Journal of Visualized Experiments

Mass spectrometry-based proteomics analyses using OpenProt database to unveil novel proteins translated from non-canonical open reading frames

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE59589R1
Full Title:	Mass spectrometry-based proteomics analyses using OpenProt database to unveil novel proteins translated from non-canonical open reading frames
Keywords:	OpenProt; ORF; alternative ORF; altORF; alternative protein; altProt; proteomics; mass spectrometry; translation; protein-coding gene; non-coding gene; pseudogene
Corresponding Author:	Marie Brunet Universite de Sherbrooke Sherbrooke, CANADA
Corresponding Author's Institution:	Universite de Sherbrooke
Corresponding Author E-Mail:	Marie.Brunet@usherbrooke.ca
Order of Authors:	Marie A. Brunet
	Xavier Roucou
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Open Access (US$4,200)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Sherbrooke, Quebec, Canada

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

To Philip Steindel, PhD
Review Editor
JoVE
1 Alewife Center | Suite 200 | Cambridge | MA | 02140
Phone: 617-401-7677 Ext: 850 | Fax: 866.381.2236

Cover Letter: Mass spectrometry-based proteomics analyses using OpenProt database to unveil novel proteins translated from non-canonical open reading frames

Dear Dr Steindel,

Following initial revision of our manuscript detailing the use of the OpenProt database for mass spectrometry-based proteomics, please find attached a revised version.
Thank you for the opportunity to submit this improved version and we are looking forward to hearing from you,

Regards,

Dr Marie A. Brunet and Dr Xavier Roucou

Manuscript

Click here to access/download;Manuscript;59589_R1_012819.docx

TITLE:

Mass Spectrometry-Based Proteomics Analyses Using the Openprot Database to Unveil Novel Proteins Translated from Non-Canonical Open Reading Frames

AUTHORS AND AFFILIATIONS:

Marie A. Brunet[1,2], Xavier Roucou[1,2]

[1]Department of Biochemistry, Université de Sherbrooke, Sherbrooke, Québec, Canada;
[2]PROTEO, Quebec Network for Research on Protein Function, Structure, and Engineering;

Corresponding author:

Marie A. Brunet (Marie.brunet@usherbrooke.ca)

Email Address of Co-Author:

Xavier Roucou (Xavier.roucou@usherbrooke.ca)

KEYWORDS:

OpenProt, ORF, alternative ORF, altORF, alternative protein, altProt, proteomics, mass spectrometry, translation, protein-coding gene, non-coding gene, pseudogene

SUMMARY:

OpenProt is a freely accessible database that enforces a polycistronic model of eukaryotic genomes. Here, we present a protocol for the use of OpenProt databases when interrogating mass spectrometry datasets. Using OpenProt database for analysis of proteomic experiments allows for discovery of novel and previously undetectable proteins.

ABSTRACT:

Genome annotation is central to today's proteomic research as it draws the outlines of the proteomic landscape. Traditional models of open reading frame (ORF) annotation impose two arbitrary criteria: a minimum length of 100 codons and a single ORF per transcript. However, a growing number of studies report expression of proteins from allegedly non-coding regions, challenging the accuracy of current genome annotations. These novel proteins were found encoded either within non-coding RNAs, 5' or 3' untranslated regions (UTRs) of mRNAs, or overlapping a known coding sequence (CDS) in an alternative ORF. OpenProt is the first database that enforces a polycistronic model for eukaryotic genomes, allowing annotation of multiple ORFs per transcript. OpenProt is freely accessible and offers custom downloads of protein sequences across 10 species. Using OpenProt database for proteomic experiments enables novel proteins discovery and highlights the polycistronic nature of eukaryotic genes. The size of OpenProt database (all predicted proteins) is substantial and need be taken in account for the analysis. However, with appropriate false discovery rate (FDR) settings or the use of a restricted OpenProt database, users will gain a more realistic view of the proteomic landscape. Overall, OpenProt is a freely available tool that will foster proteomic discoveries.

INTRODUCTION:

Over the past decades, mass spectrometry (MS-)based proteomics has become the golden technique to decipher proteomes of eukaryotic cells[1–5]. This method relies on current genome annotations to generate a reference protein sequence database that outlines the scope of possibilities[6–8]. However, genome annotations hold arbitrary criteria for ORF annotation, such as a minimum length of 100 codons and a single ORF per transcript[9,10]. An increasing number of studies challenge the current annotation model and report discoveries of unannotated functional ORFs in eukaryotic genomes[8,11–14]. These novel proteins are found encoded in allegedly non-coding RNAs, in the 5' or 3' untranslated regions (UTR) of mRNAs, or overlapping the canonical coding sequence (cCDS) in an alternative frame. Although most of these discoveries have been serendipitous, they demonstrate the caveats of current genome annotations and the polycistronic nature of eukaryotic genes[8].

Here, we highlight the use of OpenProt databases for MS-based proteomics. OpenProt is the first database to hold a polycistronic annotation model for eukaryotic transcriptomes. It is freely available at www.openprot.org. It currently contains transcriptome annotations for 10 species and reports every possible ORF longer than 30 codons[15]. A proportion of these predicted ORFs would be random and non-functional, which is why OpenProt cumulates experimental and functional evidence to increase confidence. Experimental evidence include protein expression (by MS) and translation evidence (by ribosome profiling)[15]. Functional evidence include protein orthology (with an In-Paranoid like approach) and functional domain prediction[15].

OpenProt offers the possibility to download several databases, from containing only well-supported proteins to custom-made databases. Here, we will present a pipeline for the use of OpenProt databases and will offer insights into which database to choose considering the experimental aim. The proteomics analysis pipeline presented here is supported by the Galaxy framework as it is open-access and easy-to-use, but the databases can work with any workflow[16–18]. We will also present how to use the OpenProt website for gathering further information on novel proteins detected by MS. Using OpenProt databases will provide a more exhaustive view of the proteomic landscape and will foster proteomics and biomarkers discoveries in a more systematic way than current methods.

This protocol highlights the use of OpenProt databases[15] when interrogating MS datasets; it will not review the design of the experiment itself, which has been thoroughly reviewed elsewhere[20–22]. In an effort to remain fully open-source, the protocol is freely available (**Supplementary Material S1–S4**). For easier reading, all terms used in OpenProt and hereby throughout this protocol are defined in **Table 1**.

PROTOCOL:

1. OpenProt database download

NOTE: Custom databases based on RNA-seq data for example can also be obtained and the procedure is detailed in the second section of this protocol. If a custom database is needed, please skip to the next section.

1.1. Go to the OpenProt website: www.openprot.org and open the Downloads page using the link from the top page menu.

1.2. Click on the species of interest based on the analyzed experimental data.

1.3. Click on the protein type desired.

NOTE: OpenProt offers three classifications: RefProt, Isoforms and AltProt. As shown in **Figure 1**, this parameter will vary based on the research objective.

1.3.1. Click on **RefProt alone** to generate files containing only known proteins.

1.3.2. Click on **AltProt and Isoforms** to generate files containing only novel proteins – either novel isoforms of known proteins (Isoforms) or coded by an alternative ORF (AltProts). Please note that OpenProt enforces a minimum ORF length of 30 codons[15].

1.3.3. Click on **AltProts, Isoforms and RefProts** to generate files containing all protein types present in the OpenProt database – known and novel proteins.

1.4. If available, click on the annotation from which protein sequences are drawn.

NOTE: OpenProt offers a more exhaustive proteomic landscape by combining multiple annotations. Transcriptome annotations have a minimal overlap; thus, the selected annotation can substantially affect the visualized proteomic profile[15, 23].

1.5. Click on the level of supporting evidence necessary for protein consideration. As shown in **Figure 1**, this parameter will vary based on the research objective.

1.5.1. Click on **minimum of two unique peptides detected** to generate files containing only the most confident proteins.

NOTE: A criterion of two unique peptides is currently considered a gold standard in proteomics for protein expression. If the experimental aim is to detect known and well-supported proteins, the use of this parameter is recommended.

1.5.2. Click on **minimum of one unique peptides detected** to generate files containing proteins that have already been seen at least once among the mass spectrometry experiments re-analyzed by OpenProt.

NOTE: This allows for consideration of the shorter length of AltProts and the probability that some of them may contain only one unique tryptic peptide[8,11].

1.5.3. Click on **all predicted** to generate files containing all of OpenProt predictions.

NOTE: This setting is recommended only if the experimental aim is to discover novel proteins (**Figure 1**). The subsequent substantial increase in the search space calls for an adapted analysis pipeline as discussed below[7,15].

1.6. Click on the desired file format to download. For proteomic analyses, choose the Fasta (protein) file. The readme file contains all necessary information on the file format.

2. Custom OpenProt database download

NOTE: This section details how to obtain a custom database. If no custom database is needed, skip to the next section.

2.1. Go to the OpenProt website (www.openprot.org) and open the Search page using the link from the top page menu.

2.2. Click on the species of interest based on the experimental data analyzed.

2.3. Enter a list of genes or transcripts of interest.

2.3.1. When using a list of genes, enter it in the **Gene** query box.

2.3.2. When using a list of transcripts, enter it in the **Transcript** query box.

2.4. Tick any box that applies to the desired database.

2.4.1. Do not click on any box to obtain a table containing all types of protein supported by OpenProt: RefProt, Isoforms and AltProts.

2.4.2. Click on **Show only proteins with experimental evidence** to obtain a table containing all types of proteins (RefProts, Isoforms and AltProts) that have been detected at least once by MS and/or for which translation evidence has been collected from ribosome profiling data.

2.4.3. Similarly, click on **Show only proteins detected by MS** or on **Show only proteins detected by ribosome profiling** to obtain a table containing all types of proteins that have been detected at least once by MS or by ribosome profiling respectively.

2.4.4. Click on **Show only AltProts** or on **Show only isoforms** to obtain a table containing only AltProts or only Isoforms respectively.

2.4.5. Click on both **Show only AltProts** and **Show only Isoforms** to obtain a table containing both types of proteins.

NOTE: All combinations of filters are possible.

2.5. Once all desired parameters are set, click on Search. The table output will appear below the search query fields.

2.6. Click on the **Download Fasta** button at the right top corner of the output table. This will generate a Fasta file containing all proteins resulting from the queried list of genes or transcripts.

2.7. Please note that for computational reasons, OpenProt holds a maximum of 2,000 elements to be queried (genes or transcripts) at a time. In the event of a list above that limit, several fasta can be generated and then concatenated (as detailed below); or simply download the whole OpenProt database and filter the obtained file as desired.

2.7.1. Bin the whole list of genes or transcripts into sub-lists of 2,000 entries or less. For each sub-list, download a Fasta file as described above (step 3.3 to 3.6).

2.7.2. Log in to the European Galaxy instance (or any other instance where proteomics tools are available), https://usegalaxy.eu/.

2.7.3. Create a new history and import all of the downloaded OpenProt databases (one per sub-list of genes or transcripts) by clicking on the upload logo at the left top of the screen.

2.7.4. Use the **Fasta Merge Files and Filter Unique Sequences** tool developed by the GalaxyP developers (https://github.com/galaxyproteomics/). Select the **Merge all Fasta** option and input all of the imported OpenProt databases.

NOTE: Each tool can be searched by using the query box on the left side of the screen

2.7.5. Select the **accession only** option to assess sequence unicity and copy the OpenProt identifier parse rule (**>(.*)\|**), then click on **Execute**.

2.7.6. Note that all files have been concatenated into a unique Fasta file with no redundancy that now appears in the history panel on the right side of the screen. This constitutes the working database.

3. Database handling

NOTE: From now on, the Galaxy platform will be used, but the same principles can be applied to other proteomic software.

3.1. Log in to the European Galaxy instance (or any other instance where proteomics tools are available), https://usegalaxy.eu/.

3.2. Create a new history and import the downloaded OpenProt database by clicking on the upload logo at the left top of the screen.

3.3. Go to the workflow page and import the Database Handling workflow (**Supplementary Material S1**) by clicking on the upload logo at the left top of the middle panel.

3.4. Click on **Run the workflow** and select the imported OpenProt database as input.

NOTE: This workflow will append the CRAPome repository to the OpenProt fasta and generate decoy sequences (reverse sequences)[24]. If a shuffle decoy list is desired, it can be done by changing this parameter on the DecoyDatabase tool.

3.5. Rename the obtained Fasta file to something meaningful. The database is ready to be used for proteomics analyses.

4. Mass spectrometry file preparation

NOTE: Most of the proteomics tools available on Galaxy instances use the mzML format, and peptide search engines prefer data in centroid mode.

4.1. Open the freely available MSConvert tool from the ProteoWizard suite and upload the data file to be analyzed[25].

4.2. Choose the directory for the output and the desired file format to mzML.

4.3. Set a peak picking filter using the wavelet based algorithm (CWT) on MS1 and MS2 levels, and start the conversion[26].

5. Peptide and protein identification/quantification

NOTE: This part of the pipeline uses tools from the OpenMS suite, a versatile and easy-to-use framework[18].

5.1. Log in to the European Galaxy instance (or any other instance where proteomics tools are available), https://usegalaxy.eu/.

5.2. Create a new history and transfer the previously created database (step 4.5) to this new history with a drag-and-drop.

5.3. Import the transformed mzML data file (step 5.3) by clicking on the **Upload** logo at the left top of the screen.

5.4. Go to the workflow page and import the desired workflow by clicking on the upload logo at the left top of the middle panel.

NOTE: MS experiments are differently designed based on the desired final output. Workflows are provided here for two frequent designs: protein identification and protein quantification based on stable isotope labeling (SIL). However, the Galaxy instance contains many other tools that will support other types of proteomic analyses[27,28].

5.4.1. For a protein identification design, import the workflow provided in **Supplementary Material S2**.

5.4.2. For a protein quantification based on stable isotope labeling design, import the workflow provided in **Supplementary Material S3**.

5.5. Select **run the workflow** and review the different parameters.

5.5.1. Select the imported mzML data file as input, and the previously created database (step 4.5) as the database Fasta file.

5.5.2. Since the workflow uses the X!Tandem search engine, import the X!Tandem default configuration file (provided in **Supplementary Material S4**)[29] by clicking on the upload logo at the left top of the screen.

5.5.3. The workflow uses multiple search engines (MS-GF+ and X!Tandem). Append other search engines or choose a single one simply by adding or removing the tools from the workflow[30,31].

NOTE: Using multiple search engines is recommended as it increases sensibility and sensitivity of the analysis[32].

5.5.4. In order to account for the substantial increase in size when using the whole OpenProt database, use a stringent FDR[15]. By default, the provided workflow is set for a 0.001% FDR, adequate for the use of the whole OpenProt database. For other databases, this can be edited to any desired value.

NOTE: Be sure to adapt the parameters of the different tools depending on the mass spectrometer used and the experimental protocol (precursor ion and fragment error, fixed and variable modifications, used enzyme, etc.).

5.6. Optionally, download output for each step of the workflow for storage or quality control analysis by clicking on the chosen step from the history panel, then clicking on the **Save** logo that will appear underneath.

6. Quality control

NOTE: Because MS-based proteomics is the result of a complex process where each step needs to be optimized to produce reproducible results, quality control is a necessary procedure in the workflow[33].

6.1. Several metrics are common benchmark of performance, such as the number of peptide-spectrum matches (PSM), the number of identified peptides and proteins. Run the **File Info** tool on the IDFilter output (indicated in green in **Figure 2**) to provide such metrics.

6.2. . Although not applicable to every identification, especially with large datasets, reports of novel proteins should always be carefully evaluated. Inspection of the protein score, the sequence coverage, and the spectra supporting the finding is of vital importance. Use the TOPPview tool from the OpenMS framework to do this; it is freely available and well documented[18,34,35].

7. OpenProt database mining

NOTE: Once a confident identification of a novel protein predicted by OpenProt (accession numbers starting with IP_ for AltProts and II_ for novel Isoforms) has been made, more biological information can be gathered from the OpenProt website[15].

7.1. Go to the OpenProt website: www.openprot.org and open the Search page using the link at the top page menu.

7.2. Click on the species of interest (same as the one in which the protein was identified) and enter the protein accession number in the **Protein** query box.

7.3. Click on search and a table containing basic information on the queried protein will appear. The table features: the protein length (in amino acid), its molecular weight (kDa) and isoelectric point, supporting experimental evidence by MS or ribosome profiling (Translation Evidence, TE), and functional predictions such as predicted domains and protein orthology (across the 10 species supported by OpenProt, v1.3). The table also contains information about the related gene and transcript and the localization of the protein within the transcript.

7.4. Click on the **Details** link to gather further information. The newly opened page contains a genome browser which is centered on the queried protein, and information such as the genomic and transcriptomic coordinates and the presence of a Kozak or high-efficiency translation initiation site (TIS) motif[36,37].

7.5. Click on the **Protein** or **DNA** links from the info tab, to obtain protein or DNA sequences respectively.

7.6. Browse detailed information about MS evidence, ribosome profiling detection, conservation and identified protein domains by clicking onto the top tabs[15].

REPRESENTATIVE RESULTS:

The workflow described above was applied to a MS dataset available on the PRIDE repository[38,39]. The original study developed a method (iMixPro), using stable isotope labeling of amino acids in cell culture (SILAC), to eliminate false positives from affinity-purification MS (AP-MS) experiments[38]. In brief, an AP-MS experiment consists of using beads-bound antibodies to fetch a protein of interest (bait) and its interactors (preys). The collected proteins are then digested and prepared for MS. The sample preparation method and the instrument settings are described in the original study and on the PRIDE repository (PXD004246). A challenge in such experiments is the abundance of false positives, notably from proteins binding to the beads but not the bait. Here, we used SILAC to generate different isotope ratios between true preys and false positives: 3 control samples (no bait) cultured in light medium, 1 sample expressing the bait cultured in light medium, and 1 sample expressing the bait cultured in heavy medium are processed with the beads and further mass spectrometry analysis. With such design, non-specific proteins binding to the beads will have an heavy-to-light ratio of 1:4; when true preys will have a ratio of 1:1[38].

We re-analyzed their AP-MS data using the OpenProt database; the baits included three endogenous proteins (PTPN14, JIP3 and IQGAP1), and two over-expressed proteins (RAF1 and RNF41). Since the experiments used SILAC, the Galaxy workflow for protein quantification was used (**Supplementary Material S3**, **Figure 2**). The workflow was run using the whole OpenProt database (OpenProt_all) or a restricted OpenProt database (OpenProt_2pep, including only proteins previously detected with a minimum of two unique peptides).

Protein identification and quantification were good and reproducible across the different used databases. As shown in **Figure 3**, most proteins identified in the original paper were also identified using either the OpenProt_2pep or OpenProt_all database (a detailed list is available in **Supplementary Material S5**). This result shows that the pipeline described here and the OpenProt databases are able to produce protein identification and quantification comparable to that of current procedures based on the UniProtKB databases[40]. However, the use of OpenProt databases has the unique advantage of allowing detection of novel and previously undetectable proteins, as demonstrated in this case study.

11 well-supported proteins (1 Isoform and 10 AltProts), yet currently not annotated in databases, were identified across all datasets, with confident peptides, using the OpenProt_2pep database (all protein accessions, along with the number of supporting peptides, are available in **Supplementary Material S5**). This database allows the use of a traditional 1% FDR as the search space increase remains moderate. These 11 proteins were not identified in the original study as they were absent from the database.

29 novel proteins (16 isoforms and 13 AltProts) were discovered across all datasets, with confident peptides, using the OpenProt_all database (all protein accessions, along with the number of supporting peptides, are available in **Supplementary Material S6**). As shown in **Figure 3**, the recommended stringent FDR did not affect the most confident protein identifications, although it did decrease the total number of identified proteins. Comparatively to the OpenProt_2pep database, a higher number of novel proteins can be confidently identified. All of these novel proteins are absent from the OpenProt_2pep database. This highlights the crucial role of the chosen database for MS-based proteomics.

One novel protein was discovered as an interactor of the RAF1 protein (IP_637643). Using the OpenProt website, one can see this protein had not been detected by MS nor ribosome profiling until now (OpenProt v1.3). The protein is 46 amino acids long and can only give two unique peptides upon tryptic digestion. The peptide detected in the RAF1 AP-MS dataset (fraction 18) had a good quality spectrum, as shown in **Figure 4**, and displayed a heavy-to-light ratio of 1,09. The protein is encoded in the *NANOGNBP1* gene, which is a pseudogene of *NANOGNB*. The transcript (ENST00000448444), currently annotated as non-coding, was detected across several tissues according to the GTEx portal[40]. The protein contains a predicted functional domain associated with DNA binding (Gene Ontology GO:0003677)[41].

FIGURE AND TABLE LEGENDS:

Figure 1: Database choice for proteomics analyses chart. Analyses of MS data, notably the database choice, depend on the research objectives. Three common objectives are outlined in blue (classic proteomic pipeline), green (exhaustive proteomic search) and orange (proteomic discovery). Each objective depends on an appropriate database and pipeline. A single identification tool may be used for an exhaustive and classical proteomics pipelines. For the proteomic discovery pipeline, we strongly recommend using multiple identification engines. Recommended FDRs are indicated in red, and protein database sizes are indicated in grey boxes.

Figure 2: Graphical representation of the Galaxy workflow used. Step-by-step representation of the proteomic analysis workflow used for re-analysis of Eyckerman et al. data[38]. Input files, peptide search, and protein quantification are indicated by orange boxes. Blue boxes correspond to the tools used and grey boxes correspond to the output files generated. The different search engines (MS-GF+ and X!Tandem) are indicated by different colors (respectively red and purple) as well as the arrows indicating their necessary inputs and outputs. The green box highlights the tool generating a list of protein identifications. When multiple outputs are generated, the one used for downstream steps is indicated as the closest to the arrow. This workflow is freely available in **Supplementary Material S2**. The X!Tandem default parameters configuration file is available in **Supplementary Material S4**.

Figure 3: Comparison of interactor identification per bait using different databases. Venn diagrams of protein identifications using the most confident OpenProt database (in orange, supporting evidence of minimum 2 unique peptides, OpenProt_2pep) with a 1% FDR, or the whole OpenProt database (in blue, OpenProt_all) with a 0.001% FDR, or as reported in the original paper (in grey)[38]. Each diagram corresponds to identified interactors for the mentioned bait: RAF1, RNF41, PTPN14, JIP3 and IQGAP1.

Figure 4: MS/MS spectrum of identified MDNLWAK$^{(13C6)}$ peptide from novel protein IP_637643. Intensity is relative (0 to 100%). Selected peaks are indicated in red, y ions annotations are in dark red and b ions annotations in green. Extracted from the TOPPview software[34]. Precursor Error = 2.70 ppm, PEP score = 0.12.

Table 1: Definition of terms used in OpenProt and throughout the protocol

Supplementary Material S1: Galaxy workflow for database handling. This will append the CRAPome and decoy sequences (reverse) to the input database. Output is a Fasta file.

Supplementary Material S2: Galaxy workflow for protein identification. This will identify proteins from a mass spectrometry data file using two search engines (MS-GF+ and X!Tandem). Each parameter can be tuned as desired before running the workflow.

Supplementary Material S3: Galaxy workflow for protein quantification using stable isotope labeling (SIL). This will identify and quantify proteins from a mass spectrometry data file using two search engines (MS-GF+ and X!Tandem). Each parameter can be tuned as desired before running the workflow.

Supplementary Material S4: X!Tandem default parameters configuration file. This XML file is necessary for running the X!TandemAdapter tool on the Galaxy platform.

Supplementary Material S5: Quantified proteins from iMixPro datasets. Data files from Eyckerman et al. 2016[38] were processed using OpenProt databases and quantified proteins are listed for each condition. Baits are PTPN14, JIP3, IQGAP1, RAF1 and RNF41. Gene names indicated in green correspond to proteins also identified in the original paper[38]. Gene names indicated in orange correspond to known interactors according to BioGrid that were not reported in the original paper. Gene names indicated in light blue correspond to novel proteins identified as interactors (the corresponding protein accession number is indicated in brackets). Gene names indicated in light grey and italics correspond to likely contaminants (keratin proteins).

Supplementary Material S6: Identified novel proteins from iMixPro datasets. Data files from Eyckerman et al. 2016[38] were processed using OpenProt databases and novel identified proteins are listed for each condition. Baits are PTPN14, JIP3, IQGAP1, RAF1 and RNF41. Protein accession numbers are listed, starting with II_ for novel isoforms of a known protein, and with IP_ for novel proteins from an alternative ORF (AltProt). The number of supporting peptides are indicated in brackets.

DISCUSSION:

When analyzing data from mass spectrometers, the quality of protein identification partly relies on the accuracy of the used database[6,20]. Current approaches traditionally use UniProtKB databases, yet these support the genome annotation model of a single ORF per transcript and a minimum length of 100 codons (with the exception of previously demonstrated examples)[40]. Multiple studies relate the shortcomings of such databases with the discovery of functional ORFs from allegedly non-coding regions[8,11–13]. Now, OpenProt allows for more exhaustive protein identification as it draws protein sequences from multiple transcriptome annotations. OpenProt retrieves NCBI RefSeq (GRCh38.p7) and Ensembl (GRCh38.83) transcriptomes and UniProtKB annotations (UniProtKB-SwissProt, 2017-09-27)[40,42,43]. As current annotations present little overlap, OpenProt thus displays a more exhaustive view of the potential proteomic landscape than when limited to one annotation[15].

Furthermore, as OpenProt enforces a polycistronic model, it allows for multiple protein annotations per transcript. For statistical and computational reasons, OpenProt still holds a minimum length threshold of 30 codons[15]. Yet, it predicts thousands of novel protein sequences, thereby widening the scope of possibilities for protein identification. With this approach, OpenProt supports proteomic discoveries in a more systematic manner.

The quality of protein identification can also be affected by the parameters that are used. MS-based proteomics analyses typically hold a 1% protein FDR. However, the whole OpenProt database contains about 6 times more entries (**Figure 1**). To account for this substantial increase in the search space, we recommend using a more stringent FDR of 0.001%. This parameter was optimized using benchmark studies and manual evaluation of randomly selected spectra[15]. False positive are still a possibility, though, and we encourage thorough inspection and validation of supporting evidence for a novel protein. A recommended standard could be the identification of a protein from two different MS runs, as background data and false positives vary between datasets[15].

The pipeline provided here and used for the case study can be modified as pleased to fit the experimental design and parameters. We would recommend using multiple search engines as it increases sensibility and sensitivity of peptide identification[32]. Furthermore, we encourage using the database corresponding best to the experimental aim (**Figure 1**). As using the whole OpenProt database comes with a stringent FDR, true identifications may be lost. Thus, the whole database should be intended for discovery of novel proteins, whilst classical proteomics profiling should be using the smaller OpenProt databases (such as OpenProt_2pep used in the case study above).

OpenProt currently predicts sequences starting with an ATG codon, whereas several studies highlighted translation initiation at other codons[44,45]. When a novel protein is identified by one or several unique peptides, it is possible the true initiation codon is not the presumed ATG. Users can look for translation evidence on the OpenProt website. Currently, OpenProt only reports translation events if they concern the entire predicted protein sequence (100% overlap)[15]. Thus, absence of translation evidence would not mean the protein is not translated, but that the start codon may not be the alleged ATG.

Despite its current limitations, OpenProt offers a more exhaustive view of eukaryotic genomes' coding potential. OpenProt databases foster proteomic discoveries and the understanding of proteomic functions and interactions. Future developments of the OpenProt database will include annotation of other species, translation evidence from non-ATG start codon and development of a pipeline to include novel proteins in whole genome and exome sequencing studies.

ACKNOWLEDGMENTS:

We thank Vivian Delcourt for his help, discussions and advice on this work. X.R. is a member of the Fonds de Recherche du Québec Santé (FRQS)-supported Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke. This research was supported by a Canada Research Chair in Functional Proteomics and Discovery of Novel Proteins to X.R. and CIHR grant MOP-137056. We thank the team at Calcul Québec and Compute Canada for their support with the use of the supercomputer mp2 from Université de Sherbrooke. Operation of the mp2 supercomputer is funded by the Canada Foundation of Innovation (CFI), le ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and les Fonds de Recherche du Québec – Nature et technologies (FRQ-NT). The Galaxy server that was used for some proteomics calculations is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)).

DISCLOSURES:

The authors declare no conflict of interests.

REFERENCES:

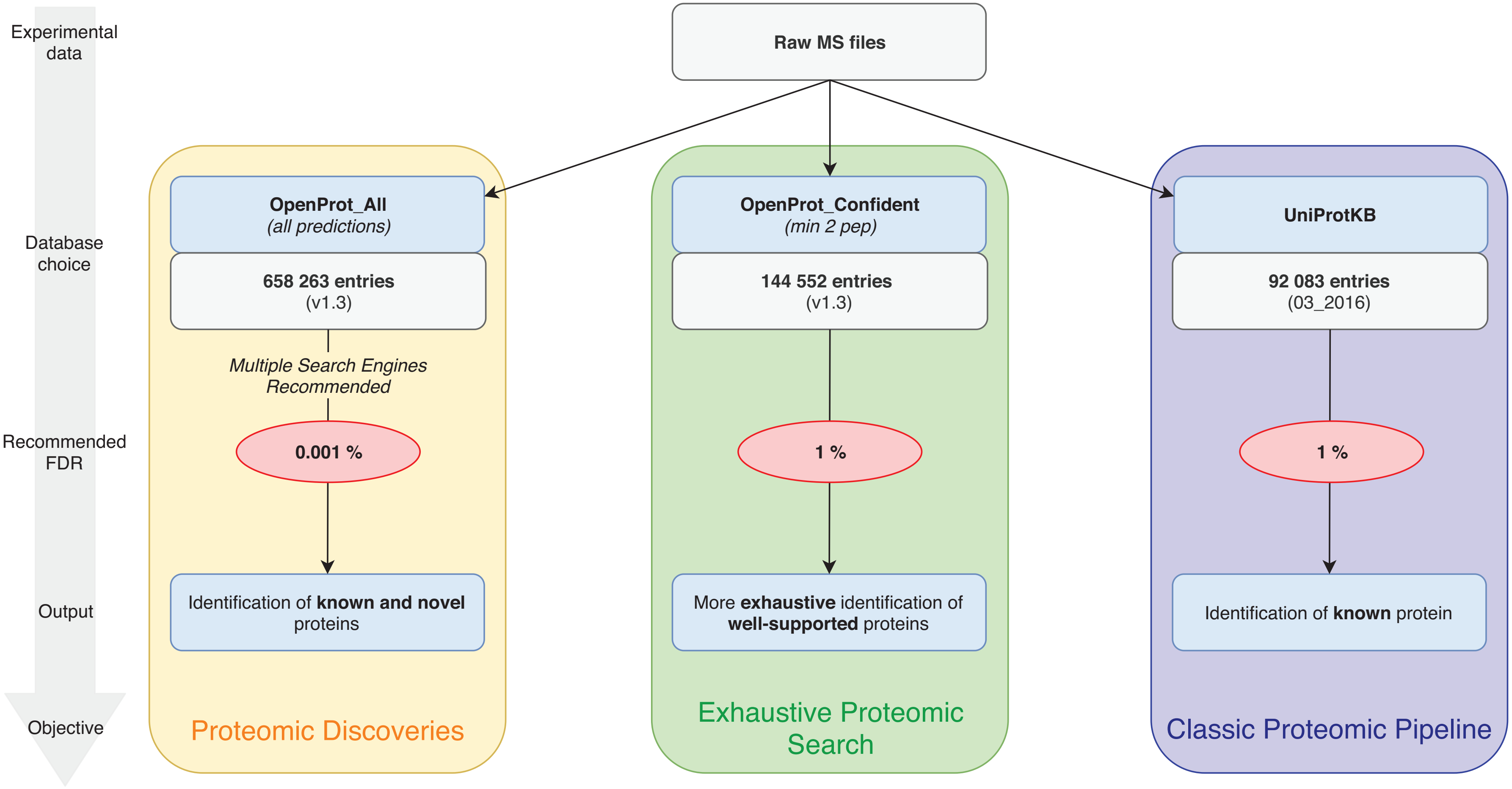
1. Kim, M.-S. et al. A draft map of the human proteome. *Nature*. **509** (7502), 575–581, doi: 10.1038/nature13302 (2014).
2. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. *Nature*. **509** (7502), 582–587, doi: 10.1038/nature13319 (2014).
3. Hein, M.Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*. **163** (3), 712–723, doi: 10.1016/j.cell.2015.09.053 (2015).
4. Huttlin, E.L. et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*. **162** (2), 425–440, doi: 10.1016/j.cell.2015.06.043 (2015).
5. Huttlin, E.L. et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*. **545** (7655), 505–509, doi: 10.1038/nature22366 (2017).
6. Kumar, D., Yadav, A.K., Dash, D. Choosing an Optimal Database for Protein Identification from Tandem Mass Spectrometry Data. *Proteome Bioinformatics*. 17–29, doi: 10.1007/978-1-4939-6740-7_3 (2017).
7. Jeong, K., Kim, S., Bandeira, N. False discovery rates in spectral identification. *BMC Bioinformatics*. **13** (Suppl 16), S2, doi: 10.1186/1471-2105-13-S16-S2 (2012).
8. Brunet, M.A., Levesque, S.A., Hunting, D.J., Cohen, A.A., Roucou, X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Research*. doi: 10.1101/gr.230938.117 (2018).
9. Brent, M.R. Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Research*. **15** (12), 1777–1786, doi: 10.1101/gr.3866105 (2005).
10. Harrow, J. et al. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research*. **22** (9), 1760–1774, doi: 10.1101/gr.135350.111 (2012).
11. Samandi, S. et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*. **6**, e27860, doi: 10.7554/eLife.27860 (2017).
12. Saghatelian, A., Couso, J.P. Discovery and characterization of smORF-encoded bioactive polypeptides. *Nature Chemical Biology*. **11** (12), 909–916, doi: 10.1038/nchembio.1964 (2015).
13. Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., Roucou, X. Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome

Annotations and Current Vision of an mRNA. *Proteomics*. doi: 10.1002/pmic.201700058 (2017).

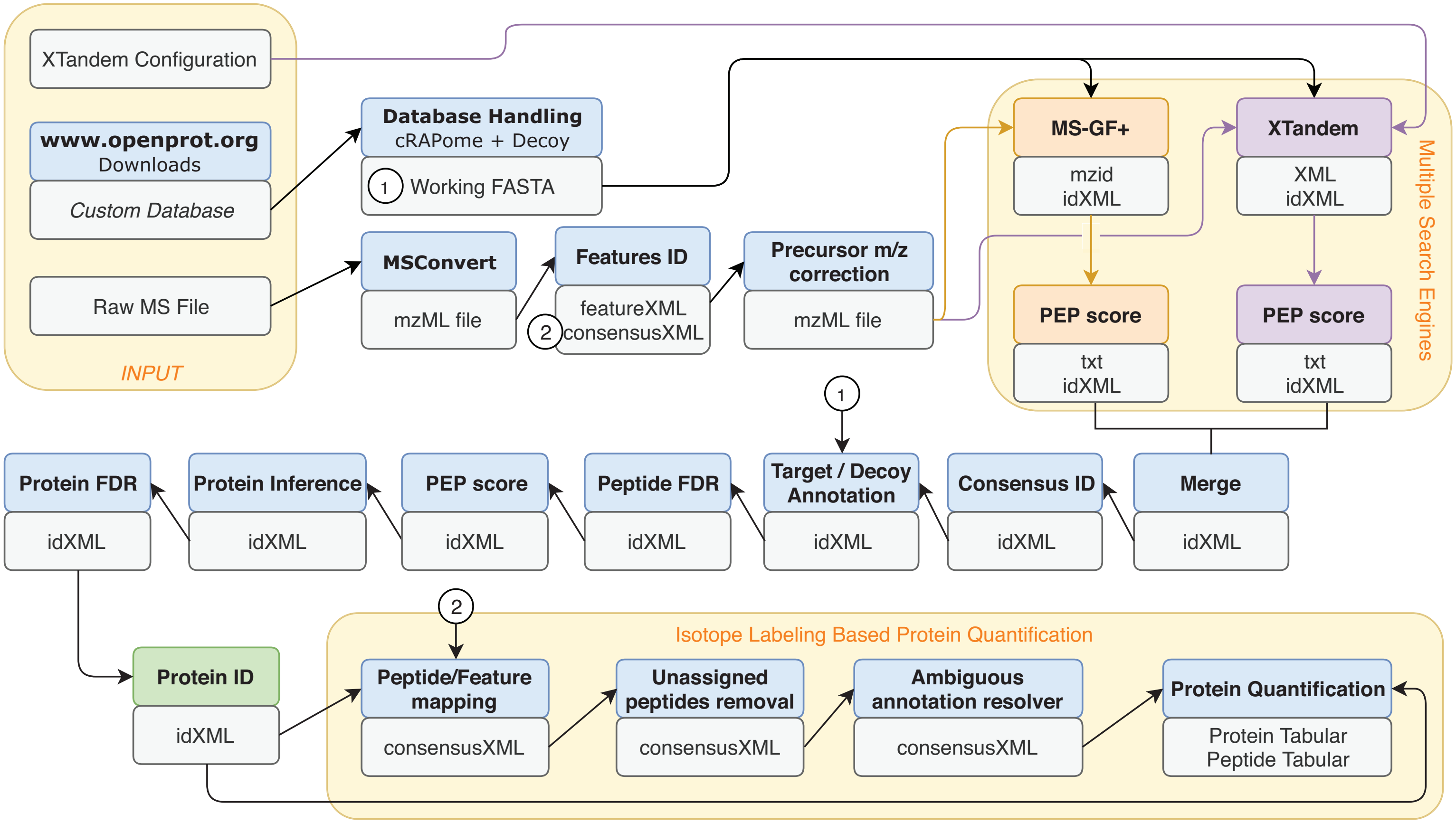
14. Plaza, S., Menschaert, G., Payre, F. In Search of Lost Small Peptides. *Annual Review of Cell and Developmental Biology*. **33** (1), null, doi: 10.1146/annurev-cellbio-100616-060516 (2017).
15. Brunet, M.A. et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Research*. doi: 10.1093/nar/gky936 (2018).
16. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*. **44** (W1), W3–W10, doi: 10.1093/nar/gkw343 (2016).
17. Afgan, E. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*. **46** (W1), W537–W544, doi: 10.1093/nar/gky379 (2018).
18. Sturm, M. et al. OpenMS – An open-source software framework for mass spectrometry. *BMC Bioinformatics*. **9** (1), 163, doi: 10.1186/1471-2105-9-163 (2008).
19. Carithers, L.J. et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*. **13** (5), 311–319, doi: 10.1089/bio.2015.0032 (2015).
20. Aebersold, R., Mann, M. Mass spectrometry-based proteomics. *Nature*. **422** (6928), 198–207, doi: 10.1038/nature01511 (2003).
21. Domon, B., Aebersold, R. Mass Spectrometry and Protein Analysis. *Science*. **312** (5771), 212–217, doi: 10.1126/science.1124619 (2006).
22. Hu, J., Coombes, K.R., Morris, J.S., Baggerly, K.A. The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales. *Briefings in Functional Genomics*. **3** (4), 322–331, doi: 10.1093/bfgp/3.4.322 (2005).
23. Wu, P.-Y., Phan, J.H., Wang, M.D. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*. **14** (11), S8, doi: 10.1186/1471-2105-14-S11-S8 (2013).
24. Mellacheruvu, D. et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature Methods*. **10** (8), 730–736, doi: 10.1038/nmeth.2557 (2013).
25. Adusumilli, R., Mallick, P. Data Conversion with ProteoWizard msConvert. *Proteomics: Methods and Protocols*. 339–368, doi: 10.1007/978-1-4939-6747-6_23 (2017).
26. French, W.R. et al. Wavelet-Based Peak Detection and a New Charge Inference Procedure for MS/MS Implemented in ProteoWizard's msConvert. *Journal of Proteome Research*. **14** (2), 1299–1307, doi: 10.1021/pr500886y (2015).
27. Kuenzi, B.M. et al. APOSTL: An Interactive Galaxy Pipeline for Reproducible Analysis of Affinity Proteomics Data. *Journal of Proteome Research*. **15** (12), 4747–4754, doi: 10.1021/acs.jproteome.6b00660 (2016).
28. Hoekman, B., Breitling, R., Suits, F., Bischoff, R., Horvatovich, P. msCompare: a framework for quantitative analysis of label-free LC-MS data for comparative candidate biomarker studies. *Molecular & Cellular Proteomics: MCP*. **11** (6), M111.015974, doi: 10.1074/mcp.M111.015974 (2012).
29. Bjornson, R.D. et al. X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers. *Journal of Proteome Research*. **7** (1), 293–299, doi: 10.1021/pr0701198 (2008).

30. Kim, S., Pevzner, P.A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nature Communications*. **5**, 5277, doi: 10.1038/ncomms6277 (2014).
31. Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A., Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics*. **11** (5), 996–999, doi: 10.1002/pmic.201000595 (2011).
32. Shteynberg, D., Nesvizhskii, A.I., Moritz, R.L., Deutsch, E.W. Combining results of multiple search engines in proteomics. *Molecular & Cellular Proteomics: MCP*. **12** (9), 2383–2393, doi: 10.1074/mcp.R113.027797 (2013).
33. Bittremieux, W. et al. Quality control in mass spectrometry-based proteomics. *Mass Spectrometry Reviews*. **37** (5), 697–711, doi: 10.1002/mas.21544 (2018).
34. Bertsch, A., Gröpl, C., Reinert, K., Kohlbacher, O. OpenMS and TOPP: Open Source Software for LC-MS Data Analysis. *Data Mining in Proteomics: From Standards to Applications*. 353–367, doi: 10.1007/978-1-60761-987-1_23 (2011).
35. Pfeuffer, J. et al. OpenMS – A platform for reproducible analysis of mass spectrometry data. *Journal of Biotechnology*. **261**, 142–148, doi: 10.1016/j.jbiotec.2017.05.016 (2017).
36. Kozak, M. Pushing the limits of the scanning mechanism for initiation of translation. *Gene*. **299** (1–2), 1–34 (2002).
37. Noderer, W.L. et al. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology*. **10**, 748 (2014).
38. Eyckerman, S. et al. Intelligent Mixing of Proteomes for Elimination of False Positives in Affinity Purification-Mass Spectrometry. *Journal of Proteome Research*. **15** (10), 3929–3937, doi: 10.1021/acs.jproteome.6b00517 (2016).
39. Vizcaíno, J.A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*. **44** (D1), D447–D456, doi: 10.1093/nar/gkv1145 (2016).
40. Bateman, A. et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. **45** (D1), D158–D169, doi: 10.1093/nar/gkw1099 (2017).
41. The Gene Ontology Consortium Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*. **45** (D1), D331–D338, doi: 10.1093/nar/gkw1108 (2017).
42. O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. **44** (D1), D733-745, doi: 10.1093/nar/gkv1189 (2016).
43. Zerbino, D.R. et al. Ensembl 2018. *Nucleic Acids Research*. **46** (D1), D754–D761, doi: 10.1093/nar/gkx1098 (2018).
44. Andreev, D.E. et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife*. **4**, e03971, doi: 10.7554/eLife.03971 (2015).
45. Jackson, R. et al. The translation of non-canonical open reading frames controls mucosal immunity. *Nature*. **564**, 434-438, doi: 10.1038/s41586-018-0794-7 (2018).

Click here to access/download;Figure;Figure1_Database_Choice.eps

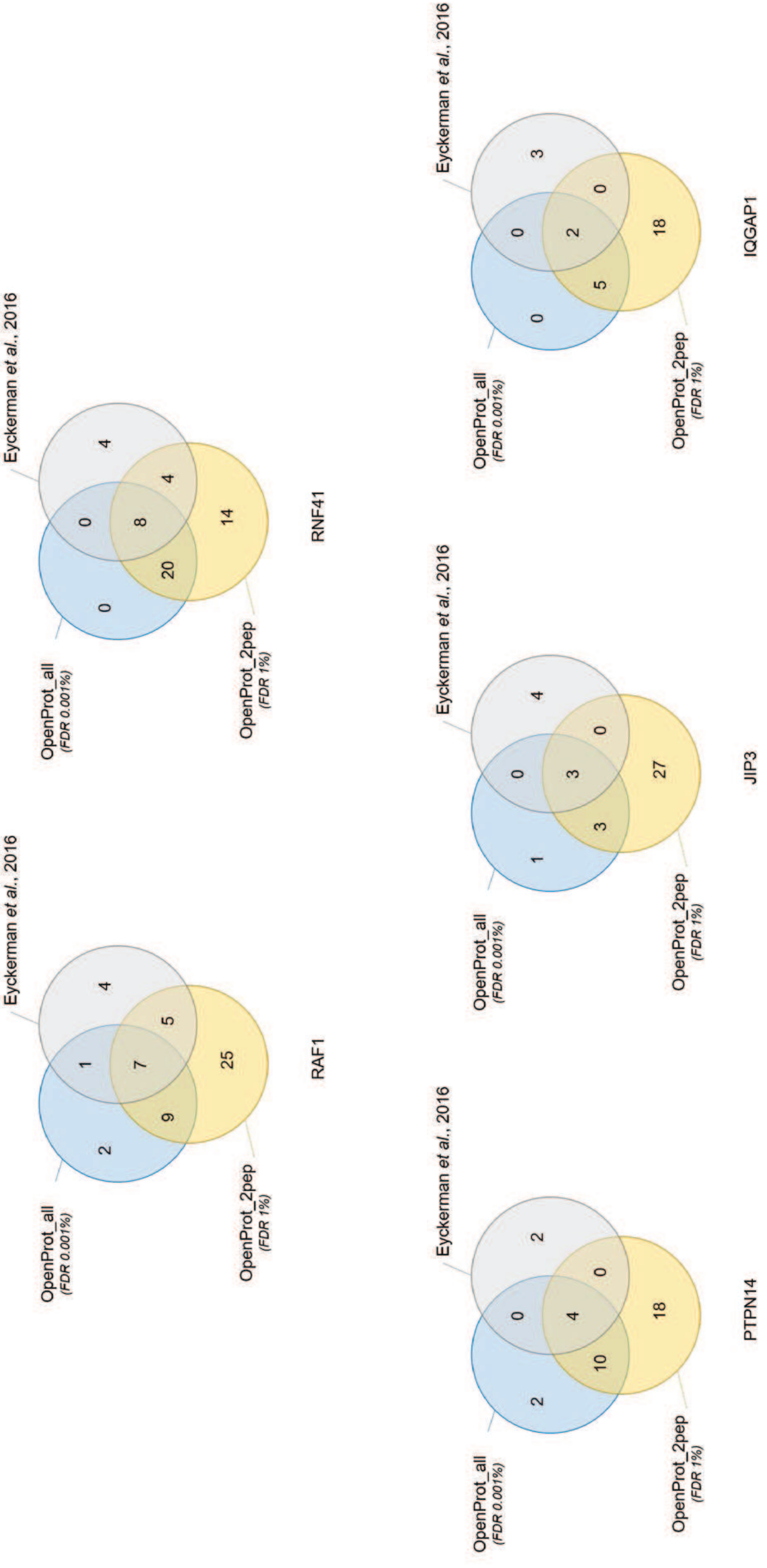


Click here to access/download;Figure;Figure2_Galaxy_OpenProt.eps



Figure

Click here to access/download;Figure;Figure3_VennDiagrams.eps



Figure

Click here to access/download;Figure;Figure4_PSM__IP_637643__RAF1.eps

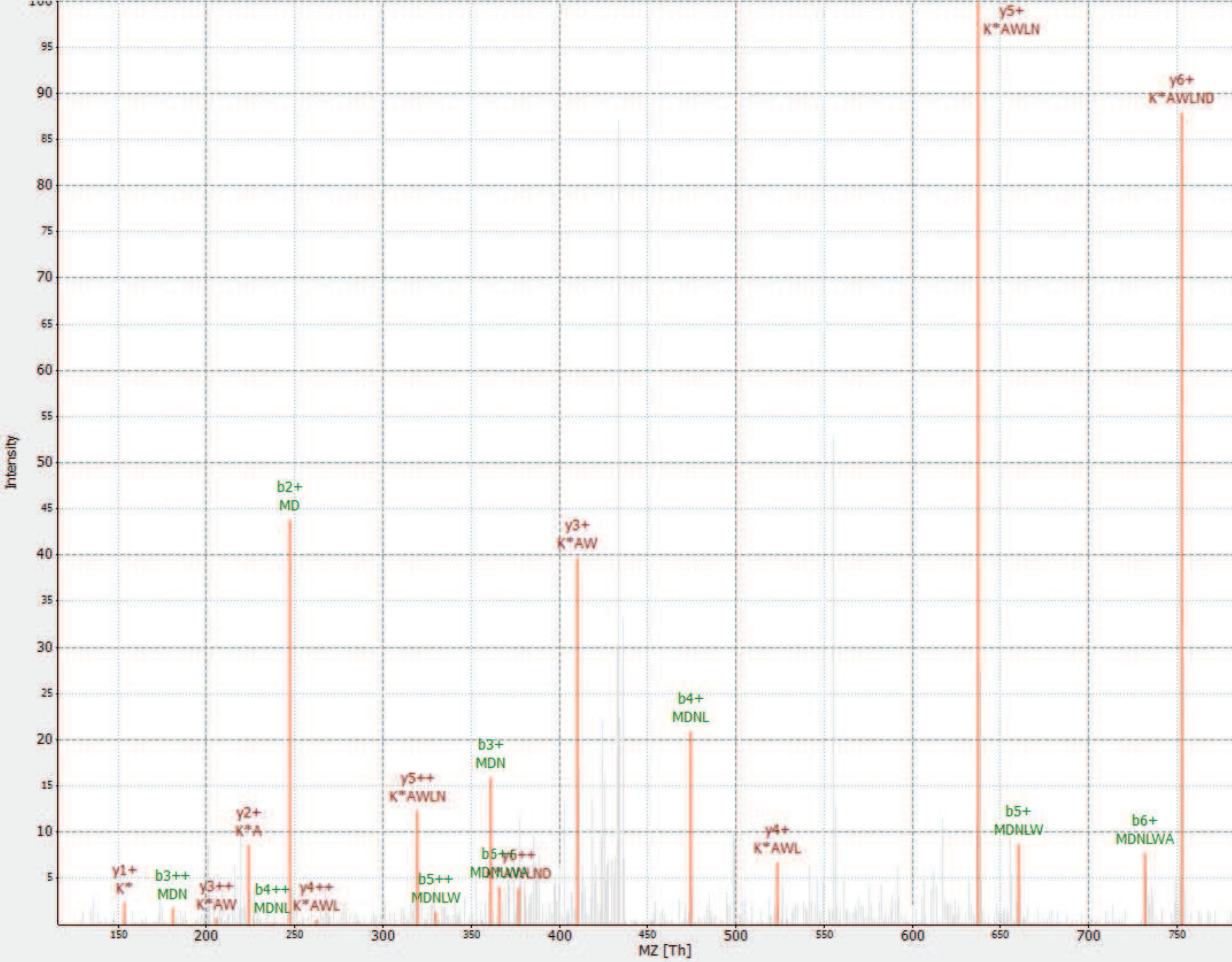


Table 1

Click here to access/download;Table;Table1_Definitions.xlsx

Term
Alternative ORF (AltORF)
Reference ORF (RefORF)
Alternative protein (AltProt)
Reference protein (RefProt)
Novel Isoform
OpenProt_2pep database
OpenProt_1pep database
OpenProt_all database

Definition	Reference
non-canonical ORF currently not annotated in genome annotations, but annotated in OpenProt.	15
canonical ORF annotated in genome annotations and OpenProt.	15
novel protein coded by an AltORF, with no significant similarity with a RefProt. Accession prefix: IP_.	15
protein currently annotated in protein sequence databases such as UniProtKB, Ensembl or NCBI RefSeq, and also in OpenProt.	15
novel protein coded by an AltORF, with a significant similarity with a RefProt. Accession prefix: II_.	15
contains the sequence of all RefProts and novel proteins predicted by OpenProt, already detected with a minimum of 2 unique peptides.	15
contains the sequence of all RefProts and novel proteins predicted by OpenProt, already detected with a minimum of 1 unique peptide.	15
contains the sequence of all RefProts and novel proteins predicted by OpenProt.	15

Table of Materials Click here to access/download;Table of Materials;OpenProt__Table_of_Materials.xlsx

Name of Material/Equipment	Company	Catalog Number	Comments/Description
OpenProt website	open source	n/a	www.openprot.org
Galaxy Server	open source	n/a	https://usegalaxy.eu/
TOPPview software	open source	n/a	www.openms.de

Author License Agreement

Click here to access/download;Author License Agreement (ALA);Author_License_Agreement.pdf

JOURNAL OF VISUALIZED EXPERIMENTS
1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:	Mass spectrometry-based proteomics analyses using OpenProt database to unveil novel proteins translated from non-canonical open reading frames
Author(s):	Marie A. BRUNET, Xavier ROUCOU

Item 1: The Author elects to have the Materials be made available (as described at http://www.jove.com/publish) via:

□ Standard Access ☒ Open Access

Item 2: Please select one of the following items:

☒ The Author is **NOT** a United States government employee.

□ The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

□ The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: “**Agreement**” means this Article and Video License Agreement; “**Article**” means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; “**Author**” means the author who is a signatory to this Agreement; “**Collective Work**” means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; “**CRC License**” means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode; “**Derivative Work**” means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; “**Institution**” means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; “**JoVE**” means MyJove Corporation, a Massachusetts corporation and the publisher of The Journal of Visualized Experiments; “**Materials**” means the Article and / or the Video; “**Parties**” means the Author and JoVE; “**Video**” means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and(c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the “Open Access” box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this **Section 6** is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Protection of the Work.** The Author(s) authorize JoVE to take steps in the Author(s) name and on their behalf if JoVE believes some third party could be infringing or might infringe the copyright of either the Author's Article and/or Video.

9. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

10. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

11. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole

discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

12. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

13. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication of the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

14. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to me one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement is required per submission.

CORRESPONDING AUTHOR

Name:	Marie A. Brunet
Department:	Biochemistry Department
Institution:	Université de Sherbrooke
Title:	Dr

Signature:		Date:	19/12/2018

Please submit a **signed** and **dated** copy of this license by one of the following three methods:

1. Upload an electronic version on the JoVE submission site
2. Fax the document to +1.866.381.2236
3. Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02140

Rebuttal Letter

Click here to access/download;Rebuttal Letter;Answers to reviews.docx

Answers to editorial and reviewer's comments - JoVE59589

Mass spectrometry-based proteomics analyses using OpenProt database to unveil novel proteins translated from non-canonical open reading frames

We would like to thank the editor and reviewers for their thoughtful comments regarding the manuscript and supporting materials. We are confident the reviewers' comments have been addressed and have helped us achieve a better version of the manuscript.

All changes are reported below.

Editorial comments:

General:

1. The manuscript has been thoroughly proofread to correct spelling/grammar mistakes.
2. All formatting recommendations have been followed in this updated version (page size, margins, font and line spacing).

Summary:

1. The Summary has been rephrased to describe the protocol and applications more clearly. Copied here is the updated version: (l. 23-26) "*OpenProt is a freely accessible database that enforces a polycistronic model of eukaryotic genomes. Here, we present a protocol for the use of OpenProt databases when interrogating mass spectrometry datasets. Using OpenProt database for analysis of proteomic experiments allows for discovery of novel and previously undetectable proteins.*"

Protocol:

1. We rephrased some steps to ensure the use of the imperative tense. (*l. 101, 103, 108, 119, 125, 131, 336 and 339)*
2. Everything in the protocol is either numbered or stated as a note. (*l.81, 89, 99, 112, 121, 128, 132, 174, 199, 209, 222, 231, 244, 258, 280, 298 and 314*)
3. We have highlighted in yellow the protocol steps for the video. This represents 2.75 pages (122 lines in total with spacing).
4. We rephrased some steps to ensure the "how" question was properly answered. (*l. 96, 98, 101, 103, 108, 111, 116, 119, 125, 136, 157, 160, 164, 168, 171, 194, 219, 221, 270, 274, 293 and 321*)

Specific Protocol steps:

1. The previously numbered step 1 (Definitions of terms used in OpenProt) has been moved to a Table (Table1) and is referenced at the beginning in a note (*l. 81-82*). All subsequent steps have been renumbered accordingly.

Figures:

1. Figure 3: Percentages now appear as '0.001%'.

Table of Materials:

1. We initially did not provide any information in the Table of Materials as all informatics resources used in this protocol are open source (and URLs are mentioned in the text). We have now filled the table with the website names and URLs.

Furthermore, we had not provided any information on the computer used as each server is hosted elsewhere, hence this protocol can be run on any computer, be it a 1 or 20 CPU, with any exploitation system. However, for full disclosure and if needed, the university computer used for the analysis is a laptop Intel® Core™ i7-7600U CPU @ 2.80 GHz, Windows® 10 system.

Reviewer #1:

Reviewer #1 had no concerns regarding the manuscript. We thank him for his review.

Reviewer #2:

Reviewer #2 had no major concerns, but pointed 3 minor concerns. We thank him for his review.

1. There are some English grammar/spelling errors. Specifically, there are some errors with verb-subject conjugations. The overall meaning is still clear.

The full manuscript has been proofread and mistakes have been corrected.

2. There is a significant reliance on the Galaxy website.

This is a good point that the protocol and supplementary materials provided partly rely on the Galaxy website. We did so to offer a ready-to-use package that someone with no expertise in bioinformatics or proteomics software could use. The Galaxy instance repeatedly pledged maintenance of their services[1, 2].

However, as mentioned in the manuscript (*l. 71-72 and 83-85*), the protocol will work with any proteomics software with minor adjustments specific to the software desired.

3. Figure 2 is a bit confusing. There is a lot going on. Perhaps break it up into 2 figures or simplify.

Admittedly, the Figure 2 is complex. We believe breaking it up would not provide an adequate solution (not easier to read and not fair to the protocol). However, we have simplified the figure, notably re-organizing it to avoid crossing arrows. We think the updated figure is now easier to read and we thank reviewer #2 for challenging us to simplify a complex workflow representation.

Suppl_S5_S6

Click here to access/download;Supplemental File (Figures, Permissions, etc.);S5_and_S6__DataMS.xlsx

Supplementary S4: Quantified proteins from iMixPro datasets.

Data files from Eyckerman *et al* ., 2016 were processed using OpenProt databases and quantif
according to BioGrid that were not reported in the original paper. Gene names indicated in lig

iMixPro paper - **Uniprot** (03.2016)			
PTPN14	JIP3	IQGAP1	RAF1
PTPN14	**JIP3**	**IQGAP1**	**RAF1**
SEPTIN11	FAM32A	CCDC47	BAG2
GPATCH8	KIF5B	**CDC42**	**CDC37**
TMPO	MYH10	DECR2	CHMP4A
VIL1	RBM34	EPB41	CHMP4B
WWC3	**SPAG9**		HSP90AA1
	WDR1		HSP90AB1
			HSP90AB4P
			HSPA8
			IQGAP2
			PDCD6IP
			USP7
			YWHAB
			YWHAE
			YWHAG
			YWHAH
			HSPA1B/1A

ied proteins are listed for each condition. Baits are PTPN14, JIP3, IQGAP1, RAF1 and RNF41. Ge
ht blue correspond to novel proteins identified as interactors (the corresponding protein acces

		OpenProt most confident (min 2 pept	
RNF41	PTPN14	JIP3	IQGAP1
RNF41	**PTPN14**	**JIP3**	**IQGAP1**
BIRC6	VIL1	SPAG9	ACTBL2
CACYBP	WWC3	MYH10	LMO7
FLII	TMPO	LMNB1	PTRF
HOMER1	LIMA1	RPLP2	NPM1
HOMER2	ALDOA	CTNNA1	APEX1
KDM3B	PDIA3	ITPR3	PHB2
KIAA1598	EIF3L	*KRT10*	*KRT1*
LIMCH1	SIPA1L1	WDR36	CALU
LRRFIP2	LMO7	RPF2	*KRT2*
MARK2	HSP90AA1	RAB5C	HNRNPK
MARK3	CLIC1	FASN	RPS28
MTCL1	VCP	SF3B3	HP1BP3
NAV1	HSP90B1	RCC1	LGALS3
NAV2	TRIM28	HMGB1	RAB1A
SOGA1	ENO1	DHX15	CDC42
	GIGYF2	EIF3C	ATP1B3
	PEBP1	C1QBP	*KRT9*
	FASN	PPIA	PGK1
	PTGES3	RP11-402J6.3 (IP_613981	SFPQ
	DHX29	CALU	SYNPO
	KHSRP	SH3BGRL2	HSPA4
	AGR2	PKM	ASNS
	IMPDH2	SPAG9 - isoform	PTGES3
	TENM4	SSH1	C15ORF52
	RALY	ATP5A1	
	HDGF	PFN2	
	ECH1	FKBP4	
	DDX50	EIF3M	
	HNRNPH3	TFG	
	DYNC1H1	ALB	
	TARS	CNDP2	
		EIF3G	
		KRT9	
		PDIA3	
		EIF3F	

ne names indicated in green correspond to proteins also identified in the original paper [20]. Ger
sion number is indicated in brackets). Gene names indicated in light grey and italics correspon

tides - v1.3)			**Op**
RAF1	RNF41	PTPN14	JIP3
RAF1	MYH9	**PTPN14**	**JIP3**
HSP90AA1	MYH10	VIL1	SPAG9
YWHAE	SOGA1	WWC3	MYH10
HSP90AB1	CACYBP	TMPO	LMNB1
HSPA1A/1B	SHTN1	LIMA1	ITPR3
CDC37	NAV1	LMO7	SAPG9 - isoform
YWHAB	MYL6B	CCT2	*KRT10*
YWHAZ	SPECC1L	ALDOA	PKM
YWHAH	BIRC6	DYNC1H1	
YWHAG	MTCL1	HSP90AB1	
HSPA1B	CORO1C	PDIA3	
CHMP4B	**RNF41**	AGR2	
MPRIP	FLII	FASN	
BAG2	TPM3	ENO1	
RCN2	HOMER1	HSP90B1	
RUVBL2	IGF2BP1	VCP	
SPECC1L	HOMER2		
ACIN1	ACTN4		
KRT1	AIF1L		
DARS	MYH14		
GNAZ	RAI14		
MAPK1IP1L	PPIB		
DDX1	BASP1		
SNRNP200	KDM3B		
DDX50	ACTR2		
FLNB	TPM1		
EPB41L4A	POGLUT1		
HNRNPK	MYH11		
RPL5	PAWR		
ATP6V0D1	ACTB		
RAVER1	PRKAR2A		
RPS21	YWHAG		
RAB3GAP1	RPL11		
ATP6V1A	RBM10		
EEF2	SRSF1		
IGF2BP1	MARK2		
ZNF600	WDR1		
STMN1	LRRFIP2		
RANBP1	EIF5B		

CHMP4A	SSFA2
SPTBN1	FLNA
PDCD6IP	MYO18A
FLNA	KNDC1
EIF3A	RAB3GAP2
SNRNP70	TPM4
TRAP1	GSN

ıe names indicated in orange correspond to known interactors
d to likely contaminants (keratin proteins).

ıe**nProt** all predictions (v1.3)		
IQGAP1	RAF1	RNF41
IQGAP1	**RAF1**	MYH10
ACTBL2	HSP90AA1	MYH9
LMO7	HSPA1B	CACYBP
HSPA4	YWHAH	SOGA1
KRT9	YWHAG	SHTN1
KRT1	HSP90AB1	**RNF41**
CDC42	HSPA1A/1B	SPECC1L
	HSPA8	MYL6B
	CHMP4B	FLII
	TRAP1	MTCL1
	KRT1	NAV1
	DDX1	TPM3
	HNRNPK	HOMER1
	PAFAH1B1	CORO1C
	RPL5	MYH14
	RPS21	IGF2BP1
	NANOGNBP1 (IP_637643	ACTN4
	BAG2	BASP1
	SNRNP70	PRKAR2A
		YWHAG
		HOMER2
		WDR1
		PPIB
		AIF1L
		SSFA2
		TPM1
		POGLUT1
		MYH11

Supplementary S5: Identified novel proteins from iMixPro datasets.

Data files from Eyckerman *et al* ., 2016 were processed using OpenProt databases and novel identifi
starting with II_ for novel isoforms of a known protein, and with IP_ for novel proteins from an alter

OpenProt most confident (min 2 peptides - v1.3)					
PTPN14	**JIP3**	**IQGAP1**	**RAF1**	**RNF41**	**PTPN14**
II_772633 (1)	IP_559603 (1)	IP_557834 (1)	IP_689722 (1)	IP_671454 (1)	II_093821 (1)
IP_595308 (1)	IP_613981 (1)	IP_622407 (1)		IP_671456 (1)	II_134861 (2)
	IP_775400 (1)	IP_624921 (2)			II_772633 (1)
					IP_559557 (3)
					IP_564617 (1)

ied proteins are listed for each condition. Baits are PTPN14, JIP3, IQGAP1, RAF1 and RNF41. P
native ORF (AltProt). The number of supporting peptides are indicated into brackets.

OpenProt all predictions (v1.3)			
JIP3	**IQGAP1**	**RAF1**	**RNF41**
II_134861 (4)	II_566112 (1)	II_083225 (1)	II_150058 (17)
II_576004 (3)	II_590131 (49)	II_590131 (17)	II_590131 (28)
II_604356 (1)	II_711657 (22)	II_711657 (15)	II_711657 (7)
IP_559603 (1)	II_711659 (3)	IP_637643 (1)	IP_132426 (1)
IP_746392 (2)	IP_557834 (1)		IP_2304182 (1)
IP_788439 (2)	IP_622407 (1)		IP_2370385 (1)
	IP_624921 (2)		IP_743029 (1)

rotein accession numbers are listed,

Suppl_S1

Click here to access/download
Supplemental Coding Files
Supplementary_S1_Galaxy-Workflow_Database_Handling.ga

Suppl_S2

Click here to access/download
Supplemental Coding Files
S2_Galaxy_Pep_Prot_Id.ga

Suppl_S3

Click here to access/download
Supplemental Coding Files
S3_Galaxy_SILAC_Quant.ga

Suppl_S4

Click here to access/download

Supplemental Coding Files

S4_XTandem_Config.xml