Aug 31, 2022

Proteoform Identification and Quantitation with TopPIC and TDPortal for Human Tissues

James M Fulcher[1], Yen-Chen Liao[2], Mowei Zhou[2], Ljiljana.PasaTolic[2]

[1]Pacific Northwest National lab; [2]Pacific Northwest National Laboratory

In Development Share

dx.doi.org/10.17504/protocols.io.3byl4bpj2vo5/v1

Human BioMolecular Atlas Program (HuBMAP) Method Development Community

PNNL-TTD

Mowei Zhou
Pacific Northwest National Laboratory

ABSTRACT

This protocol describe a workflow for top-down proteomics analysis. Top-down proteomics data are processed with two separate software packages TopPIC and TDPortal. Proteoform identifications were merged from the two software with unified FDR to increase coverage. TopPICR was separately used to cluster TopPIC proteoform to extract abundances for label-free quantitation.

DOI

dx.doi.org/10.17504/protocols.io.3byl4bpj2vo5/v1

PROTOCOL CITATION

James M Fulcher, Yen-Chen Liao, Mowei Zhou, Ljiljana.PasaTolic 2022. Proteoform Identification and Quantitation with TopPIC and TDPortal for Human Tissues. **protocols.io** https://protocols.io/view/proteoform-identification-and-quantitation-with-to-ceqztdx6

FUNDERS ACKNOWLEDGEMENT

National Institutes of Health (NIH) Common Fund, Human Biomolecular Atlas Program (HuBMAP)
Grant ID: UG3CA256959-01

CREATED

Aug 02, 2022

LAST MODIFIED

Aug 31, 2022

PROTOCOL INTEGER ID

68089

PARENT PROTOCOLS

In steps of

Overall protocol for MicroPOTS LCMS top down proteomics of kidney tissue sections

Cited in

Top Down Proteomics Data Collection for Microdissected Kidney Tissue Functional Units

Top Down Proteomics Data Collection for Microdissected Pancreas Tissue Functional Units

TopPIC Processing

1 Convert Instrument raw data to mzML using MSConvert

MSConvert

2 Analyze mzML files using the TopPIC Suite (version 1.4.13.1) .

TopPIC Suite 1.4.13.1

source by Xiaowen Liu

2.1 TopFD Parameters---------------------------------

Citation: James M Fulcher, Yen-Chen Liao, Mowei Zhou, Ljiljana.PasaTolic Proteoform Identification and Quantitation with TopPIC and TDPortal for Human Tissues https://dx.doi.org/10.17504/protocols.io.3byl4bpj2vo5/v1

```
Spectral data type:                Centroid
Maximum charge:                     30
Maximum monoisotopic mass:                50000 Dalton
Peak error tolerance:                0.02 m/z
MS1 signal/noise ratio:               3
MS/MS signal/noise ratio:               1
Thread number:                   10
Precursor window size:                2 m/z
Use Env CNN model:                 No
Miss MS1 spectra:                No
Generate Html files:              Yes
Do final filtering:             Yes
```

2.2 TopPIC 1.4.13 Parameters----------------------------------

```
********************** Parameters **********************
Protein database file:     ID_008032_8627C6BD.fasta.zip
Spectrum file:                xxxxxxxxxxxxxxxxx_ms2.msalign
Number of combined spectra:              1
Fragmentation method:              FILE
Search type:                 TARGET
Fixed modifications:               None
Use TopFD feature file:              True
Maximum number of unexpected modifications:  1
Error tolerance for matching masses:       15 ppm
Error tolerance for identifying PrSM clusters:  0.8 Da
Spectrum-level cutoff type:             EVALUE
Spectrum-level cutoff value:             0.05
Proteoform-level cutoff type:             EVALUE
Proteoform-level cutoff value:            0.05
Allowed N-terminal forms:
NONE,NME,NME_ACETYLATION,M_ACETYLATION
Maximum mass shift of modifications:       275 Da
Minimum mass shift of modifications:       -150 Da
Thread number:                  14
E-value computation:               Generating function
Common modification file name:          Dynamic_mods.txt
MIScore threshold:                0.15
Executable file directory:
Version:                  1.4.13
```

The protein fasta contains human proteome from UniProt with both SwissProt and TREMBL sequences. Decoy sequences were added as well. Unzip the attachment to use it.

3 TopPIC outputs proteoform spectrum matches (PrSMs) as tab-separated files (...toppic_prsm.tsv) and quantification data within MS1 feature files (..._ms1.feature). These are both imported into the R environment for post-processing with TopPICR.

4 TopPICR is used for post-processing to improve proteoform identification and quantification. All functions are documented within the TopPICR R package.

TopPICR
source by Evan Martin

4.1 First, result files are read into R using the read_toppic(file_path = path, file_name = names) function in TopPICR, where the "path" is the path to the directory containing the TopPIC PrSM files and "names" is a character vector specifying the PrSM files to import. This function can also be utililzed to import the MS1 feature files into a separate object.

4.2 Next, the data is further processed with the augment_annotation() and rm_false_gene() functions to account for ambiguity in proteoform identifications

4.3 False discovery rate (FDR) filtering is accomplished by finding the appropriate E-value cutoff to filter the results to 1% FDR at the isoform and protein level. This is provided by the find_evalue_cutoff() and apply_evalue_cutoff() functions.

4.4 Proteoform inference is performed with infer_pf() function and the proteoform level is determined with set_pf_level() function .

Smith LM, Thomas PM, Shortreed MR, Schaffer LV, Fellers RT, LeDuc RD, Tucholski T, Ge Y, Agar JN, Anderson LC, Chamot-Rooke J, Gault J, Loo JA, Paša-Tolić L, Robinson CV, Schlüter H, Tsybin YO, Vilaseca M, Vizcaíno JA, Danis PO, Kelleher NL (2019). A five-level classification system for proteoform identifications.. Nature methods.
https://doi.org/10.1038/s41592-019-0573-x

Citation: James M Fulcher, Yen-Chen Liao, Mowei Zhou, Ljiljana.PasaTolic Proteoform Identification and Quantitation with TopPIC and TDPortal for Human Tissues https://dx.doi.org/10.17504/protocols.io.3byl4bpj2vo5/v1

4.5 Retention time alignment is processed with the form_model() and align_rt() functions.

4.6 Mass calibration is accomplished with the calc_error() and recalibrate_mass() functions

4.7 Clustering and deisotoping error correction is performed with the cluster() and create_pcg() functions.

4.8 Metadata for each proteoform cluster is generated with the create_mdata() function.

4.9 Steps 4.5 and 4.6 are applied to the MS1 feature files as well before features are matched and combined (for MBR) with the match_features() and combine_features() functions.

5 The final table of proteoform identification and quantitation results from TopPIC Suite and TopPICR are exported as comma-separated value (.csv) files.

TDPortal Processing

6 Request TDPortal access and follow their instructions to set up an account.

TDPortal
by Northwestern University

TDPortal search process
6.1 Upload data
6.2 Search on TDPortal

TDportal has an option for label-free quantitation, but it is not used in this workflow.

6.1 Upload data
1. Connect to Northwestern through VPN.

(https://kb.northwestern.edu/page.php?id=94726)

2. Copy the files to your user folder. (Eg. \\resfiles.northwestern.edu\NU-PCEDATA\external_users\XXXXX)
3. The system will ask you to log in. Please use "ads\your id" with your password to log into your folder.
4. Create a sub-folder under your user folder with each search.
5. Put raw files to the sub-folder accordingly and do not have more folders under the sub-folder. (https://kb.northwestern.edu/page.php?id=70525).

6.2 Search on TDPortal (https://portal.nrtdp.northwestern.edu/static/TDPortalSOP_043_20180301.pdf)

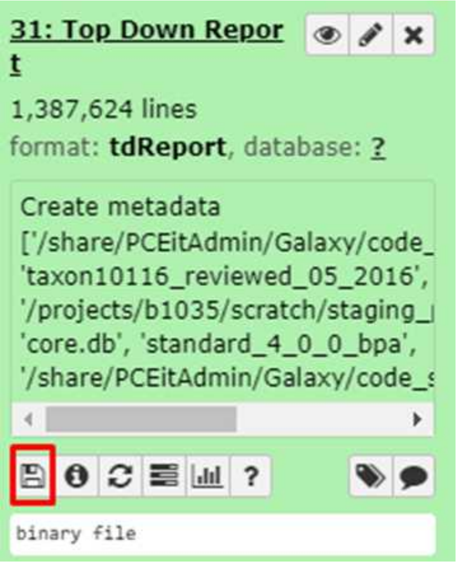
1. Connect website(https://portal.nrtdp.northwestern.edu/)
2. Log in TDPortal with “your email address” and “your password”
3. Your subfolder's name will show as each dataset.
4. Select files into the “Input files” under the selected dataset.
5. Select organism "human".
6. Set parameters as follow:

User empirical P-score: False
Filter by FDR: True
Create SAS input sheet for quant: Select True when we need to.
Precursor resolution: High resolution
Fragmentation Type: Auto (or the type we used on MS).
Code set: Standard 4.0.0
Include ProSight Error Tolerance Search: False (select “true” when we want to allow one unknown mass shift in the proteoform).

7 Exporting TDPortal results

TDViewer
by Northwestern University

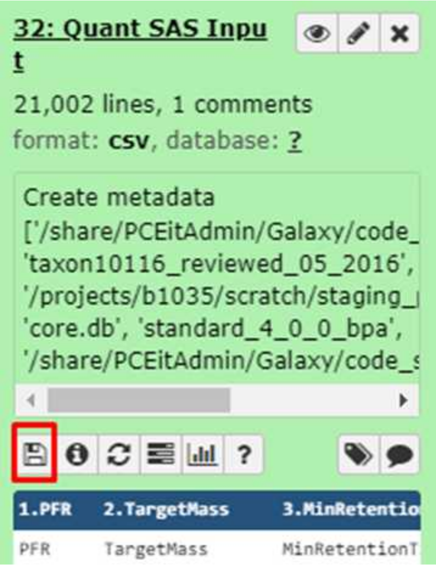
1. Download *.tdReports file. Note: There can be two separate processes created in the queue. One for ID results in the TDReport. Another is the CSV file for quantitation (if enabled).
2. Click the download icon to download these files.



```
31: Top Down Repor
t
1,387,624 lines
format: tdReport, database: ?

Create metadata
['/share/PCEitAdmin/Galaxy/code_
'taxon10116_reviewed_05_2016',
'/projects/b1035/scratch/staging_
'core.db', 'standard_4_0_0_bpa',
'/share/PCEitAdmin/Galaxy/code_s

binary file
```



```
32: Quant SAS Inpu
t
21,002 lines, 1 comments
format: csv, database: ?

Create metadata
['/share/PCEitAdmin/Galaxy/code_
'taxon10116_reviewed_05_2016',
'/projects/b1035/scratch/staging_
'core.db', 'standard_4_0_0_bpa',
'/share/PCEitAdmin/Galaxy/code_s

1.PFR    2.TargetMass     3.MinRetentio
PFR      TargetMass       MinRetentionT
```

3. Open TDReport by TDViewer 2.0(http://tdviewer2.northwestern.edu/)
4. Read and export proteoform ID results from TDViewer with 1% FDR cutoff.

Combining Results

8 Results (proteoform spectral matches) from TopPIC and TDPortal are then merged using a function written in R that is openly available on GitHub. The input proteoform tables from each software was pre-filtered with FDR cutoff of 1% (adjusted FDR in TopPICR for TopPIC, and the default FDR in TDPortal).

TDPortal_TopPIC_Join
source by James M Fulcher

Final output

9 Results for proteoform spectral matches (merged from TopPIC and TDPortal) and proteoform quantitation (TopPICR) are uploaded to HIVE.

Citation: James M Fulcher, Yen-Chen Liao, Mowei Zhou, Ljiljana.PasaTolic Proteoform Identification and Quantitation with TopPIC and TDPortal for Human Tissues https://dx.doi.org/10.17504/protocols.io.3byl4bpj2vo5/v1