the control from the raw data.	04_XIC_filter
5	Precursor variant search This pipeline will create the precursor mass variants and execute the OMSSA searches.	05_Precursor_variants

The output files of a step are used as input for the subsequent step. We strongly recommend centroiding (step 1) and performing the filtering steps 2 to 4 although they are not needed to run the precursor variant search.

The pipelines along with a small sample data set are available from:
http://open-ms.sourceforge.net/workflow-integration/toppasworkflows/

2.1 Step 1: Data centroiding/peak picking

Several of the following pipelines (alignment and XIC filter) need the data in centroided format (control as well as UV irradiated sample). As an additional advantage, the file size is reduced considerably (sometimes over 90%), increasing speed of subsequent data analysis steps and easing file handling. Furthermore, centroided files can be used to evaluate the search results as they still contain all spectra but are considerably smaller. Note that the sample data has already been centroided to reduce the size of the download. If you use the RNP^{xl} sample data please skip step 1 and proceed with step 2.

We provide two sample pipelines set up for centroiding either only MS^1 or both MS^1 and MS^2. The appropriate pipeline has to be chosen according to the mode in which the data was acquired. Picking of centroided data leads to undesired data loss and has to be avoided.

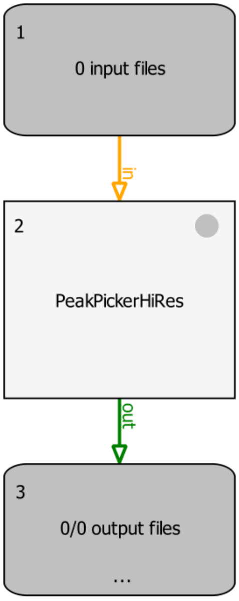
There might occur a warning “Unhandled cvParam ‘MS:1000927’ in tag ‘scan’”. This is a conversion artifact and can be ignored.

All files can be submitted at once and will be consecutively processed by the pipeline.

Peak picking - MS^1 only

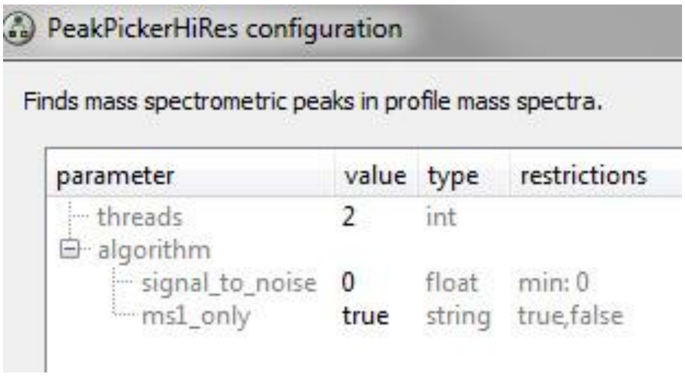
In many cases, MS^2 spectra are already acquired in centroid mode. Therefore, only the MS^1 spectra have to be picked. To this end, the “ms1_only” option of the PeakPickerHiRes node must be set to “true”, which is the case in the 01_PeakPicker_MS1only sample pipeline.

Pipeline



Supplementary Note Figure 4: Peak picking pipeline, MS^1 only

Parameters



PeakPickerHiRes configuration

Finds mass spectrometric peaks in profile mass spectra.

parameter	value	type	restrictions
threads	2	int	
algorithm			
signal_to_noise	0	float	min: 0
ms1_only	true	string	true,false

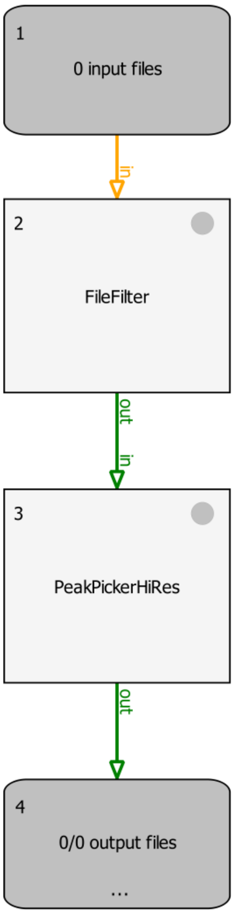
Supplementary Note Figure 5: Parameters of the PeakPickerHiRes tool

To run the pipeline, nothing except the input files need to be changed.

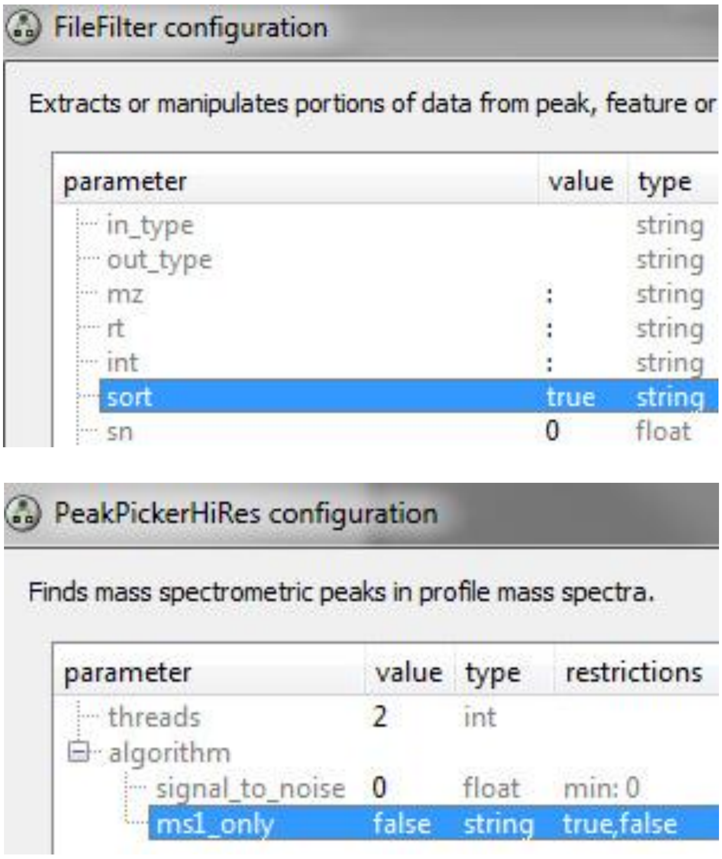
Output is a mzML file with the picked data. To differentiate the original and the picked data, it is advisable to rename the output file, e.g. with the suffix _picked behind the filename (e.g. `input_UV.mzML` should be manually renamed to `input_UV_picked.mzML`).

Peak picking - MS^1 and MS^2

If both MS^1 and MS^2 are acquired in profile mode, both levels need to be centroided. For this, the MS^{1}only pipeline can be modified by setting the "ms1_only" to false (as it is set in the 01_PeakPicker_MS1andMS2 sample pipeline). The provided sample pipeline for both MS^1 and MS^2 contains an additional node, the FileFilter with the "sort" option enabled. This is due to specifics of raw data acquired on a Q Exactive instrument and is obsolete if data is recorded e.g. with an Orbitrap Velos. I.e., for data acquired on a Q Exactive with both MS^1 and MS^2 in profile mode, use the pipeline below.



Supplementary Note Figure 6: Peak picking pipeline, MS^1 and MS^2



FileFilter configuration

Extracts or manipulates portions of data from peak, feature or

parameter	value	type
in_type		string
out_type		string
mz	:	string
rt	:	string
int	:	string
sort	true	string
sn	0	float

PeakPickerHiRes configuration

Finds mass spectrometric peaks in profile mass spectra.

parameter	value	type	restrictions
threads	2	int	
algorithm			
signal_to_noise	0	float	min: 0
ms1_only	false	string	true,false

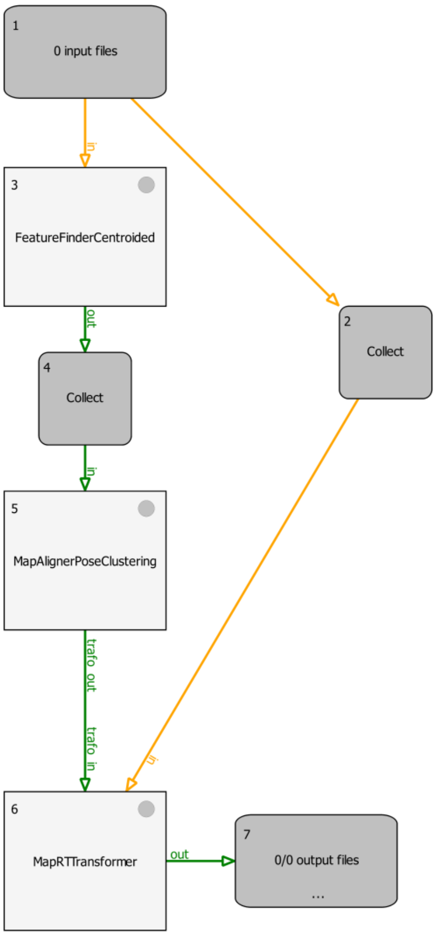
Supplementary Note Figure 7: Parameters for the PeakPickerHiRes and FileFilter tool

In the input node, the raw mzML files are specified.

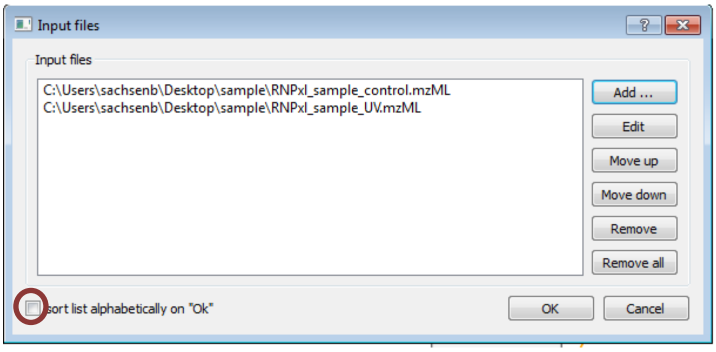
Output is a mzML file with the picked data. To differentiate the original and the picked data, it is advisable to rename the output file with the suffix _picked behind the filename (e.g. `input_UV.mzML` should be manually renamed to `input_UV_picked.mzML`).

2.2 Step 2: Map alignment

This pipeline can correct for minor retention time shifts between control and UV irradiated sample. This might be necessary for the XIC filter pipeline (step 3) to give better results.



Supplementary Note Figure 8: Map alignment pipeline


Input files

Input files

C:\Users\sachsenb\Desktop\sample\RNPxl_sample_control.mzML
C:\Users\sachsenb\Desktop\sample\RNPxl_sample_UV.mzML

□ sort list alphabetically on "Ok"

Supplementary Note Figure 9: Input files

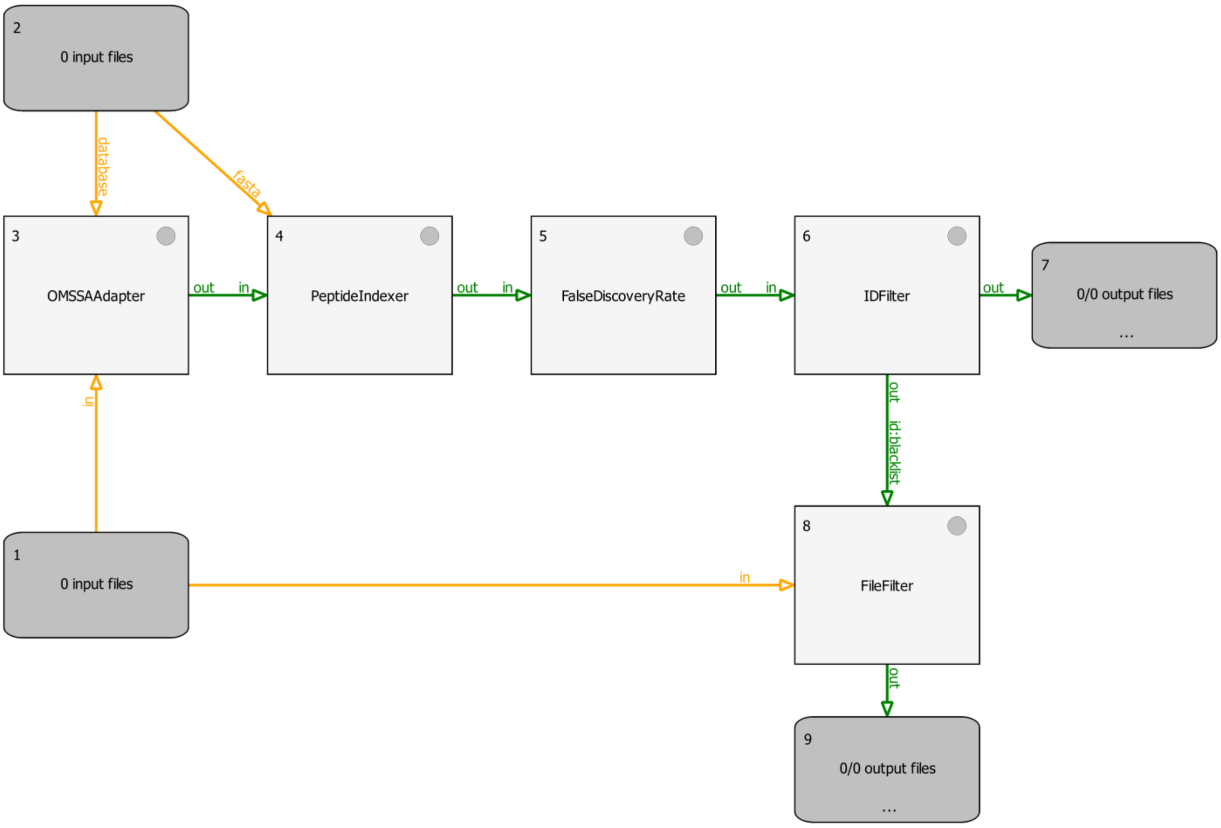
As input files, choose **first** the control and **second** the UV sample. Be sure the “sort list alphabetically on OK” option in the input file dialogue is **not** chosen as this might reverse the order depending on the file names that were chosen (see Supplementary Note Figure 9). The order ensures that the control is shifted relative to the UV and the retention times of the UV remain unchanged. This enables direct comparison of processed data and data in vendor format based on retention time.

Important: It is not possible to submit more than one pair of control and UV samples. Note that the output files will have the same name, i.e. that of the control. The first output file will be the UV with the name of the control and extension .0.mzML, the second output file will be the control with its correct name. This is unfortunate but requested to be fixed in further releases of the workflow system. To work around this issue, manually rename the second file with the suffix `_aligned.mzML`, to distinguish it from the original, unaligned data. The UV file is unchanged and can be deleted. Since the file size does not significantly change due to alignment, it is an additional reference for correct assignment of output file to control and UV sample

In case of doubt whether the data needs to be aligned or not, the XIC filter (see below) can be run with and without alignment. The results can be evaluated with respect to number of remaining spectra (see useful tips – FileInfo tool) and thus it can be verified whether the alignment leads to an increased number of filtered spectra.

2.3 Step 3: Filtering by identifications

This filtering step will remove MS² spectra with a good match to a pure peptide from the raw data (false discovery rate FDR < 1%).



Supplementary Note Figure 10: Identification filter pipeline

This pipeline will read a mzML file and a fasta database and produce two output files incorrectly assigned with the name of the database:

- Node 9 will give the new mzML file where all MS² spectra with good (in terms of low q-value) peptide identification are removed. The output file should be renamed, e.g. to sample_IDfiltered.
- Node 7 will give an idXML file with all peptide identifications that led to filtering of the corresponding spectra. This file can later be used to annotate the original mzML for manual validation of this filtering step but is not needed for further analysis. Again, the output file should be renamed, e.g. to sample_resultsIDfilter.

The fasta database has to contain target and decoy sequences and can contain contaminant sequences (see section about preparation of protein database).

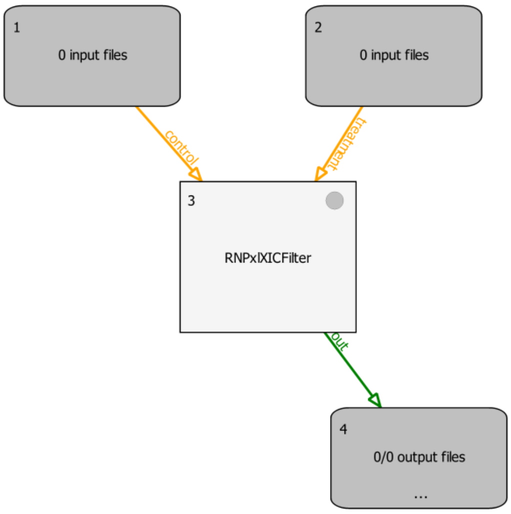
In the OMSSAAdapter node, the parameters for the OMSSA database search are specified. Mass accuracy for both MS^1 and MS^2 as well as posttranslational modifications should be adjusted to the experimental workflow. In addition, the maximum allowed missed cleavage sites and other search parameters can be changed if desired or needed. In case an enzyme other than trypsin has been used, this needs to be specified in both the OMSSAAdapter and the PeptideIndexer node. The correct path for omssa_executable should be confirmed. The corresponding folders should not contain any spaces, i.e. OMSSA cannot be installed into the standard “Program Files” folder.

PeptideIndexer annotates for each search result whether it is a target or a decoy hit and also links peptide hits to proteins. FalseDiscoveryRate computes q-values for the identifications. IDFilter selects only those identifications with a q-value of less than 0.01 (the FDR threshold can be changed using the score:pep parameter). FileFilter takes the filtered peptide list and the original mzML and removes the MS^2 spectra that gave rise to a confident peptide identification.

2.4 Step 4: Filtering by XIC

The XIC filter pipeline will remove MS^2 spectra from the UV sample measurement if the same precursor appears in the control with comparable intensity and retention time. If a considerable retention time shift is observed between control and UV, the files should be aligned (step 2) prior to applying this pipeline.

The XIC filter pipeline needs the control mzML file as left input and UV mzML as right input. Not more than one pair of control and UV irradiated sample should be submitted.



Supplementary Note Figure 11: XIC filtering pipeline

IMPORTANT: The output node will produce an mzML file that is incorrectly named after the control mzML. Rename back to UV and change the filename to distinguish it from the original mzML (e.g. _UV_XICfiltered). This behavior is caused by internal handling of filenames in the workflow system and will be addressed in further updates of TOPPAS.

2.5 Step 5: Precursor variant search

After the original data has been processed, aligned and filtered, the actual search for precursor mass variants can take place. Note that the previous steps are not a prerequisite, but are advisable as they significantly reduce the number of false positive results. Correct filtering can be evaluated by running RNP^{xl} on filtered and original data, checking whether true positives are incorrectly filtered.

The RNP^{xl} tool uses an .ini file which contains the parameters for the database search as a template for all subsequent precursor variant searches.

Preparing the OMSSAAdapter .ini file

We provide the OMSSAAdapter .ini file as a starting point. Alternatively, it can be created from scratch using a call to:

```
    OMSSAAdapter –write_ini IniFile.ini
```

on the command line.

The parameters in the ini file can be edited using a call to:

```
    INIFileEditor IniFile.ini
```

This starts the INIFileEditor which allows to enter or adapt the search parameters. Alternatively, open the program from the Windows start menu or Windows Explorer, it is installed alongside OpenMS in the same folder. In contrast to standard database search, e.g. as performed in the ID filter, several considerations need to be taken into account:

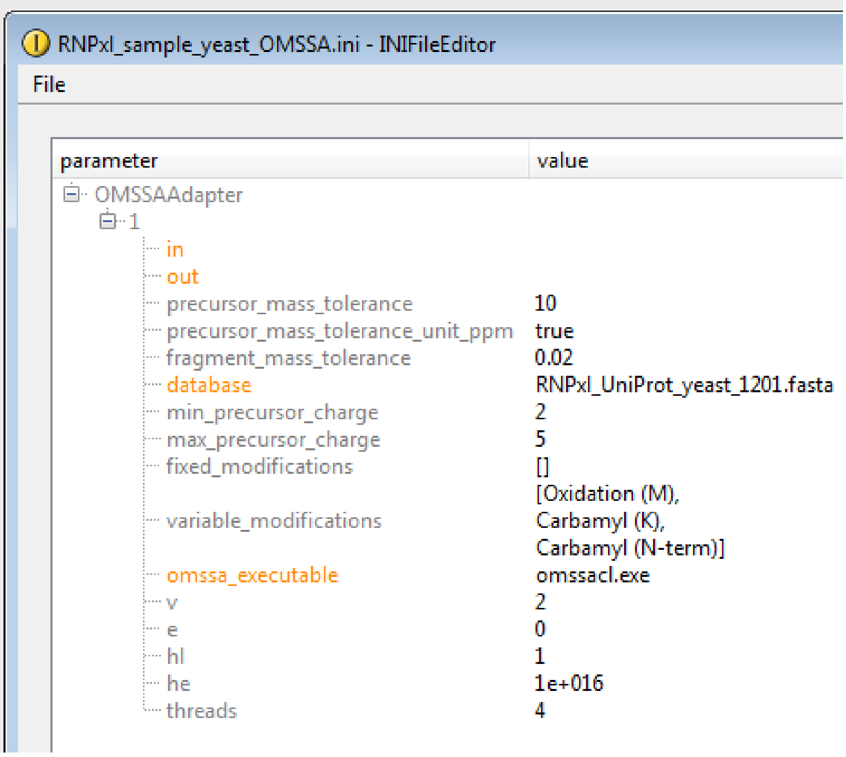
- The precursor mass accuracy should not be limited too strictly. After subtraction of the calculated RNA mass, the absolute error remains the same. Since the relative error is then calculated relative to the smaller precursor mass variant, it will be greater than in reference to the experimental precursor mass. In a very simplified example, both peptide and RNA have a mass of 600 Da, the additive cross-link therefore has a mass of 1200 Da. The experimental mass of the cross-link could be determined to be 1200.0060 Da, the absolute error of 60 mDa corresponds to a relative error of 5 ppm. For the precursor variant search, 600 Da will be subtracted from the experimental precursor mass 1200.0060 Da to give 600.0060 Da. The absolute error of 60 mDa now corresponds to a relative error of 10 ppm. Therefore, the mass accuracy for searches of precursor mass variants should be chosen considerably larger than the maximum relative error expected from the mass spectrometer. We recommend setting the mass accuracy to 10 ppm for Orbitrap instruments.
- The number of posttranslational modifications should be very limited. A great number of PTMs will increase search time and number of false positives results. In addition, certain PTMs have the same elemental composition as RNA modifications. For example, phosphorylation of peptides and phosphate groups of RNA termini both have the composition HPO_3. If phosphorylation is considered as a peptide PTM and loss of terminal phosphate as an RNA modification, both combinations of unmodified peptide and RNA and phosphorylated peptide and RNA without terminal phosphate have the same mass. Unfortunately, OMSSA is often not capable to distinguish both cases correctly, correct placement of the phosphate group needs to be confirmed manually. Since the observation of a phosphorylated cross-link is unlikely, we recommend to omit phosphorylation from precursor variant searches and only include it in the ID filter if desired. Of note, adenosine and guanosine differ by one oxygen atom. If oxidation is considered as a peptide PTM, e.g. on methionine, it has to be kept in mind that this can also lead to incorrect placement of the oxygen. However, this is rarely observed and usually easily spotted by comparing the cross-linked RNA to the observed marker ions. We usually only allow oxidation

of methionines and carbamylation of primary amines (N-terminus and lysines, chemical modification due to hydrolysis in the presence of urea) as posttranslational modifications in precursor variant searches.

The database should only be specified in the RNP^{xl} tool but left empty in the OMSSA ini file.

The correct path for omssa_executable should be chosen (again: no spaces, see previous section).

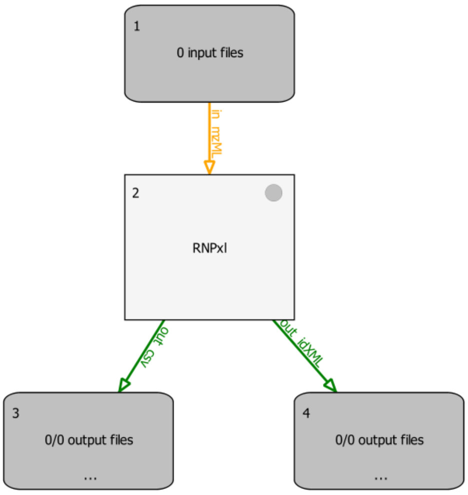
For an explanation of the other parameters take a look at the description in the INIFileEditor or the OpenMS manual. Figure 11 shows the parameters used in our analysis.



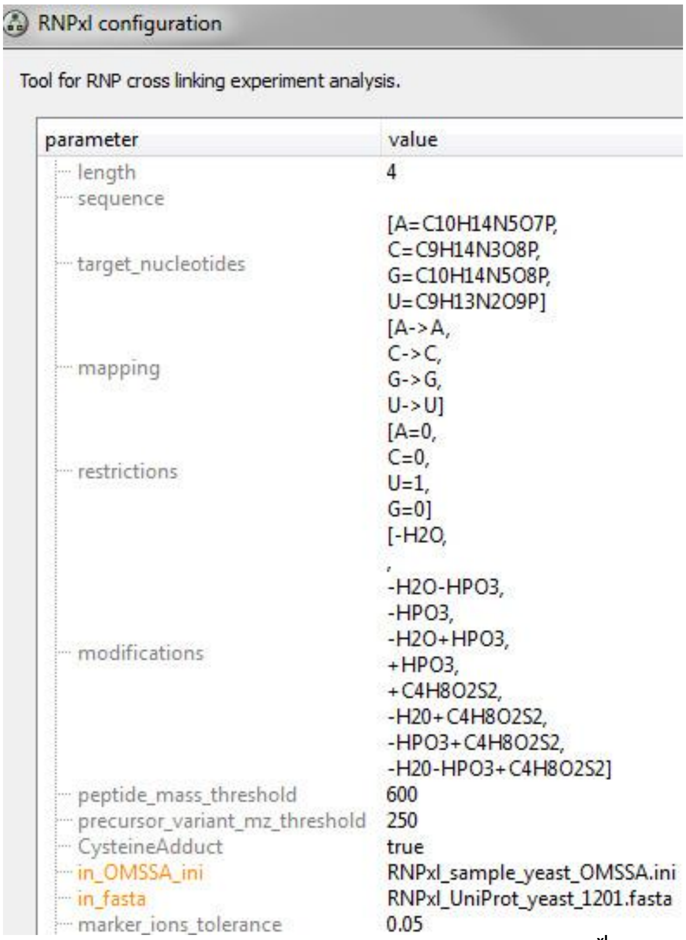
parameter	value
OMSSAAdapter	
1	
in	
out	
precursor_mass_tolerance	10
precursor_mass_tolerance_unit_ppm	true
fragment_mass_tolerance	0.02
database	RNPxl_UniProt_yeast_1201.fasta
min_precursor_charge	2
max_precursor_charge	5
fixed_modifications	[]
variable_modifications	[Oxidation (M), Carbamyl (K), Carbamyl (N-term)]
omssa_executable	omssacl.exe
v	2
e	0
hl	1
he	1e+016
threads	4

Supplementary Note Figure 12: OMSSAAdapter template .ini file as used by the RNP^{xl} tool.

Configuring and searching for precursor variants with the RNP^{xl} tool



Supplementary Note Figure 13: Precursor variation search pipeline


RNPxl configuration

Tool for RNP cross linking experiment analysis.

parameter	value
length	4
sequence	

target_nucleotides	[A=C10H14N5O7P,
	C=C9H14N3O8P,
	G=C10H14N5O8P,
	U=C9H13N2O9P]

mapping	[A->A, C->C, G->G, U->U]

restrictions	[A=0,
	C=0,
	U=1,
	G=0]

modifications	[-H2O, , -H2O-HPO3, -HPO3, -H2O+HPO3, +HPO3, +C4H8O2S2, -H20+C4H8O2S2, -HPO3+C4H8O2S2, -H20-HPO3+C4H8O2S2]
peptide_mass_threshold	600
precursor_variant_mz_threshold	250
CysteineAdduct	true
in_OMSSA_ini	RNPxl_sample_yeast_OMSSA.ini
in_fasta	RNPxl_UniProt_yeast_1201.fasta
marker_ions_tolerance	0.05

Supplementary Note Figure 14: Parameters of the RNP^{xl} tool

Input is a (filtered or not, centroided or not) mzML file. One output node (4) produces the idXML, the other (3) the csv file.

The RNP^{xl} tool is highly customizable and allows a large number of different cross-linking experiments. This degree of customizability is achieved by introduction of a range of parameters used to describe the experiment.

Length gives the maximum length for the oligonucleotides that will be created and later used to generate precursor mass variants. It is strongly advised to keep this value between 1 and 4 as search time increase exponentially with this parameter and confident identification of cross-links to longer oligonucleotides after ESI-MS is highly unlikely.

Under sequence, an RNA sequence can be defined if desired. Only RNA combinations that actually appear in the sequence are then considered for precursor variant generation. Especially for experiments with short oligonucleotides, this can significantly reduce the number of RNA combinations and thus greatly speed up analysis time and limit the number of false positives. When empty, all RNA combinations will be considered. If the sequence contains substituted nucleotides, use the same one letter code as specified in the target_nucleotide section.

Under target_nucleotides, the nucleotides considered in the precursor variant generation are specified. Nonstandard nucleotides are added by assigning a one letter code (e.g. X) and providing the sum formula (full nucleotide, not chain form). This allows experimental setups with substituted nucleotides like 4-thio-U or 6-thio-G, DNA, with or without isotopic labeling, etc. See the RNP^{xl} configuration examples in the appendix of this document for further information.

Under mapping, all nucleotides specified in the target_nucleotide parameter should be listed in a mapping rule. If no substitution is used, each nucleotide is simply mapped to itself as done in the example above. If e.g., all uracils of the sequence are substituted by a nucleotide X, the mapping rule U->X is used. If the labeling isn't complete but considered random, both mappings U->U and U->X are added to consider both variants.

Under restrictions, RNA combinations used for precursor variant generation can be required to contain a certain nucleotide. For example, after setting U=1, only oligonucleotides which contain at least one U will be considered.

Under modifications, the modifications which should be considered for all nucleotide combinations are specified. In the example above, the typical modifications for a standard experiment with the cysteine adduct 152 are shown. The empty entry in the list corresponds to no modification, which must be explicitly provided. Additional modifications can be defined by giving their chemical sum formulas.

Under peptide_mass_threshold, the mass value is specified which is used to filter the precursors before the precursor mass variants will be generated. In the example, all precursors below 600 Da will be filtered. In a typical instrument method, only precursors with a minimal charge of two will be fragmented, and the MS^1 is recorded starting at *m/z* 350. Consequently, each fragmented precursor should have a mass of at least 700 Da. Therefore, this option is only applied to precursors with incorrect charge state assignment during data acquisition and is typically obsolete for modern instruments.

Under precursor_variant_mz_threshold the minimal *m/z* required for a precursor mass variant AFTER subtraction of the calculated oligonucleotide masses is specified. Bases on our experimental data, the smallest observed value was 350, so 250 or 300 would be a conservative choice.

CysteinAdduct should be set to “true” if the 152 Da modification should be considered as a modification that can occur without any nucleotide.

Finally, the OMSSA ini file and the database in fasta format must be specified.

Note that the database should not contain contaminants or decoy sequences. No cross-links are expected to originate from either sequences. Including contaminants and decoy sequences for precursor variant searches will considerably slow down the searches and increase the number of false positives. Therefore, different databases are used for ID filtering and in the RNP^{xl} tool.

All files to be processed with the same parameters for precursor variant generation and database search can be submitted at once and will be consecutively handled by the pipeline.

3 RNP^{xl} configuration examples

In the examples listed below, only the parameters affecting the oligonucleotide generation are shown.

Example 1a: Standard experiment with input RNA sequence and cross-links to uracil

The maximum length of the cross-linked oligonucleotide is limited to four. An input RNA sequence is defined, only nucleotide combinations that appear in this sequence are considered. The four standard nucleotides are defined as targed nucleotides and mapped on themselves. The restriction “U=1” requires one uridine in all RNA combinations considered for precursor mass variant generation. Typical modifications as well as cysteine adducts are considered for each RNA combination.

length	4
sequence	ACUGCAUGAG
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P
mapping	A->A, C->C, G->G, U->U
restrictions	A=0, C=0, G=0, U=1
modifications	-H2O, , -H2O-HPO3, -HPO3, +C4H8O2S2, -H2O+C4H8O2S2, -HPO3+C4H8O2S2, -H2O-HPO3+C4H8O2S2
CysteineAdduct	true

Example 1b: Standard experiment without limiting RNA sequence or cross-linked nucleotide

Keep the sequence parameter empty to generate all possible RNA fragments. In addition, the “restriction” for U is set to zero to allow RNA combinations without uridine. All other parameters are kept as described above.

length	4
sequence	
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P
mapping	A->A, C->C, G->G, U->U
restrictions	A=0, C=0, G=0, U=0
modifications	-H2O, , -H2O-HPO3, -HPO3, +C4H8O2S2, -H2O+C4H8O2S2, -HPO3+C4H8O2S2, -H2O-HPO3+C4H8O2S2
CysteineAdduct	true

Example 2: Experiment with a site-specifically substituted oligonucleotide

In this example, the cross-linked RNA is site-specifically labeled with 4-thio-uridine at position 7. This nucleotide is defined as “Y” in “sequence” and “mapping”, as well as with its elemental composition in the “target_nucleotides”. Only 4-thio-uridine is expected to form cross-links, therefore only RNA combinations with 4-thio-uridine are considered for precursor mass variant generation by setting “Y=1” in “restrictions”. The RNA “modifications” are adjusted, omitting the cysteine adduct and adding loss of H_2S (see Kramer et al. [11] for details on 4-thio-uridine specific losses).

length	4
sequence	ACUGCAYGAG
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P, Y=C9H13N2O8PS
mapping	A->A, C->C, G->G, U->U, Y->Y
restrictions	A=0, C=0, G=0, U=0, Y=1
modifications	-H2O, , -HPO3, -H2O-HPO3, -H2S, -H2S-HPO3
CysteineAdduct	False

Example 3: Experiment with isotopically labeled nucleotides

In this example, nucleotide six of the input sequence is substituted site-specifically with adenosine containing one heavy 13C. Input sequence, target nucleotides, and mapping are adjusted accordingly from the Experiment 1a.

length	4
sequence	ACUGCYUGAG
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P, Y=(13)C1(12)C9H14N5O7P
mapping	A->A, C->C, G->G, U->U, Y->Y
restrictions	A=0, C=0, G=0, U=0, Y=0
modifications	-H2O, , -H2O-HPO3, -HPO3, +C4H8O2S2, -H2O+C4H8O2S2, -HPO3+C4H8O2S2, -H2O-HPO3+C4H8O2S2
CysteineAdduct	True

4 Interpreting the output of the RNP^{xl} tool

The RNP^{xl} tool was run with following parameters on the sample dataset (without prior XIC or ID filtering):

length	3
sequence	
target_nucleotides	A=C10H14N5O7P, C=C9H14N3O8P, G=C10H14N5O8P, U=C9H13N2O9P
mapping	A->A, C->C, G->G, U->U
restrictions	A=0, C=0, G=0, U=1
modifications	-H2O, ,
CysteineAdduct	true

Output that has been printed in the log window of TOPPAS and can also be found in the toppas.txt file created in the same folder as the TOPPAS_out folder containing all results files. Note that the toppas.txt will be overwritten if another pipeline is run in the same folder, it could be renamed if one wants to keep it for future reference.

13:28:36 NOTICE: RNPxl of node #2 started. Processing ... Min. count restrictions: min. count: U 1	**At least one U must be contained in RNA fragment**
Modification: H2O subtractive: 1 Modification:	**Loss of water is considered as modification**
target sequence(s):1	**≥1 if substitutions are specified in the mapping**
target nucleotide: A modifications: A-H2O1 C10H12N5O6P1 modifications: A C10H14N5O7P1 target nucleotide: C modifications: C-H2O1 C9H12N3O7P1 modifications: C C9H14N3O8P1 target nucleotide: G modifications: G-H2O1 C10H12N5O7P1 modifications: G C10H14N5O8P1 target nucleotide: U modifications: U-H2O1 C9H11N2O8P1 modifications: U C9H13N2O9P1	**All considered nucleotides (including modifications), how they are reported, and their elemental composition**
Filtering on restrictions... 1 C18H22N4O16P2 612.051 (UU-H2O1) 2 C18H23N5O15P2 611.067 (CU-H2O1) 3 C18H24N4O17P2 630.061 (UU) 4 C18H25N5O16P2 629.077 (CU) 5 C19H23N7O14P2 635.078 (AU-H2O1) 6 C19H23N7O15P2 651.073 (GU-H2O1) 7 C19H25N7O15P2 653.088 (AU) 8 C19H25N7O16P2 669.083 (GU) 9 C1H2N3O6 151.994 (cysteine adduct) 10 C27H33N6O24P3 918.076 (UUU-H2O1) 11 C27H34N7O23P3 917.092 (CUU-H2O1) 12 C27H35N6O25P3 936.086 (UUU) 13 C27H35N8O22P3 916.108 (CCU-H2O1) 14 C27H36N7O24P3 935.102 (CUU) 15 C27H37N8O23P3 934.118 (CCU) 16 C28H34N9O22P3 941.103 (AUU-H2O1) 17 C28H34N9O23P3 957.098 (GUU-H2O1) 18 C28H35N10O21P3 940.119 (ACU-H2O1) 19 C28H35N10O22P3 956.114 (CGU-H2O1) 20 C28H36N9O23P3 959.114 (AUU) 21 C28H36N9O24P3 975.109 (GUU) 22 C28H37N10O22P3 958.13 (ACU) 23 C28H37N10O23P3 974.125 (CGU) 24 C29H35N12O20P3 964.13 (AAU-H2O1) 25 C29H35N12O21P3 980.125 (AGU-H2O1) 26 C29H35N12O22P3 996.12 (GGU-H2O1) 27 C29H37N12O21P3 982.141 (AAU) 28 C29H37N12O22P3 998.136 (AGU) 29 C29H37N12O23P3 1014.13 (GGU) 30 C9H11N2O8P1 306.025 (U-H2O1) 31 C9H13N2O9P1 324.036 (U) Finished generation of modification masses.	**Actual RNA fragments (including modifications) used to generate the precursor variants**
Theoretical precursor variants: 31	**Maximum number of precursor variants per spectrum considered**
Tandem spectra: 1080 8.41751% … 92.5926%	**Number of MS^2 spectra in the data**
RNPxl_sample_UV: Spectra filtered by fractional mass: 11	**Number of spectra filtered by the fractional mass filter**
RNPxl_sample_UV: Spectra filtered by peptide weight: 0	**Number of spectra filtered by peptide mass threshold**
RNPxl_sample_UV: Precursor variants filtered by m/z: 14351	**Number of precursor variants filtered by *m/z***

	threshold
RNPxl_sample_UV: Before filtering: 33480 theoretical precursor variants.	**Number of precursor variants before filtering for fractional mass, peptide weight and precursor *m/z* ratio**
RNPxl_sample_UV: After filtering: 18788 theoretical precursor variants.	**Number of precursor variants after filtering for fractional mass, peptide weight and precursor *m/z* ratio**
…	…
Cleaning up. Removed 1069 temporary mzML files and 1069 temporary idXML files.	**Diagnostic information**
RNPxl took 21:48 m (wall), 07:54 m (CPU), 7.60 s (system), 07:46 m (user).	**Total running time**
13:50:25 NOTICE: RNPxl finished!	**Diagnostic information**
13:50:25 NOTICE: Output file 'Q:\TOPPAS_out\003-RNPxl\RNPxl_sample_UV.csv' written.	**Position of the tabular output file.**
13:50:25 NOTICE: Output file 'Q:\TOPPAS_out\004-RNPxl\RNPxl_sample_UV.idXML' written.	**Position of the idXML output file.**
13:50:25 NOTICE: Entire pipeline execution finished!	**Diagnostic information**

5 Visualization of search results in TOPPView

TOPPView, the graphical viewer for mass spectrometry, data can be used to visualize and validate identification results.

Annotating spectra with identifications

Before identifications in idXML format can be visualized they have to be annotated to the experimental data. To this end,

- Select `File->Open file` from the top menu and select the .mzML file corresponding to the UV measured sample. Open it in 1D view.
  Alternatively, open the desired mzML file as any other file by double-clicking in Windows Explorer.
  It is advisable to open the file containing centroided data due to its considerably smaller size.
- Now select `Tools->Annotate with identification` and load the .idXML file generated by the RNP^{xl} tool.
- Click on the `Identification view` tab to visualize the identifications along with the spectra. Note that if the space for the identification view is too small to display all columns, the entire window can be dragged below the spectrum visualization (as done in Supplementary Note Figure 16).
- To display cross link annotations, select the `Show advanced annotations` checkbox.

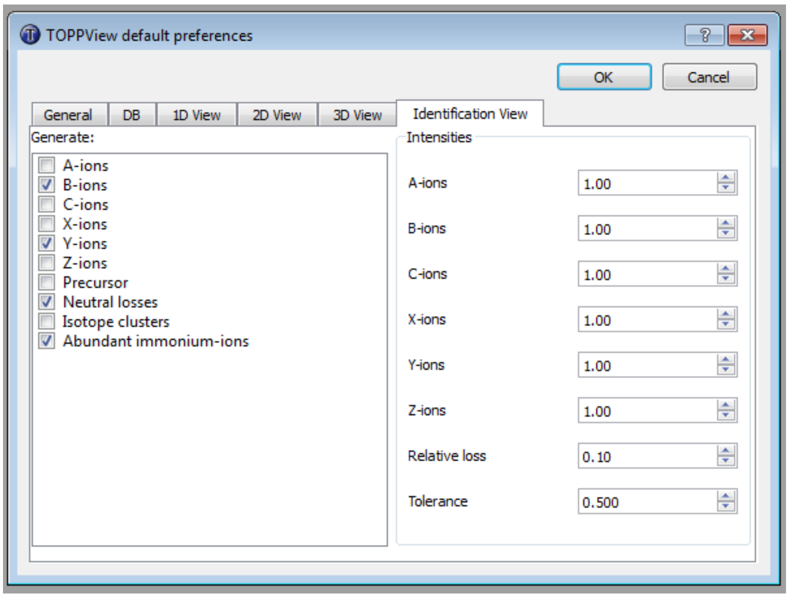
Left mouse click on the column headers sorts the identifications for charge, score, etc. Clicking on a cell in the `precursor m/z` column will display the corresponding MS^1 spectrum, the charge and the precursor isolation window. This can be used to manually control correct assignment of monoisotopic peak and charge state as well as cofragmentation of unrelated peaks. Clicking on a field in a different column displays the annotated spectrum.
Dragging over a certain area in a spectrum while pressing the control key allows zooming into the spectrum. Zooming in and out can also be done with the mouse wheel.

Configuring the identification view

The identification view allows annotating raw fragment ions with identified sequences. To use it effectively with the data produced by our workflow, some adaptation of default parameters has to be performed. The identification view parameters can be accessed as follows:

Open `File->Preferences` and select the `Identification View` tab

In the list to the left of the configuration dialog (see Supplementary Note Figure 15) the expected ion types can be selected. For Orbitrap HCD spectra, annotation of a-, b- and y-ions, neutral losses and abundant immonium ions is recommended. In the lower right, the tolerance used to annotate fragment mass ions can be specified in Thomson. We suggest 0.01-0.02 as default value for HCD fragmentation on an Orbitrap Velos or Q Exactive. The intensity scaling (“Intensities”) of theoretical fragments can be ignored.



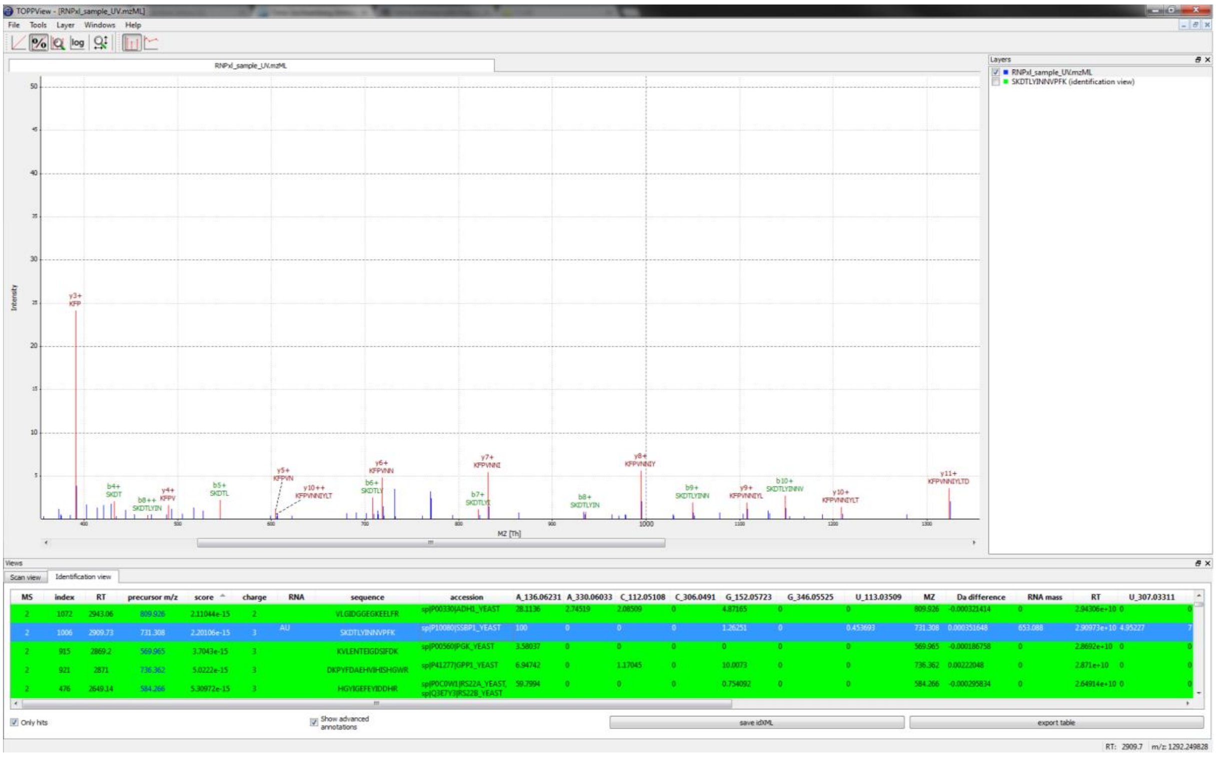
Generate:

- ☐ A-ions
- ☑ B-ions
- ☐ C-ions
- ☐ X-ions
- ☑ Y-ions
- ☐ Z-ions
- ☐ Precursor
- ☑ Neutral losses
- ☐ Isotope clusters
- ☑ Abundant immonium-ions

Intensities

A-ions	1.00
B-ions	1.00
C-ions	1.00
X-ions	1.00
Y-ions	1.00
Z-ions	1.00
Relative loss	0.10
Tolerance	0.500

Supplementary Note Figure 15: Identification view configuration dialog



MS	index	RT	precursor m/z	score	charge	RNA	sequence	accession	A_136.06231	A_330.06033	C_112.05108	C_306.0491	G_152.05723	G_346.05525	U_113.03509	MZ	Da difference	RNA mass	RT	U_307.03311	
2	1072	2943.06	809.926	2.11044e-15	2		VLGIDGGEGKEELFR	sp|P00330|ADH1_YEAST	28.1136	2.74519	2.08509	0	4.87165	0	0	809.926	-0.000321414	0	2.94306e+10	0	0
2	1006	2909.73	731.308	2.20106e-15	3	AU	SKDTLYINNVPFK	sp|P10080|SSBP1_YEAST	100	0	0	0	1.26251	0	0.453693	731.308	0.000351648	653.088	2.90973e+10	4.95227	7
2	915	2869.2	569.965	3.7043e-15	3		KVLENTEIGDSIFDK	sp|P00560|PGK_YEAST	3.58037	0	0	0	0	0	0	569.965	-0.000186758	0	2.8692e+10	0	0
2	921	2871	736.362	5.0222e-15	3		DKPYFDAEHVIHISHGWR	sp|P41277|GPP1_YEAST	6.94742	0	1.17045	0	10.0073	0	0	736.362	0.00222048	0	2.871e+10	0	0
2	476	2649.14	584.266	5.30972e-15	3		HGYIGEFEYIDDHR	sp|P0C0W1|RS22A_YEAST, sp|Q3E7Y3|RS22B_YEAST	59.7994	0	0	0	0.754092	0	0	584.266	-0.000295834	0	2.64914e+10	0	0

☑ Only hits

☑ Show advanced annotations

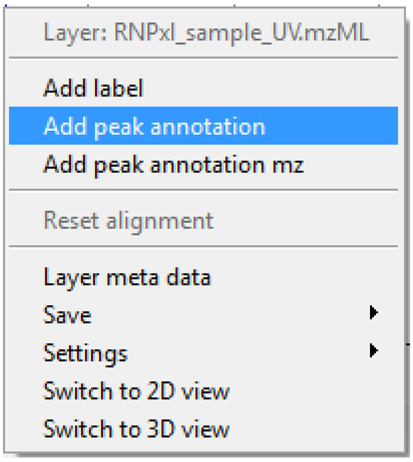
Supplementary Note Figure 16: TOPPView showing an annotated spectrum and the table of identified sequences.

Editing peak labels

Experimental m/z values can be annotated by moving the mouse on the desired peak (red cross on peak tip appears), left clicking and selecting `Add peak annotation mz`.

To annotate a peak with a label, e.g. a manual annotation, hover over it, left click and select `Add peak annotation` from the context menu (see Supplementary Note Figure 17). Labels can be moved and re-edited by left- or right-clicking, respectively.

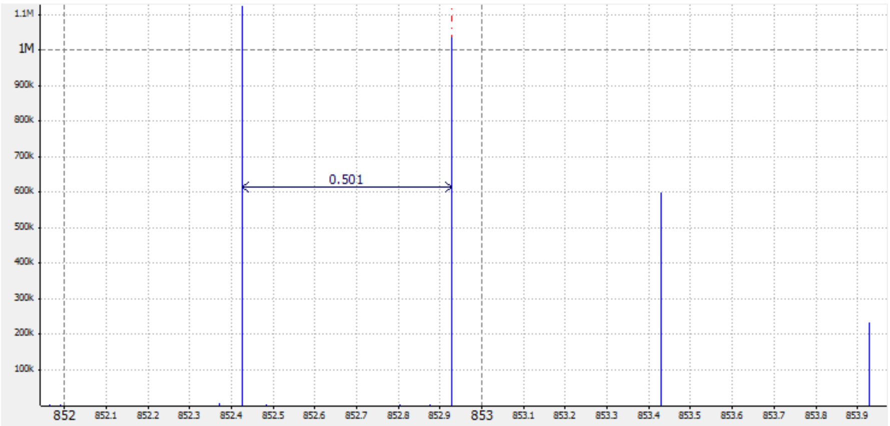
It has to be noted that labels cannot be saved, so all annotations manually added will be deleted when TOPPView is closed. Therefore it is advisable to fully annotate a spectrum and export it as image (see below).



Supplementary Note Figure 17: Context menu

Measuring peak distances

Holding the shift key and dragging (left mouse button pressed) from one peak to another allows measuring and annotating peak distances (see Supplementary Note Figure 18). This function is extremely useful for manual validation or spectra interpretation.



Supplementary Note Figure 18: Annotated peak distance

Manual validation of cross-links

Initial evaluation of cross-links is typically started with candidates that have a good score (which corresponds to a low number in the score column). Therefore, the results can be sorted by score and the first cross-links (having an entry in the RNA column) can be evaluated.

Candidates with a good score (< 10^{-10}) have a high number of matches and are typically validated easily. Nonetheless, XIC comparison (step 2) should be performed and the presence of RNA marker ions and RNA adducts (step 4) should be noted.

Candidates with intermediate score (between 10^{-9} and 10^{-5}) are increasingly likely to be false positives and have to be evaluated carefully.

Candidates with a bad score (> 10^{-5}) are very likely to be false positive matches. If evaluated at all, these matches have to be validated very carefully.

Since score distributions depend on many factors (sample amount and complexity, cross-linking yield, instrument type, fragmentation mode, database size, protein/RNA modifications…) the above mentioned classification into good, intermediate, and bad scores can be different for each experiment.

Validation follows the following steps:

1. Rough assessment of the fragment spectrum and the quality of the match to the search result.
   - Evaluation of spectrum quality.
     Spectra with low quality (low number of signals and low signal-to-noise ratio) are unlikely to allow unbiased cross-link identification. These spectra are not evaluated in detail but regarded as false positives.
   - Matches between experimental signals and theoretical peptide fragments (automatically annotated and color-coded by TOPPView).
     If the majority of signals are annotated as peptide fragments, validation is straightforward, which is the case for high scoring cross-link candidates.
     If a low number unassigned medium to high intensity signals are observed, they should be manually assigned as internal ions, RNA fragments or adducts of peptide and RNA (fragments) if possible, see below.
     If a high number of unassigned signals is observed, manual validation should only be attempted after XIC comparison (step 2) and evaluation of the precursor (step 3) have not identified the potential cross-link as a false positive.
2. Comparison of extracted ion chromatograms (XICs) in UV irradiated and control samples.
   - only if appropriate control is available
   - unfortunately not possible in TOPPView but only in the respective vendor program
   - cross-link candidate should have a significantly higher intensity/XIC peak area in the UV irradiated sample compared to the control, otherwise it can be regarded as a false positive hit
3. Evaluation of precursor
   - correct assignment of monoisotopic peak and charge state should be confirmed
   - co-fragmenting peaks should be noted as they can explain the presence of unassigned fragments in the MS2 spectrum
4. manual assignment of high intensity peaks not corresponding to peptide fragments
   - RNA marker ions should be compared to the candidate RNA sequence.

If more than one nucleotide is contained in the RNA sequence, high intensity signals are expected for adenine (m/z 136.0623), cytosine (112.0511) and guanine (152.0572) as well as adenosine and cytidine minus water (m/z 306.0491 and 330.0603). Uracil marker ions (113.0351 and 307.0331) are typically much less intense if observed at all.

- Additional peptide fragments, e.g. internal ions, can be calculated with online tools like ProteinProspector.
- Remaining unassigned peaks could correspond to adducts of peptide and RNA fragments. Distances to assigned peaks might help in assignment. If a series of signals in the higher m/z range is observed, it might be a good idea to check if distances between peaks correspond to amino acid masses. These can then be compared to the assigned peptide sequence to identify the peptide fragments. RNA adducts are then identified by calculating the mass difference between observed experimental m/z and calculated peptide fragment m/z.

  If, however, an amino acid sequence is identified that does not appear in the sequence reported, the hit might have to be discarded as a false positive result.

In all validation steps, parameters used for precursor variant generation and database search should be kept in mind. As described above, certain combinations of RNA combinations or modifications alone or together with posttranslational modifications defined for the database search can introduce ambiguity. For example, if a cross-link to GU is identified but an A marker ion is observed, it is possible that the database search overlooked an oxidation on the peptide (either incorrectly or because it was not chosen as a PTM) and the cross-link has to be manually assigned to the same but oxidized peptide to AU. Fortunately, these ambiguities are usually very rare if modifications are chosen appropriately and are easily spotted with some experience.

Once a cross-link has been validated, it might be desirable to check whether there are additional hits for the same peptide. By first clicking the score column, then the RNA column, and finally the peptide column, all cross-links for the same peptide will appear together. Equally, all cross-links of this peptide to the same RNA will be together, with the best scoring spectrum on top. This allows quick comparison of several spectra from the same cross-link or cross-links of the same peptide to different RNAs.

No manual remarks can be added to the identification view table. These can be added to the csv file which contains the same information and can be opened by appropriate programs like Microsoft Excel. For example, validated or false positive results can be color coded. In addition, results can be filtered, e.g. creating a separate table for all cross-link candidates by deleting all lines that do not have an entry in the RNA column.

There are various strategies imaginable for the validation of search results. The points described above represent the validation workflow that was followed for the presented data. Basic understanding of peptide fragmentation under beam-type CID/HCD conditions is required for complete and fast validation. As starting point on general fragmentation rules we recommend Michalski et al. [12] and references therein.

Exporting images

To export images select `Save->As image` from the context menu. Images can be exported as raster image (*.png, *.jpg) as well as vector image (*.svg).

6 Useful tips

Determine the number of MS^2 spectra in a measurement

The FileInfo tool prints information on an experiment including the number of MS^1 and MS^2 spectra. Open the command line window; go to the folder with the raw data in mzML format and type:

```
FileInfo –in filename.mzML
```

7 References

1. Sturm, M. et al. OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* **9**, 163 (2008).
2. Orchard, S., Hermjakob, H. & Apweiler, R. The proteomics standards initiative. *Proteomics* **3**, 1374-1376 (2003).
3. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24**, 2534-2536 (2008).
4. Geer, L.Y. et al. Open mass spectrometry search algorithm. *Journal of proteome research* **3**, 958-964 (2004).
5. Junker, J. et al. TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data. *Journal of proteome research* **11**, 3914-3920 (2012).
6. Sturm, M. & Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *Journal of proteome research* **8**, 3760-3763 (2009).
7. Martens, L. et al. mzML--a community standard for mass spectrometry data. *Molecular & cellular proteomics : MCP* **10**, R110 000133 (2011).
8. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology* **26**, 1367-1372 (2008).
9. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).
10. Camacho, C. et al. BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).
11. Kramer, K. et al. Mass-spectrometric analysis of proteins cross-linked to 4-thio-uracil- and 5-bromo-uracil-substituted RNA. *Int J Mass Spectrom* **304**, 184-194 (2011).
12. Michalski, A., Neuhauser, N., Cox, J. & Mann, M. A systematic investigation into the nature of tryptic HCD spectra. *Journal of proteome research* **11**, 5479-5491 (2012).