nature protocols

PROTOCOL UPDATE

https://doi.org/10.1038/s41596-022-00690-x

Simple, efficient and thorough shotgun proteomic analysis with PatternLab V

Marlon D. M. Santos[1], Diogo B. Lima[2], Juliana S. G. Fischer[1], Milan A. Clasen[1], Louise U. Kurt[1], Amanda Caroline Camillo-Andrade[1], Leandro C. Monteiro[1], Priscila F. de Aquino[3], Ana G. C. Neves-Ferreira[4], Richard H. Valente[4], Monique R. O. Trugilho[4,5], Giselle V. F. Brunoro[6], Tatiana A. C. B. Souza[1], Renata M. Santos[1,7], Michel Batista[8], Fabio C. Gozzo[9], Rosario Durán[10], John R. Yates III[11], Valmir C. Barbosa[12]✉ and Paulo C. Carvalho[1]✉

Shotgun proteomics aims to identify and quantify the thousands of proteins in complex mixtures such as cell and tissue lysates and biological fluids. This approach uses liquid chromatography coupled with tandem mass spectrometry and typically generates hundreds of thousands of mass spectra that require specialized computational environments for data analysis. PatternLab for proteomics is a unified computational environment for analyzing shotgun proteomic data. PatternLab V (PLV) is the most comprehensive and crucial update so far, the result of intensive interaction with the proteomics community over several years. All PLV modules have been optimized and its graphical user interface has been completely updated for improved user experience. Major improvements were made to all aspects of the software, ranging from boosting the number of protein identifications to faster extraction of ion chromatograms. PLV provides modules for preparing sequence databases, protein identification, statistical filtering and in-depth result browsing for both labeled and label-free quantitation. The PepExplorer module can even pinpoint de novo sequenced peptides not already present in the database. PLV is of broad applicability and therefore suitable for challenging experimental setups, such as time-course experiments and data handling from unsequenced organisms. PLV interfaces with widely adopted software and community initiatives, e.g., Comet, Skyline, PEAKS and PRIDE. It is freely available at http://www.patternlabforproteomics.org.

This protocol is an update to *Nat. Protoc.* 11, 102–117 (2015): https://doi.org/10.1038/nprot.2015.133

Introduction

Shotgun proteomics is the mainstream method for analyzing and quantitating proteins in complex protein mixtures such as those originating from cell lysates, tissues or body fluids[1]. This approach, developed during the 1990s in the Yates Lab, back then at the University of Washington, consists in digesting the protein mixtures and chromatographically separating the peptides online with tandem mass spectrometry. The tandem mass spectra originating from fragmented peptides (MS2) were, in those years, identified with SEQUEST[2], the first software to correlate experimental spectra with those theoretically generated from a sequence database[2]. This strategy provided sequence identification and enabled large-scale proteomic analyses. Following identification, protein sequences were inferred by matching identified peptide sequences to those from a sequence database. As peptides can match more than one protein in the database, a list of proteins was also typically reported according to a maximum parsimony approach (i.e., an approach reporting only the minimum number of proteins that can justify all peptide identifications[3]).

[1]Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute, Fiocruz Paraná, Curitiba, Brazil. [2]Department of Structural Biology, Leibniz-Forschungsinstitut für Molekulare Pharmakologie (FMP), Berlin, Germany. [3]Leonidas and Maria Deane Institute, Fiocruz Amazonas, Manaus, Brazil. [4]Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil. [5]Center for Technological Development in Health, Fiocruz, Rio de Janeiro, Brazil. [6]Research and Development in Biotechnology, Blau Farmacêutica S/A, Cotia, Brazil. [7]Chemistry Institute, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil. [8]Mass spectrometry facility, Carlos Chagas Institute, Fiocruz Paraná, Curitiba, Brazil. [9]Dalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil. [10]Analytical Biochemistry and Proteomics Unit, Instituto de Investigaciones Biológicas Clemente Estable/Institut Pasteur, Montevideo, Uruguay. [11]Laboratory for Biological Mass Spectrometry, The Scripps Research Institute, La Jolla, CA, USA. [12]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil.
✉e-mail: valmircbarbosa@gmail.com; paulo@pcarvalho.com

At first, confident identifications were achieved by selecting peptide spectrum matches (PSMs) with XCorr (the cross-correlation score used by SEQUEST to compare spectra) >2.5 and ΔCN (a normalized difference between the first and the second XCorr matches) >0.1 for 2+ peptide ions. These scores were empirically defined and later became known as the Washburn criteria[1]. Since such an empirical approach provided no statistics, it was hard to compare results acquired on different mass spectrometers or originating from different research groups. With this as motivation, more sophisticated approaches for filtering confident identifications soon emerged, such as PeptideProphet[4] and DTASelect[5], among others. PeptideProphet initially relied on modeling negative hits with a gamma distribution and positives with a normal, but newer versions allow one to opt for extreme value distributions for the negatives and can take advantage of kernel density-based distributions if decoys are at hand (i.e., if artificially generated protein sequences were included in the database). Alternatively, DTASelect requires prior searching on a target–decoy sequence database, so a false-discovery rate (FDR) can be applied by considering the distribution of decoy-hit scores as a surrogate for an error distribution. The score distribution is then used to converge to a confident result under an acceptable FDR[4]. Barboza et al.[6], however, showed that it is possible to obtain untrustworthy estimates under an FDR system (i.e., a number of false identifications higher than that predicted by the model can be obtained), so new tools should be extensively evaluated[5]. This occurs mostly because the software filtering function may overfit the data (the software is too attuned to the decoys), but can also be due to imbalances in how the ratio of target peptides in the search database compares to that of decoys. PatternLab's filtering function was evaluated using databases that included an additional layer of decoys (unlabeled decoys) not known to the software. Our results showed that the number of unlabeled decoys roughly matched that of decoys, and thus no overfitting occurred[7]. We perform this test from time to time when updates are done to the filtering module. Modern target–decoy database generation also preserves the decoy-to-target ratio of peptides, as imbalances may lead to inaccuracies in the FDR estimation as well[8,9].

It is worth noting there exist several other search engines and proteomic computational environments; some of the most well known are Mascot[10], PEAKS[9], the Trans-Proteomic Pipeline (TPP)[11] and OpenMS[12]. Each has specific advantages and disadvantages; for example, Mascot has a web interface, while PEAKS runs on a desktop environment, and both offer commercial support; TPP and OpenMS run on several platforms, and most of the code is open source. Computational environments for proteomics tend to overlap one another, but there are differences that one should keep an eye on before closing in on a solution for one's lab. For example, TPP and Mascot are tailored for working on computer clusters, which tends to make them more burdensome for the user to install and upgrade, but as an advantage, all the computational heavy lifting is done on the server side, releasing the user's desktop from any burden. In contrast, OpenMS and PatternLab are tailored toward local execution, with software updates controlled by the user (PatternLab V (PLV) issues a warning when updates are available). Planning carefully when to update is advisable, as some updates might affect ongoing studies and might have an immediate (and sometimes undesired) impact. This can be seen as an advantage of desktop-contained solutions as the end user decides when to upgrade and roll-backs are possible. Finally, today's high-end desktops (and even notebooks) are powerful enough to carry out all large-scale proteomic analyses.

Development of the protocol

As with many labs in different countries, the vast majority of proteomic analyses in Brazil in the early 2000s were performed via two-dimensional electrophoresis. Spots of interest were excised from the gel, and through a laborious procedure the proteins within were trypsinized and analyzed, typically by MALDI-TOF mass spectrometry. While PhD students, two of us (P.C.C. and J.S.G.F., respectively in computer science and biochemistry) visited the Yates Lab, relocated to The Scripps Research Institute in California, to learn shotgun proteomics. The lab was unique in its ability to congregate different disciplines: chemists to advance the analytical and mass spectrometric approaches, biologists to focus on the relevant questions and analyze results, and computer scientists to enable data analysis. The continued convergence of these three disciplines enabled the lab to put forward many breakthroughs in the field. The lab had more than ten mass spectrometers, most of them linear traps. Data analysis was a lot less straightforward back then as users had to convert the raw mass spectral data into an open text format, namely MS2[13], transfer the data to Garibaldi, one of Scripps's 'superclusters', and then use a command-line interface to the cluster to follow a complex set of routines and run several independent software modules. At that time, ProLuCID was the main search engine in use in the lab, as it could easily be run on the cluster[14]. This working pipeline was unmatched; indeed, the lab was one of the very few specialized in proteomics, besides having access to the unparalleled computational resources provided by the institution. At the time, it was difficult for this type of data-analysis workflow to be widely adopted by institutions worldwide, particularly those in developing countries. The amount of data generated in large-scale proteomic experiments was still too much to be analyzed using a typical desktop computer; moreover, there were no straightforward, user-friendly desktop software solutions at the time.

However, advances in desktops were happening at an impressive pace. With this as motivation, a shotgun proteomic computational environment was conceived to be fully executed on a desktop, being easy to install and equipped with a graphical user interface (GUI). One goal was to have the system ready to enable data analysis by the time the said institutions could acquire their first Orbitraps dedicated to proteomics. The resulting software, PatternLab for Proteomics[15], was ready right in time to be one of the first to demonstrate the effectiveness of differential proteomics using spectral counting[16] in a real but controlled experiment where known markers were spiked in a yeast lysate[17]. Indeed, this was one of the first studies to rely on label-free proteomics.

Extracted ion chromatograms (XIC)-based quantitation, today's mainstream label-free approach, was not popular back then, as the proteomic workhorses were mostly linear trap quadrupoles (LTQs); these instruments did not have enough resolution and mass accuracy to provide trustworthy XICs for complex protein mixtures. Two years later, as the software continued to evolve, the first protocol on using PatternLab (then in version 2) was published[18]. PatternLab began to incorporate independent data processing modules, one of the first being the Search Engine Processor (SEPro)[7].

The SEPro module aimed to filter search results to match a user-specified FDR; it first used a three-tier filtering approach with roots in Bayesian discriminant functions. As a result, it identified ~20% more peptides on a yeast lysate than other tools[7]. Most importantly, some tools could overfit and report a list of results that, even though satisfying a given FDR, were still not trustworthy[6]. SEPro was benchmarked using a semi-labeled decoy approach to ensure the trustworthiness of its results and was later certified by other groups[19]. A new protocol was published[20] reflecting the changes that led to PatternLab 3. This version unified several new tools under the hood, such as a TFold test to combine variable fold changes with P values to shortlist differentially abundant proteins[21], a probabilistic Venn diagram[22] and a module for analyzing time-course experiments[23].

A major request from the community was for user experience to be further improved and for dependencies to be reduced; for example, PatternLab 3 used ProLuCID, which required installation of the Java Runtime Environment. Several PhD students at Brazilian institutions began to develop projects that resulted in new modules integrated into the PatternLab suite. For instance, Leprevost and collaborators developed PepExplorer, a module for the analysis of unsequenced organisms by interpreting de novo sequencing data[24], and Santos and collaborators created a module to deal with isobaric labeling (e.g., iTRAQ and TMT) and the XD Scoring system for evaluating confidence of phosphopeptides[25]. The integration of these new modules, plus several other adaptations, for example, switching from ProLuCID to Comet[26] to reduce dependencies, culminated in 2016 with the release of PatternLab 4 (PL4)[27]. While PL4 has become adopted in many countries, its dissemination through South America has been particularly effective and helped ignite many collaborations in the region.

From the perspective of its development, one of the most valuable additions to PL4 was a user forum to assist the broader community in providing suggestions and quickly reporting any system instabilities. As a natural consequence of such crucial input from the community, the software quickly evolved, adding >100 updates to PL4. At the beginning of 2020, a major rewrite was undertaken based on the latest available frameworks (.NET 5). User experience has since been completely revamped by reducing the time to run several data analysis modules by at least 50% and by improving the GUI. New data analysis features were also included in most modules. The resulting PLV is the biggest and most important update so far. Most recent projects will only be compatible with PLV (not with PL4), especially those analyzing data with mixed-data acquisition[28], which combines data-dependent and data-independent acquisitions in the same run. Therefore, the up-to-now current PL4 protocol has been surpassed and replaced by the unified PLV protocol presented here.

PLV protocol

The present protocol describes how to work with the redesigned interfaces and features included during the past 4 years. The user experience has been substantially simplified, and all items are now loaded through what we have called a ‘universal loader’, accessible through the ‘Load’ menu item. This function automatically directs the user to the desired module, be it a search engine result or a Thermo.raw file browser. The search engine and filtering steps are now joined and simplified, and no longer require the configuration of technical parameters. Though editing all the advanced parameters is still possible, all parameters default to commonly used values so data analysis can be carried out as soon as the setup of the mass spectrometer used has been specified.

PLV relies on a new SEPro format (with file-name extension.sepr2) to store new features, e.g., PSM retention times for improved compatibility with SkyLine[29] in PRM or mixed-data acquisition for data-independent acquisition experiments[28]. The revamped SEPro also includes new features, such as supporting protein families and their inference, which has proven useful especially in simplifying the analysis of highly redundant data (e.g., as in venomics). Improvements to the XIC algorithm include higher throughput and better XIC extraction. The XIC Browser went through a complete rewrite, with data now loaded significantly (more than ten times) faster and with the possibility to easily scale to extensive proteomic experiments (>100 raw files), thanks to an embedded NoSQL database.

Applications

The PatternLab for proteomics software has been tested on millions of spectra by various groups and has addressed a broad range of biological questions by identifying, quantitating and helping interpretation in several types of proteomic experiments, including labeled and label-free proteomics. The applicability of previous versions of PatternLab is very broad; some examples of works that resorted to our software for helping to interpret proteomic data are on (a) uncovering the role of the BAF complex heterogeneity in maintaining the transcriptional network of pluripotency in mouse embryonic stem cells[30], (b) highlighting exosomes as extracellular mediators promoting tumor progression[31], (c) unveiling mechanisms of Z-ring assembly and regulation in bacteria[32], (d) describing how methylation aids in bacterial adhesion and host cell invasion[33] and (e) investigating the effects of quinoa bioester in epidermal tissue[34].

Experimental design

Replicates

PLV is adaptable to analyze a broad range of proteomic experimental designs; however, the software requires that the dataset include both biological and, ideally, technical replicates. We recommend a minimum of three biological replicates and, especially for more complex studies (e.g., assessing differences in biopsies of diseased versus control tissue), as many as possible. The topics on preparing proteomics samples and generating mass spectrometry data are broad and beyond this protocol’s scope; we refer the reader to the protocol published by Richards et al.[35] as a starting point.

Sequence databases

PLV is designed to address the requirements of most shotgun proteomic experiments, typically comparing alterations in the protein abundance profiles of organisms, tissues, fluids and other biological materials in different biological states. The first step generally relies on downloading a sequence dataset for the study at hand; we recommend referring to UniProt for this[36] yet note there are other widely adopted sequences databases, including NCBI and NeXtProt[37], to name two. The UniProt knowledge base comprises both Swiss-Prot, with manually annotated and reviewed sequences, and TrEMBL, with automatically annotated but not reviewed sequences that await full annotation.

While downloading both Swiss-Prot and TrEMBL separately is acceptable (and widely done), there are advantages and disadvantages to be considered. Notably, the joint database is expected to contain higher identification rates, but on the other hand its increased redundancy may create additional challenges in the interpretation of downstream quantitative results. Very large sequence databases (>100,000 sequences) or simultaneously considering several variable post-translational modifications (PTMs) while searching may negatively impact the sensitivity of peptide identification owing to an increased number of high-scoring random hits[38].

Peptide spectrum matching

The PSM approach is carried out by the Comet–SEPro duo; Comet is responsible for comparing the experimental spectra with those theoretically generated from the sequence database, and SEPro used for shortlisting trustworthy results, typically abiding to a 1% FDR. The Comet–SEPro duo is now embedded into PLV, and the two modules now work together, in a single module. It is important to note that only sequences found in the sequence database will be identified. The user must specify the expected amino acid modifications to be considered by the search engine, be they of natural or artifactual occurrence (e.g., cysteine carbamidomethylation and lysine dimethylation).

Peptide sequencing from the mass spectrum

If there is no sequence database available for the organism, or only minimal data are available, PLV can still be used through the PepExplorer module. Briefly, in this approach, the user resorts to a de novo sequencing algorithm to first interpret the tandem mass spectra and provide sequence candidates.

We recommend Novor, a freely available de novo sequencing tool[39]. Among the limitations of this approach, de novo sequencing is not efficient for the low-resolution MS2 typically provided by instruments such as an LTQ; we recommend spectra to be acquired with a resolution of at least 15,000 and a mass accuracy <20 ppm at both MS1 and MS2 levels, as provided by a Q-Exactive Plus or later orbitrap models.

In what follows, PepExplorer aligns the de novo results against a sequence database from homologous organisms and provides an interpretable report listing similar proteins and the aligned peptides. Regardless of whether the identifications are derived from PSMs or PepExplorer, the next step is to organize the experimental files and perform quantitation; for this, PLV provides the Project Organization Module, which allows the user to categorize the experimental mass spectral files as technical or biological replicates and into different biological conditions.

Peptide and protein quantitation

The user then specifies a quantitation strategy; typical examples are spectral counting and XICs. Spectral counts serve as surrogates for relative abundance by reporting the number of spectra associated with a protein; this approach is mainly used for data generated by experiments that include online (MudPIT) or offline chromatographic prefractionation[1], as well as when low-resolution instruments are used (e.g., LTQ).

XICs are obtained by integrating the intensity of a given peptide's ion current over a very narrow (typically <10 ppm) mass-to-charge window as a function of time. XICs are the mainstream option for high-resolution and high-mass-accuracy mass spectrometers, and provide more sensitivity than spectral counting.

PLV also allows one to work with labeled strategies such as iTRAQ and TMT[40], and with strategies that produce heavy and light versions of a peptide, such as SILAC[41] or isotope labeling[42].

As soon as the quantitation is finished, PLV will automatically present the results to the user in the appropriate module; for example, if XICs are the choice method, the XIC Browser will automatically open. The XIC Browser allows the user to verify and set up several stringency parameters, such as whether contaminant or decoy identifications should be considered, whether only quantitation of unique peptides should be accounted for, the minimum number of peptides per protein, and data normalization options. When all is set, the user will finally generate the PatternLab project (.plp) file required for all downstream bioinformatics analyses.

The .plp file unlocks several data analysis modules. Typically, the Buzios module is used first as it provides a bird's-eye view of the experiment.

- Buzios displays dimension reduction plots, such as those resulting from principal component analysis (PCA) or multidimensional scaling and allows one to check whether biological conditions group together. It should be noted that interpretation of such plots is subjective; not grouping as expected suggests hidden patterns in the data
- In a similar vein, the Clustergram module relies on a hierarchical clustering of the protein profiles of the various biological conditions to present another overall view of the experiment. Clustergram's result is a heatmap where each column corresponds to a biological condition and each line represents the quantitative profile across the columns; an overarching dendrogram interconnects the columns. Samples from the same biological condition are expected to be clustered together
- The Area-proportional Venn Diagram shortlists proteins (or peptides) identified in a single condition. Note that, even though a protein may be reported as unique to a condition, it might just be the case that it was undersampled in another condition. To minimize this shortcoming, the user should consider only proteins identified in two or more biological replicates or opt to use the probabilistic Venn diagram module[22] instead

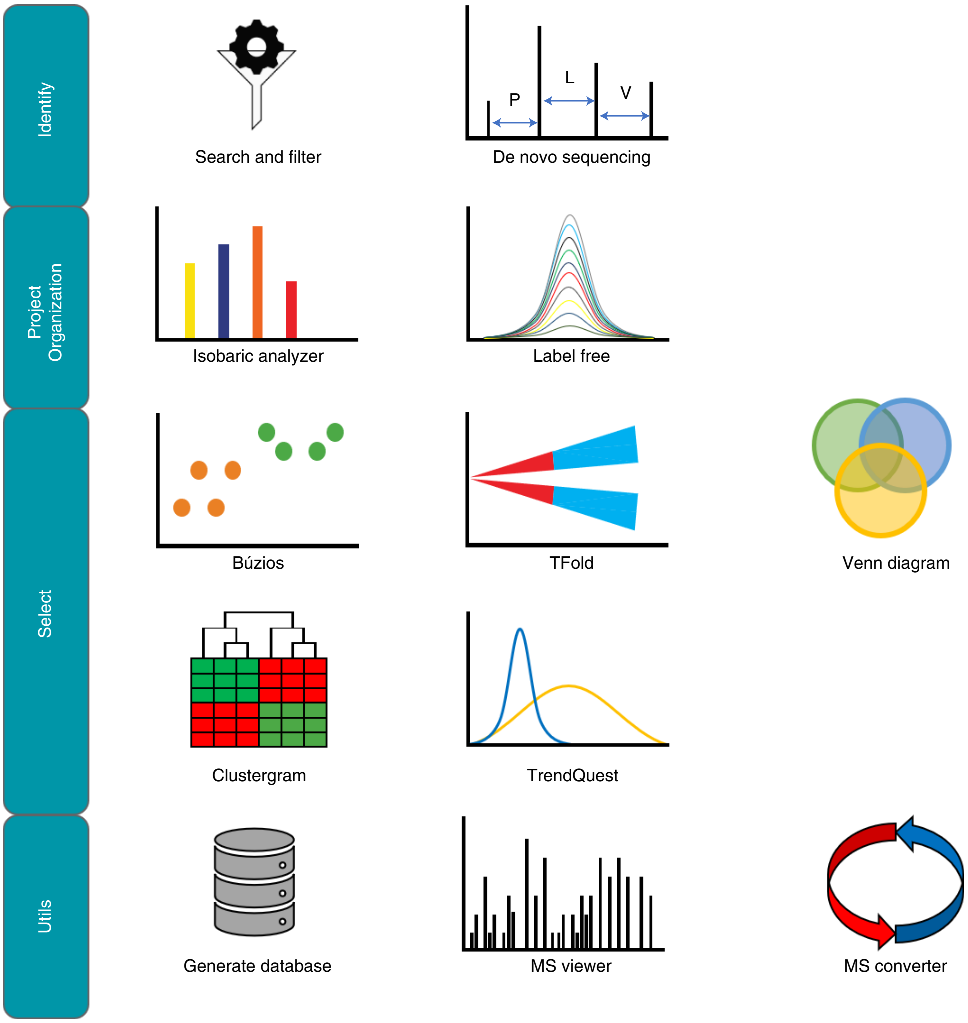


Fig. 1 | Overview of the PLV workflow. The major steps in the PLV pipeline are: 'Identify', identify proteins/peptides by either using the search engine to find matches between the data and a target-decoy database and statistically filtering for confident results or with PepExplorer for de novo sequencing interpretation; 'Project Organization', organize the project in terms of which mass spectral files belong to which biological condition and then perform label-based or label-free quantitation; 'Select', differentially abundant proteins using the various data analysis modules that are integrated into PLV; 'Utils', PLV also provides several utilities, such as viewing mass spectra directly from Thermo.raw files or converting raw files into different formats that are widely used for sharing data and enabling full PRIDE submissions.

Differential abundance

A major proportion of the proteins will typically be shared among all experimental conditions; to pinpoint which are differentially abundant, the user must resort to the TFold module[21]. In brief, TFold works by maximizing the number of identifications that satisfy a fold-change cutoff that varies with the *t*-test *P* value as a power law. TFold provides additional stringency options, such as removing proteins with low quantitation from the analysis; this option is typically used to overcome a limitation of spectral counting data, which is that proteins with very low spectral counts can artificially yield very low *P* values with acceptable fold changes. For example, say that a given protein produced a spectral count of 1 per biological replicate (total of three) in biological condition A, and counts of 4, 5, 4 for the replicates in biological condition B; these data would produce a 4.3-fold change with $P < 0.05$; nevertheless, this *P* value might be artifactual given that the quantitation values of condition A could be a result of, say, dynamic exclusion.

For experiments that comprise more than two biological conditions, such as a time-course experiment, the TrendQuest module comes in handy[34]. TrendQuest will group proteins according to their expression profiles; it is expected that proteins with similar trends sbelong, for example, to the same metabolic pathway[23]. Finally, PLV also provides tools of broader applicability, such as a raw file spectrum browser and tools for combining results across several experiments, the online user forum and the generation of results in the mzIdentML[43] format for full PRIDE[44] compatibility. An overview of the PLV workflow is presented in Fig. 1.

Materials

Equipment

Hardware requirements

- A desktop computer with at least 16 GB of RAM (32 GB strongly recommended) and a x64 processor. We do not recommend using entry-level processors (e.g., i3, Ryzen 3)
- Local storage is required while processing the mass spectrometer raw files. The requirements can vary substantially, depending on the mass spectrometer used. A typical Q-Exactive analysis can generate ~1–2 GB of data in a single run. We believe that ~100 GB of free storage on the computer should be enough

Data files

- Data files in any of these formats are acceptable: mzML[45], mzXML, MGF, MS2 (preferred)[13] or Thermo.raw (preferred)

Software requirements

- Microsoft Windows 10 or later (64-bit version). NET Desktop Runtime 6.0.3 or later (64-bit version), which is free to download from Microsoft at https://dotnet.microsoft.com/en-us/download/dotnet/6.0. If Runtime is not detected during PLV installation, an attempt will be made to install it automatically
- Visual C++ Runtime 14 or later. A link for download is available at https://www.microsoft.com/en-us/download/details.aspx?id=48145

Procedure

Installing the software ● Timing Typically 3 min if the computer already has the .NET Desktop Runtime 6.0.3 or later but might take ~15 min if installation is required

▲CRITICAL Administrative privileges are required for installation.

1 Target your preferred browser to http://www.patternlabforproteomics.org.

2 Click 'DOWNLOAD' on the website. PLV checks for updates every time it is loaded.

Downloading a sequence database ● Timing 5 min

▲CRITICAL Our preferred source for downloading protein sequences is http://www.uniprot.org/proteomes, and the steps described below relate to this source; nevertheless, there are other widely adopted sources.

3 On the website, click on the 'Proteomes' tab and then type the name of an organism of interest (e.g., *Homo sapiens*).

4 In the search results, look for an entry that, preferably, has the full RefSeq genome representation; this information is available in one of the search result columns. Check the Protein count column as this reflects how many proteins there are for the given Proteome ID. As of this writing, there are 77,027 proteins for *Homo sapiens* (human).

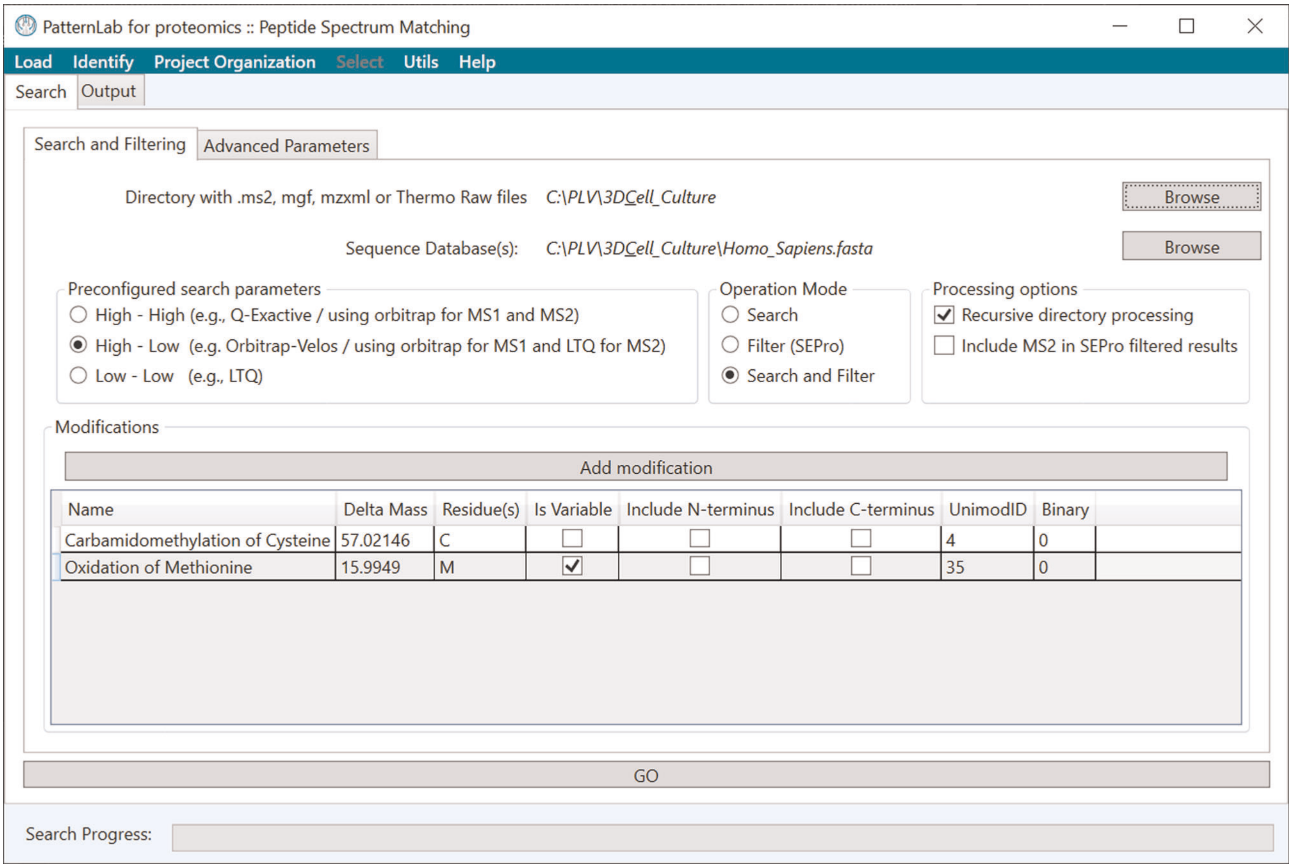
5 Click on the number of the Protein count column; this will redirect the browser to list the proteins for the organism of interest.

6 Click the 'Download' option; select 'Uncompressed' and then 'Go'.

Performing PSM with the integrated Comet search engine ● Timing <1 h to >1 d, depending on the number of samples analyzed and the equipment used

7 Organize the mass spectrometry files. A typical proteomic experiment comprises multiple biological conditions, each having several biological replicates, most often including technical replicates in attempt to reduce undersampling.

- Directory names should not include special characters (e.g., !, @,#,$, ç, á, à, ã)
- No mass spectrum file should have the same name as another, even if placed in different directories
- We recommend first creating a main directory with the experiment name, then creating a subdirectory for each biological condition; for example, if we compare cancer versus control biopsies, appropriate directory names for the biological conditions are cancer and control


PatternLab for proteomics :: Peptide Spectrum Matching
Load Identify Project Organization Select Utils Help
Search Output
Search and Filtering Advanced Parameters
Browse
Browse
Add modification
GO

Directory with .ms2, mgf, mzxml or Thermo Raw files C:\PLV\3DCell_Culture

Sequence Database(s): C:\PLV\3DCell_Culture\Homo_Sapiens.fasta

Preconfigured search parameters
○ High - High (e.g., Q-Exactive / using orbitrap for MS1 and MS2)
● High - Low (e.g. Orbitrap-Velos / using orbitrap for MS1 and LTQ for MS2)
○ Low - Low (e.g., LTQ)

Operation Mode
○ Search
○ Filter (SEPro)
● Search and Filter

Processing options
☑ Recursive directory processing
□ Include MS2 in SEPro filtered results

Modifications

Name	Delta Mass	Residue(s)	Is Variable	Include N-terminus	Include C-terminus	UnimodID	Binary
Carbamidomethylation of Cysteine	57.02146	C	□	□	□	4	0
Oxidation of Methionine	15.9949	M	☑	□	□	35	0

Search Progress:

Fig. 2 | PSM. Predefined parameters are used in the 'Search and Filtering' tab. They can be edited in the 'Advanced Parameters' tab.

- If the experiment is, for example, a time-course experiment including five timepoints, appropriate names are T0, T1, T2, T3 and T4
- Within each biological condition directory, directories for each biological replicate should be created; for example, supposing there are three biological replicates for T0, appropriate directory names would be T0B1, T0B2 and T0B3
- Place all technical replicates for a biological replicate within the same directory
- If data generation relied on some chromatographic prefractionation, such as strong cation exchange (SCX) prior to LC/MS/MS, place all the SCX salt step MS data files within the appropriate biological replicate directory

  ▲ **CRITICAL STEP** Proper organization of the mass spectrometry files prior to PSM is essential to the success of the following steps.

8 Click on the 'Identify' menu for the search engine GUI (Fig. 2), then select the 'Search and Filtering' tab.

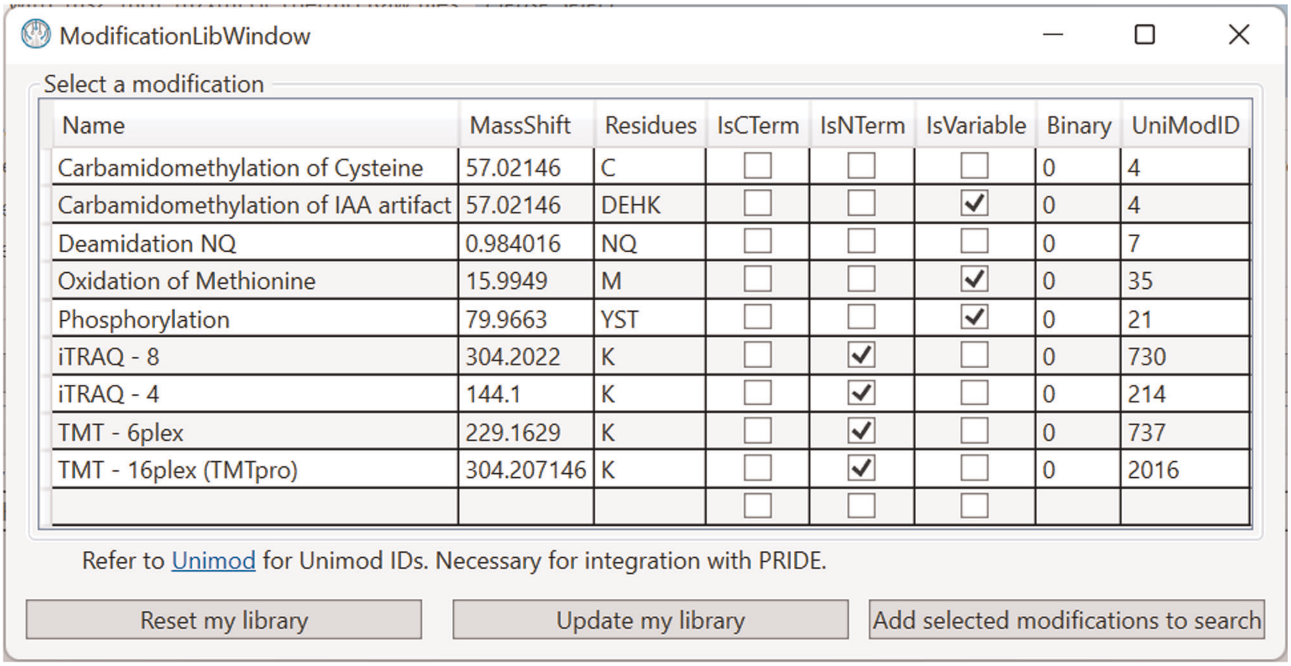
9 Click on the 'Browse' button to select 'Directory with.ms2,.mgf,.mzXML, mzML, or Thermo.raw files' and choose the directory containing the raw files (Thermo.raw files are preferred).

10 Click on the 'Browse' button for 'Sequence Database(s)', and choose a.fasta sequence database file.

11 (Optional) The software will automatically generate a target–decoy sequence database file (i.e., *.T-R) from the .fasta file. The user can optionally generate a custom target–decoy database. To do this, use the 'Generate search DB' module located in the 'Utils' menu as described in Step 64 of this protocol. The .T-R file can also be directly loaded for searching.

12 Under 'Preconfigured search parameters', check the appropriate option. These parameters must be selected according to the instrument used to generate experimental data.

- We recommend using 'High-High' for data from instruments that provide <20 ppm of mass accuracy and >15,000 resolution (e.g., the typical Q-Exactive setup)
- We recommend 'High-Low' for the typical setup of an Orbitrap-Velos, in which the MS1 are collected in the Orbitrap and the MS2 in the LTQ



Name	MassShift	Residues	IsCTerm	IsNTerm	IsVariable	Binary	UniModID
Carbamidomethylation of Cysteine	57.02146	C	☐	☐	☐	0	4
Carbamidomethylation of IAA artifact	57.02146	DEHK	☐	☐	☑	0	4
Deamidation NQ	0.984016	NQ	☐	☐	☐	0	7
Oxidation of Methionine	15.9949	M	☐	☐	☑	0	35
Phosphorylation	79.9663	YST	☐	☐	☑	0	21
iTRAQ - 8	304.2022	K	☐	☑	☐	0	730
iTRAQ - 4	144.1	K	☐	☑	☐	0	214
TMT - 6plex	229.1629	K	☐	☑	☐	0	737
TMT - 16plex (TMTpro)	304.207146	K	☐	☑	☐	0	2016
			☐	☐	☐		

Refer to Unimod for Unimod IDs. Necessary for integration with PRIDE.

Fig. 3 | PTM library. The PTM library allows selection of PTMs to be considered by the search engine. The user can also add new modifications to the library in this window.

- We recommend 'Low-Low' when both the MS1 and the MS2 are generated in a low-resolution instrument such as an LTQ

  ▲**CRITICAL STEP** These choices are critical to the experiment's success because failing to select settings according to the instrument used may provide suboptimal results.

13 Choose one of the 'Operation Mode' options.

- 'Search' will only execute the Comet search engine, providing a plain text file as the result output
- 'Filter (SEPro)' will filter Comet search results and generate the SEPro user-friendly interactive result files, with trustworthy identifications, necessary for downstream data analysis
- 'Search and Filter' is the default option as search results will typically need to be filtered to become useful

  ▲**CRITICAL STEP** Modifications used for searching should still be properly setup in the GUI, even if only using the filtering option, as they will be registered in the SEPro file; see Step 15.

14 Select 'Processing options'.

- 'Recursive directory processing' will allow all subdirectories to be searched. This is the typical setting as the user specifies the main directory with subdirectories for the biological conditions
- Select 'Include MS2 in SEPro filtered results' to include the annotated mass spectra of the identified peptides

  ▲**CRITICAL STEP** Choosing to annotate the spectra will make the results files substantially larger. For example, a results file for 12,241 identified MS2 with (without) annotated spectra has ~38 MB (5.1 MB).

15 Click on the 'Add modification' button to select a modification.

- Indicate whether the modification is variable (e.g., oxidation of methionine) or not (e.g., carbamidomethylation of cysteine) and, if applicable, to which of the two termini it applies (by checking the corresponding boxes)
- The 'Binary' option can also be used along with other modifications; this option is to specify when all residues are either modified or not modified. For example, when using iTRAQ, peptides are expected to contain the same label in the N-terminus as in the reacting aminoacids
- Other modifications can be added manually by including the required information in the last row and then clicking on the 'Update my library' button to save it (Fig. 3)

  ▲**CRITICAL STEP** When including a new modification, its Unimod ID is required to make possible a complete submission to the PRIDE Proteomics Repository. Unimod is the is the database of protein modifications for mass spectrometry. IDs can be retrieved from www.unimod.org.

Specifying parameters for Comet and SEPro ● Timing ~2 min

16 Clicking on the 'Advanced Parameters' tab allows search engine parameters to be specified for Comet (option A) and FDR filtering parameters to be specified for SEPro (option B). We note that, for high-resolution instruments, the 'Precursor mass accuracy' and 'Enzyme Specificity' have default values of 35 ppm and semi-specific, respectively; these may be considered, by some, as too relaxed. The motivation stems that most proteomic searches comprise single organism sequence databases that may not provide 'Sufficient Search Space' to guarantee robustness for estimating several search engine scores also discussed by Jimmy Eng and collaborators[46]. This notwithstanding, we recommend setting these parameters to 20 ppm and fully specific when searching sequence database containing more than one organism or more than 100,000 sequences. Note that, regardless, for high-resolution instruments, the final filtering steps as carried out by SEPro will narrow down, by default, the search tolerance to 10 ppm from the error mean.

(A) **Main search engine parameters for Comet**.

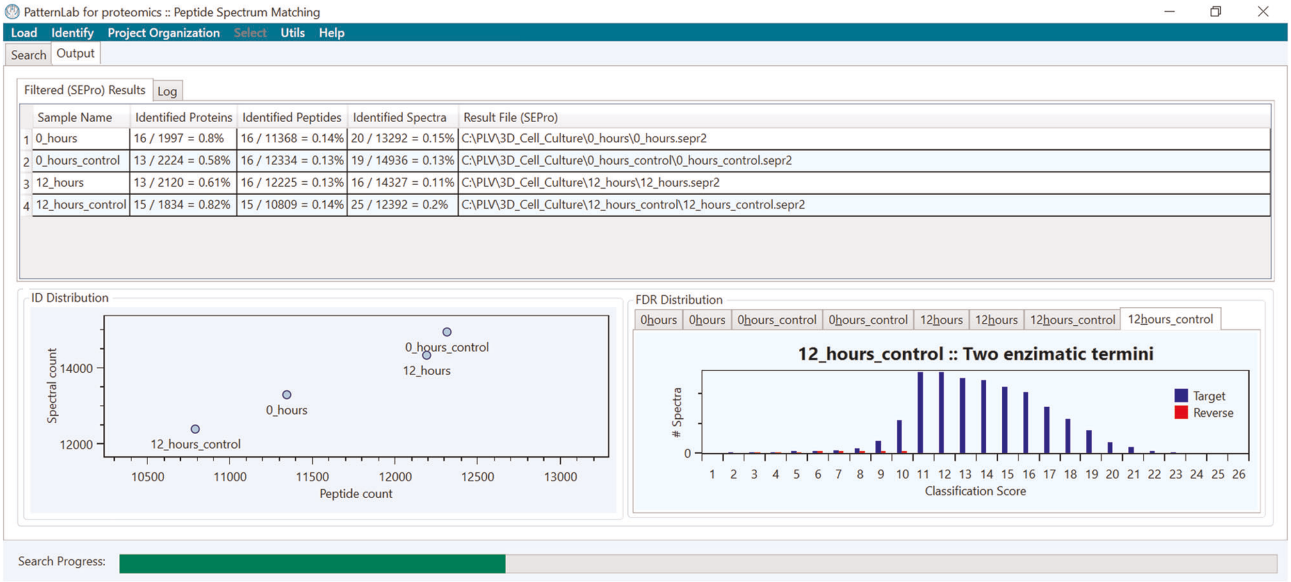
(i) Specify search engine parameters for Comet via the 'Advanced Parameters' tab. The reader is referred to http://comet-ms.sourceforge.net/parameters/parameters_202001/ for more specific parameters.

Precursor mass accuracy	This parameter is related to a +/− mass tolerance for selecting peptides whose masses match the precursor mass
Fragment bin tolerance	This parameter controls Comet's bin size associated with fragment ions related to an MS/MS peak and how it is stored internally in an array element by the search engine; this can be considered a proxy to a fragment ion tolerance. There is a correlation between this parameter and the memory used in the search; the lower the value, the higher the memory
Fragment bin offset	This parameter defines how each fragment bin starts; for example, for a 'Fragment bin tolerance' of 1, the bins would be 0.0, 1.0, ..., *n*, but with a 'Fragment bin offset' of 0.5, the bins would be 0.5, 1.5, ..., *n*. A typical configuration of 'Fragment bin tolerance' of 1.0005 with a 'Fragment bin offset' of 0.4 is used for LTQ data to avoid starting a bin in the middle of a mass spectral peak as a function of the mass defects
Clear *m/z* range (start/stop)	This parameter is typically used when searching iTRAQ/TMT data so the search engine ignores the MS/MS peaks within the specified range that are resultant from the reporter ions
Search mass range	This parameter defines the peptide mass range to search
Max. variable mods per peptide	This parameter specifies the maximum number of residues with modifications that the search engine will consider
Missed cleavages	Number of allowed missed enzyme cleavages for a peptide
Theoretical fragment ions	This parameter defines how Comet generates the theoretical mass spectrum peaks. The 'M peak only' option indicates that the fast correlation score, used by Comet, will be the sum of the intensities at each theoretical fragment mass bin; this option is preferred for LTQ data. The 'Use flanking peaks' option indicates that Comet's fast correlation score will be the sum of the intensities at each theoretical fragment mass bin and half the intensity of each flanking bin, optimal for high-resolution data
Enzyme	Used for specifying the enzyme used for digesting the sample. This parameter works in conjunction with 'Enzyme specificity' and 'Missed cleavages'
Enzyme specificity	This parameter specifies the number of enzyme termini that a peptide may have; for example, 'Fully specific' would require both termini to match the requirements of the 'Enzyme'
Ion series	Used to specify the ion series to be considered in the theoretical mass spectrum as well as neutral losses
Homo sapiens PEFF search	Used to extend the global amino acid variant and PTM analysis by using the PSI extended format[47,48]

(B) **Filtering parameters for SEPro**

(i) Choose the filtering parameters for SEPro considering the following guidelines:

- The 'Bayesian score' group box includes features provided by the search engine to generate a quadratic discriminant function to rank the hits
- The 'Pre-processing quality filters' group box lists threshold of scores used to eliminate search engine results that do not achieve a minimum quality; this filter is applied before generating the quadratic discriminant function



	Sample Name	Identified Proteins	Identified Peptides	Identified Spectra	Result File (SEPro)
1	0_hours	16 / 1997 = 0.8%	16 / 11368 = 0.14%	20 / 13292 = 0.15%	C:\PLV\3D_Cell_Culture\0_hours\0_hours.sepr2
2	0_hours_control	13 / 2224 = 0.58%	16 / 12334 = 0.13%	19 / 14936 = 0.13%	C:\PLV\3D_Cell_Culture\0_hours_control\0_hours_control.sepr2
3	12_hours	13 / 2120 = 0.61%	16 / 12225 = 0.13%	16 / 14327 = 0.11%	C:\PLV\3D_Cell_Culture\12_hours\12_hours.sepr2
4	12_hours_control	15 / 1834 = 0.82%	15 / 10809 = 0.14%	25 / 12392 = 0.2%	C:\PLV\3D_Cell_Culture\12_hours_control\12_hours_control.sepr2

Fig. 4 | Filtered results. Once the search begins, PLV will switch to the 'Output'-tab report. Identification information will be provided for each mass spectrometry file. A plot of 'Peptide count' versus 'Spectral count' will be displayed in the lower left-hand corner. Histograms showing the number of spectra ('# Spectra') versus 'Classification Score', for each file, will be displayed in the lower right-hand corner; data will be displayed independently for targets (blue) and decoys (red). SEPro results can be opened by clicking on the corresponding directory in the 'Result File (SEPro)' tab.

- The 'Acceptable False Discovery Rate Estimates' group box lists the acceptable FDRs at each level. SEPro uses a three-tier approach that filters IDs first at the spectral level, then by peptide, and lastly by protein level. SEPro uses a Gaussian discriminant analysis to rank identifications and converge to an FDR
- The 'Bayesian Score' group box reflects the features considered for the FDR: 'Normalized XCorr' (i.e., XCorr divided by peptide length), 'DeltaCN', 'DeltaMass' (ppm), 'SpecCount Score', 'Peaks Matched' and Comet's 'Secondary Score'. In this regard, specifying a 3% FDR at the spectral level does not imply that the final result will have this FDR; the three-tier approach is fully explained in the SEPro manuscript[7]
- The 'Grouping' group box provides options to group mass spectra according to charge state and by enzymatic specificity before applying the quadratic discriminant; for example, if the 'Group by charge state' option is specified, the PSMs will be separated into a group with 2+ and another with higher charge states, ultimately leading to estimating the FDR with a discriminant function for each group
- The 'Post-processing quality filters' group box provides filters to be applied after the FDR filter is applied. Here we draw attention to the 'Minimum Protein Score', which specifies a lower bound on the primary score (XCorr) for a protein to be accepted. This filter is critical, especially for filtering out one-hit-wonders that pass the FDR filter but have a low XCorr
- The 'Decoy Labels' group box specifies the header of Decoys and Unlabeled decoys in the sequence database. Unlabeled decoys are primarily used when benchmarking algorithms; more information is available[6]

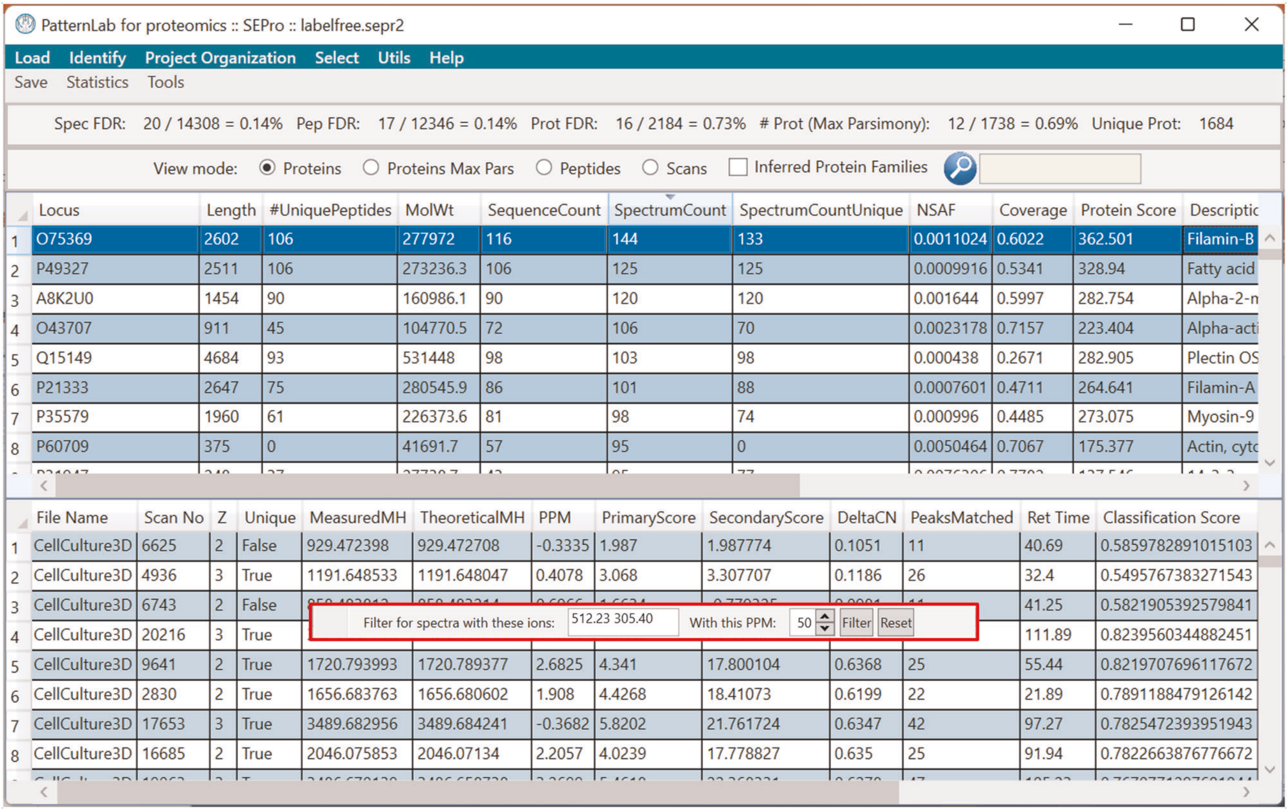
17 Click on the 'GO' button to do search and filtering.
▲**CRITICAL STEP** If the Visual C++ Runtime is not installed, no search engine results will be generated. See 'Software requirements'.

18 In the 'Filtered (SEPro) Results' tab (Fig. 4), click on the 'Result File (SEPro)' tab to access the identification results (Fig. 4). Each SEPro file will be automatically saved in the corresponding directory.

Interpretation of protein identification results ● Timing Variable; the process involves scientific interpretation that may require discussion and further research.

19 Access the SEPro results as described in Step 18 or by selecting the 'Load' option. Once the results are loaded, the SEPro GUI loads as exemplified in Fig. 5.

20 Click on a protein result (upper panel) to fill the lower panel (PSM panel) with data for all identifications related to that protein.



Spec FDR: 20 / 14308 = 0.14% Pep FDR: 17 / 12346 = 0.14% Prot FDR: 16 / 2184 = 0.73% # Prot (Max Parsimony): 12 / 1738 = 0.69% Unique Prot: 1684

View mode: ● Proteins ○ Proteins Max Pars ○ Peptides ○ Scans □ Inferred Protein Families

	Locus	Length	#UniquePeptides	MolWt	SequenceCount	SpectrumCount	SpectrumCountUnique	NSAF	Coverage	Protein Score	Descriptic
1	O75369	2602	106	277972	116	144	133	0.0011024	0.6022	362.501	Filamin-B
2	P49327	2511	106	273236.3	106	125	125	0.0009916	0.5341	328.94	Fatty acid
3	A8K2U0	1454	90	160986.1	90	120	120	0.001644	0.5997	282.754	Alpha-2-n
4	O43707	911	45	104770.5	72	106	70	0.0023178	0.7157	223.404	Alpha-acti
5	Q15149	4684	93	531448	98	103	98	0.000438	0.2671	282.905	Plectin OS
6	P21333	2647	75	280545.9	86	101	88	0.0007601	0.4711	264.641	Filamin-A
7	P35579	1960	61	226373.6	81	98	74	0.000996	0.4485	273.075	Myosin-9
8	P60709	375	0	41691.7	57	95	0	0.0050464	0.7067	175.377	Actin, cytc

	File Name	Scan No	Z	Unique	MeasuredMH	TheoreticalMH	PPM	PrimaryScore	SecondaryScore	DeltaCN	PeaksMatched	Ret Time	Classification Score
1	CellCulture3D	6625	2	False	929.472398	929.472708	-0.3335	1.987	1.987774	0.1051	11	40.69	0.5859782891015103
2	CellCulture3D	4936	3	True	1191.648533	1191.648047	0.4078	3.068	3.307707	0.1186	26	32.4	0.5495767383271543
3	CellCulture3D	6743	2	False								41.25	0.5821905392579841
4	CellCulture3D	20216	3	True								111.89	0.8239560344882451
5	CellCulture3D	9641	2	True	1720.793993	1720.789377	2.6825	4.341	17.800104	0.6368	25	55.44	0.8219707696117672
6	CellCulture3D	2830	2	True	1656.683763	1656.680602	1.908	4.4268	18.41073	0.6199	22	21.89	0.7891188479126142
7	CellCulture3D	17653	3	True	3489.682956	3489.684241	-0.3682	5.8202	21.761724	0.6347	42	97.27	0.7825472393951943
8	CellCulture3D	16685	2	True	2046.075853	2046.07134	2.2057	4.0239	17.778827	0.635	25	91.94	0.7822663876776672

Filter for spectra with these ions: 512.23 305.40 With this PPM: 50

Fig. 5 | SEPro's result browser. The upper panel lists protein identifications. The software reports identifications at the spectral, peptide and protein level. For the protein level, the results containing redundancy are displayed to the right of 'Prot FDR'. 'Prot (Max Parsimony)' displays the minimum number of proteins required to explain all peptides, and 'Unique Prot' the number of proteins containing at least one unique peptide. Clicking on any protein identification causes the lower panel to display all PSM data associated with the corresponding protein, along with the respective scores.

21 Right-click on the PSM panel to shortlist identifications related to spectra that contain a given *m/z* peak (Fig. 5, red box) (Note: this feature is available only if 'Include MS2 in SEPro filtered results' was previously selected as described in Step 14).

22 Double-click on a protein result to generate an overlay a window with a graphical and a FASTA representation of the protein's coverage and its domains inferred through the Pfam[49] or SuperFamily[50] (also known as Supfam) and a heatmap whose colors reflect the numbers of peptides shared among proteins (Fig. 6).

▲CRITICAL STEP 'Peptide Shared Matrix' appears only for proteins that contain shared peptides.

23 Double-click on a row in the lower panel (in the 'Proteins' view mode) to view the annotated mass spectrum and matching ion table pop-up (Fig. 7). This option will be available only if the user checked the 'Include MS2 in SEPro filtered results' box when performing the search.

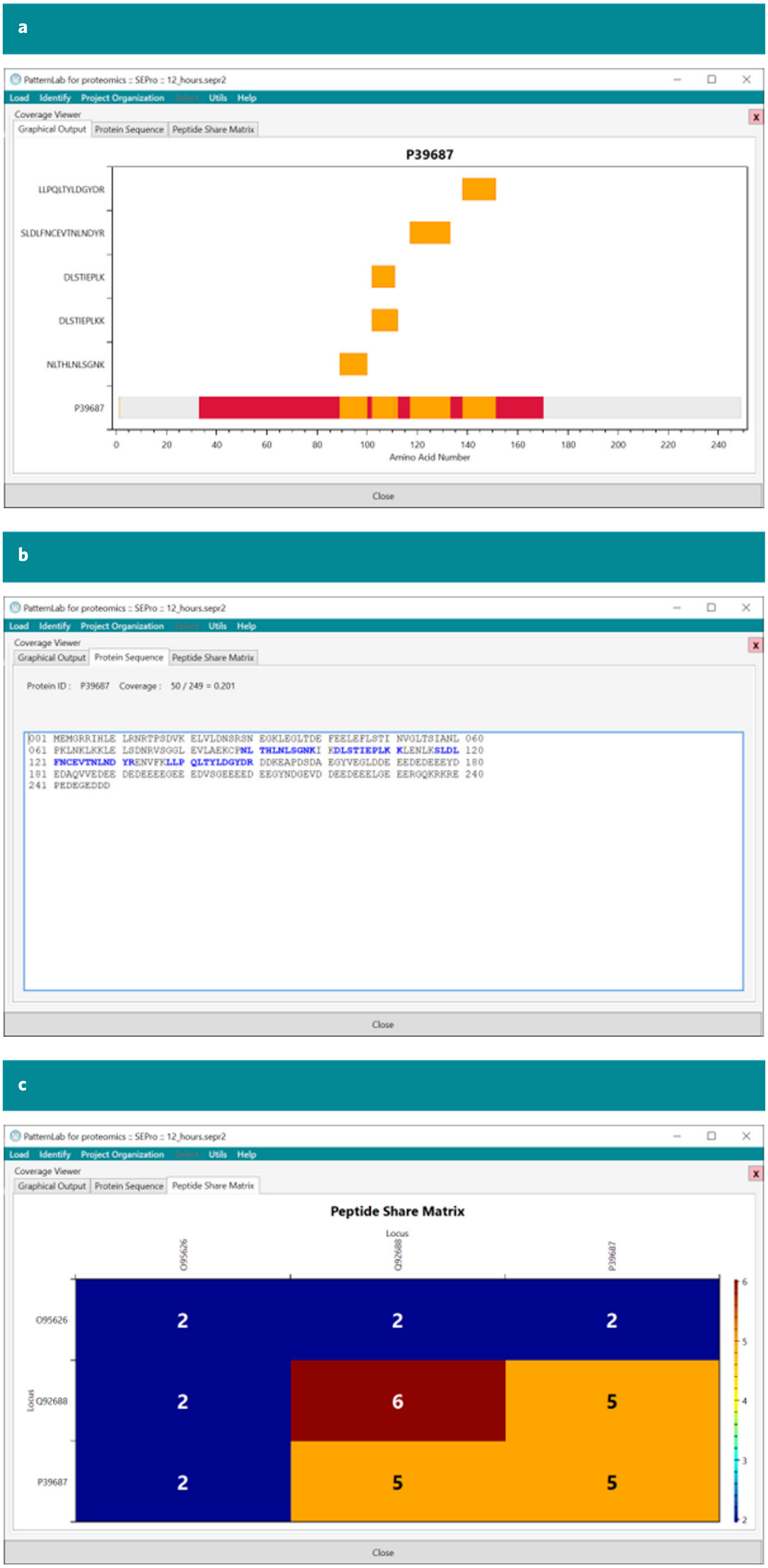
24 Click on the 'Statistics' menu and then select 'Charge State Distribution' or 'PPM Distribution' to display a histogram of the selected distribution.

▲CRITICAL STEP 'PPM Distribution' is only for data acquired with MS1 in high resolution.

25 Clicking on the 'Tools' menu and then on 'Evaluation of Enzyme Specificity' will display a window informing how many fully specific and semi-specific peptides were identified in the mixture.

26 The user can select 'Proteins', 'Proteins Max Pars', 'Peptides' or 'Scans' as view mode and then use the search box to look for a specific protein, peptide or scan number, respectively.

- The user could choose 'Inferred Protein Families' instead of 'Proteins Max Pars' (maximum parsimony)
- When the 'Inferred Protein Families' box is checked, PLV will display additional columns in the 'Proteins' view mode panel: 'Family', 'Family Type' and 'Family Spectral Count'. These options are typically used when, for example, one wishes to perform spectral count quantitation by family type, as in snake venomics studies
- Note that PLV groups proteins into families if proteins share at least one common identified peptide



Protein ID : P39687 Coverage : 50 / 249 = 0.201

```
001 MEMGRRIHLE LRNRTPSDVK ELVLDNSRSN EGKLEGLTDE FEELEFLSTI NVGLTSIANL 060
061 PKLNKLKKLE LSDNRVSGGL EVLAEKCPNL THLNLSGNKI KDLSTIEPLK KLENLKSLDL 120
121 FNCEVTNLND YRENVFKLLP QLTYLDGYDR DDKEAPDSDA EGYVEGLDDE EEDEDEEEYD 180
181 EDAQVVEDEE DEDEEEEGEE EDVSGEEEED EEGYNDGEVD DEEDEEELGE EERGQKRKRE 240
241 PEDEGEDDD
```

	O95626	Q92688	P39687
O95626	2	2	2
Q92688	2	6	5
P39687	2	5	5

◀ **Fig. 6 | Double-clicking on a protein result. a**, Graphical output of the protein sequence (gray) and peptide identifications (yellow), and automatic domain inferring (red). **b**, The identified protein's sequence with blue letters representing the identification coverage. **c**, 'Peptide Share Matrix' indicates how many identified peptides there are in common when comparing identifications of proteins belonging to the same group.

- The user can click on a family number and change it manually to make the protein belong to another family
- If the user alters the family numbering, selecting 'Save', then 'Save SEPro files', saves the changes
- 'Maximum Parsimony' refers to a minimum set of proteins that can explain all identified peptides
- In general, the number of proteins abiding by the maximum parsimony rule is expected to be larger than the number of families

Project organization and quantitation ● Timing <20 min for quantitation in a typical proteomic experiment with ~20 raw files

▲CRITICAL One of the goals of proteomics is the study of differences in protein abundance throughout different biological states. In this regard, the user must specify which samples come from which biological condition, biological replicate or experimental timepoint (in time-course experiments). It is essential that all files be previously arranged as described in Step 7. Project organization aims to compile all the experimental data into the single file required for downstream analysis.

The following steps describe the procedure for label-free data (Fig. 8). To look at data from labeled experiments, refer to Box 1 and Fig. 9.

27 Click on the 'Project Organization' menu to specify whether the quantitation will be performed on labeled or label-free data.

28 Click on the 'Add Directory' button, and select the directory for each biological condition.
▲CRITICAL STEP Directories must contain results filtered by SEPro. At least two directories must be included.

29 Decide whether quantitation is to be performed by XICs (option A) or spectral counting (option B). We recommend using XICs for single shots and spectral counting for experiments that use some kind of prefractionation step.

(A) **Parameters for XICs**

(i) 'PPM Tolerance' sets the mass tolerance value for quantitation.

(ii) 'MS1 Tolerance' allows the software to skip a failed MS1 and still properly extract XICs.

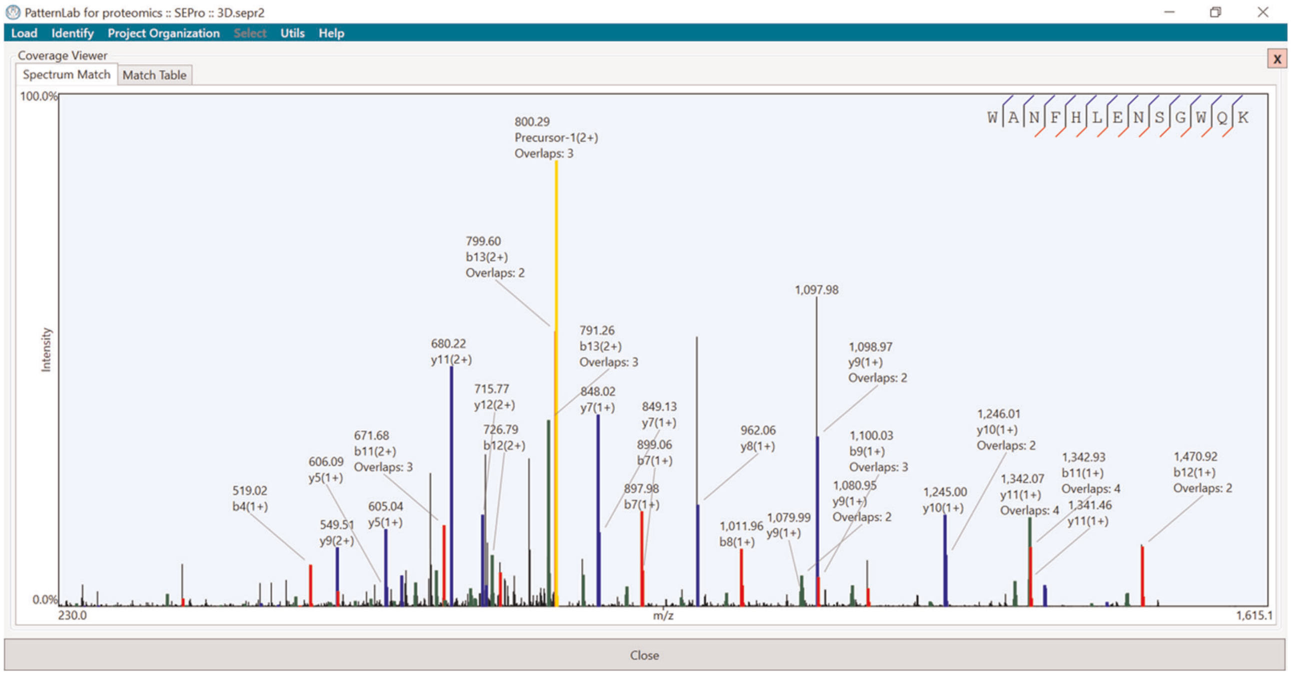
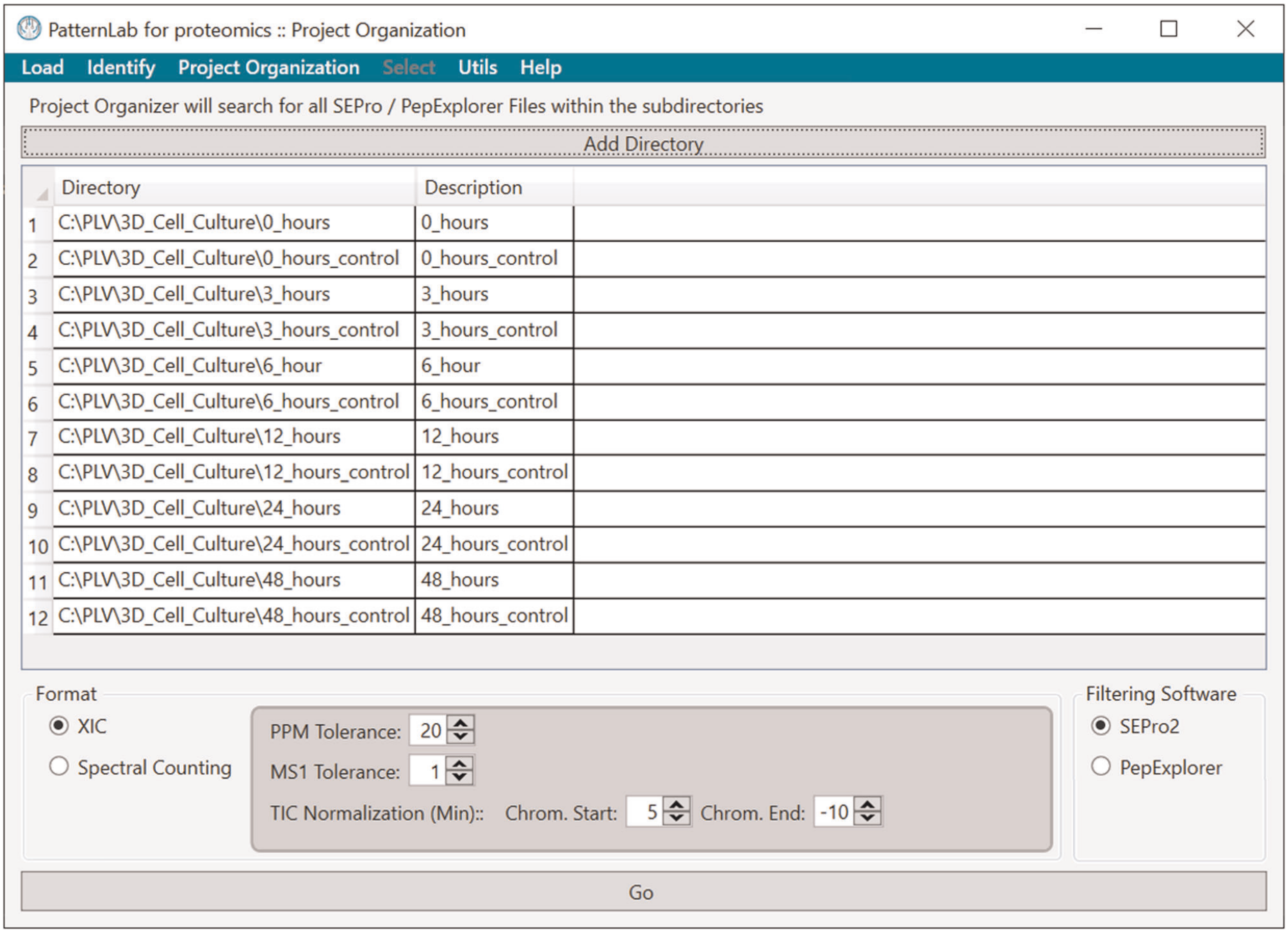


Fig. 7 | Annotated mass spectrum. The yellow, red, blue and green peaks represent precursor ion, the b series, the y series and neutral losses, respectively.



Project Organizer will search for all SEPro / PepExplorer Files within the subdirectories

	Directory	Description
1	C:\PLV\3D_Cell_Culture\0_hours	0_hours
2	C:\PLV\3D_Cell_Culture\0_hours_control	0_hours_control
3	C:\PLV\3D_Cell_Culture\3_hours	3_hours
4	C:\PLV\3D_Cell_Culture\3_hours_control	3_hours_control
5	C:\PLV\3D_Cell_Culture\6_hour	6_hour
6	C:\PLV\3D_Cell_Culture\6_hours_control	6_hours_control
7	C:\PLV\3D_Cell_Culture\12_hours	12_hours
8	C:\PLV\3D_Cell_Culture\12_hours_control	12_hours_control
9	C:\PLV\3D_Cell_Culture\24_hours	24_hours
10	C:\PLV\3D_Cell_Culture\24_hours_control	24_hours_control
11	C:\PLV\3D_Cell_Culture\48_hours	48_hours
12	C:\PLV\3D_Cell_Culture\48_hours_control	48_hours_control

Format
● XIC
○ Spectral Counting

PPM Tolerance: 20
MS1 Tolerance: 1
TIC Normalization (Min):: Chrom. Start: 5 Chrom. End: -10

Filtering Software
● SEPro2
○ PepExplorer

Fig. 8 | Project Organization. This module is responsible for joining all the information from the various biological or technical replicates of all biological conditions.

Box 1 | Quantitation analyses with labeled data

PLV simplifies the setting of parameters for the analysis of multiplex experiments labeled with isobaric tags, taking into account only the information essential to run the quantitation (Fig. 9). PLV's Isobaric Analyzer is tailored toward experiments in which all biological conditions and replicates are labeled and contained within the same analysis. For example, a three-condition experiment with four biological replicates per condition would require 12 isobaric markers. If the results are unsatisfactory, increased coverage can be achieved by using online or offline fractionation as previously described[58] and including all mass spectrometry data within the same directory for analysis.

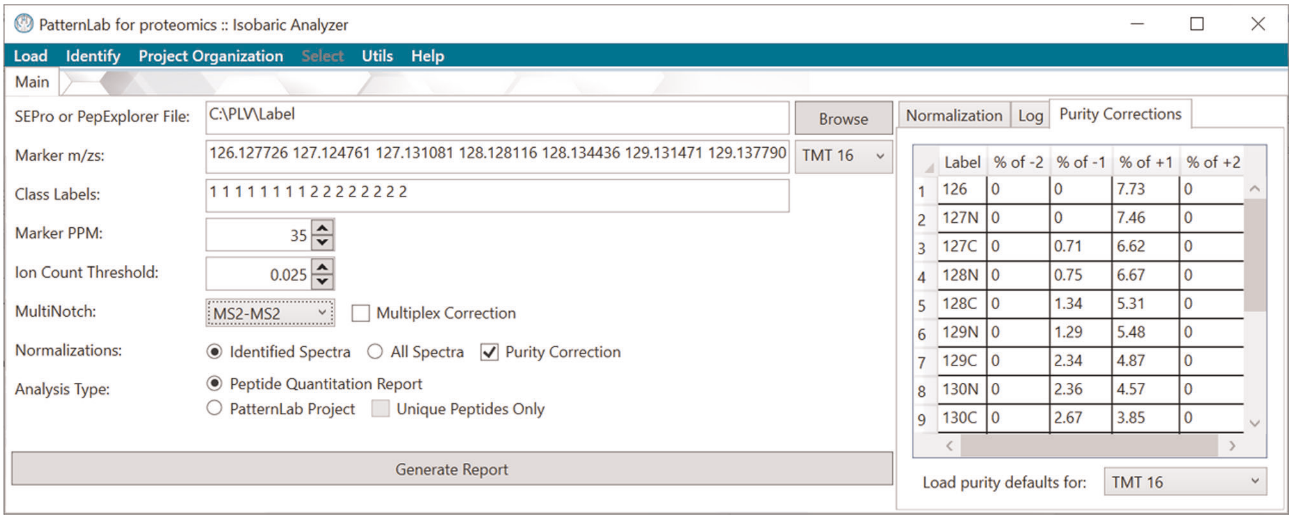
- Set the default parameters, such as 'SEPro or PepExplorer File' and the reporter ions' masses of the tags used (e.g., iTRAQ or TMT)
- Check the 'Multiplex Correction' if intending to remove multiplexed spectra
- Set the 'MultiNotch' method if the data were acquired in stepped-CID-HCD or SPS-MS3 acquisition mode
- Select a normalization algorithm. The normalization options include 'Identified Spectra' or 'All Spectra' and, optionally, 'Purity Correction'. The first two options will consider the summed ion signals of each marker as a normalization factor, while the third will consider correction factors reported in the isobaric labeling kit to account for isotopic impurities that cause cross-talk between channels
- For all the remaining parameters, follow the same pipeline described in the previous protocol[27]
- Save the results in the Peptide Quantitation Report or the PatternLab Project format. The former enables one to compare two biological states with a specialized module, as per the previous protocol[27]; the latter enables one to access all of the downstream bioinformatics analysis modules described in this protocol, including Venn diagrams, volcano plots, clustergrams and PCA

(iii) 'Chrom. Start' and 'Chrom. End' set the range (min) for extracting XIC data; for example, setting the parameters to 5 and −5, respectively will make the software ignore the first and last 5 min of the chromatography.

▲**CRITICAL STEP** The 'Chrom. End' negative (−) sign must be included.

(B) **Options for spectral counting**

(i) Select 'Proteins', 'Peptides' or 'Unique peptides' to choose which type of information the analysis will comprise.



SEPro or PepExplorer File: C:\PLV\Label

Marker m/zs: 126.127726 127.124761 127.131081 128.128116 128.134436 129.131471 129.137790 TMT 16

Class Labels: 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2

Marker PPM: 35

Ion Count Threshold: 0.025

MultiNotch: MS2-MS2 □ Multiplex Correction

Normalizations: ● Identified Spectra ○ All Spectra ☑ Purity Correction

Analysis Type: ● Peptide Quantitation Report ○ PatternLab Project □ Unique Peptides Only

	Label	% of -2	% of -1	% of +1	% of +2
1	126	0	0	7.73	0
2	127N	0	0	7.46	0
3	127C	0	0.71	6.62	0
4	128N	0	0.75	6.67	0
5	128C	0	1.34	5.31	0
6	129N	0	1.29	5.48	0
7	129C	0	2.34	4.87	0
8	130N	0	2.36	4.57	0
9	130C	0	2.67	3.85	0

Load purity defaults for: TMT 16

Fig. 9 | Isobaric Analyzer. The Isobaric Analyzer enables analysis of experiments typically labeled with iTRAQ or TMT.

(ii) Optionally, select ‘Use NSAF’ for normalized spectral abundance factor (NSAF)[51]. ‘Use Maximum Parsimony’ to consider only a minimum set of proteins that explains all peptides.

(iii) Check ‘Only unique peptides’ to consider peptides that belong to only one protein.

30 Go to ‘Filtering Software’ to specify which software was used to filter (SEPro or PepExplorer).

31 Click on the ‘Go’ button to begin performing the quantitation. A pop-up screen will ask where to save the .xic2 file containing all the XIC information of the project.
▲CRITICAL STEP Raw files must be available in the directories.

Quantitation analysis with XIC ● Timing ~5–10 min as this is an exploratory module

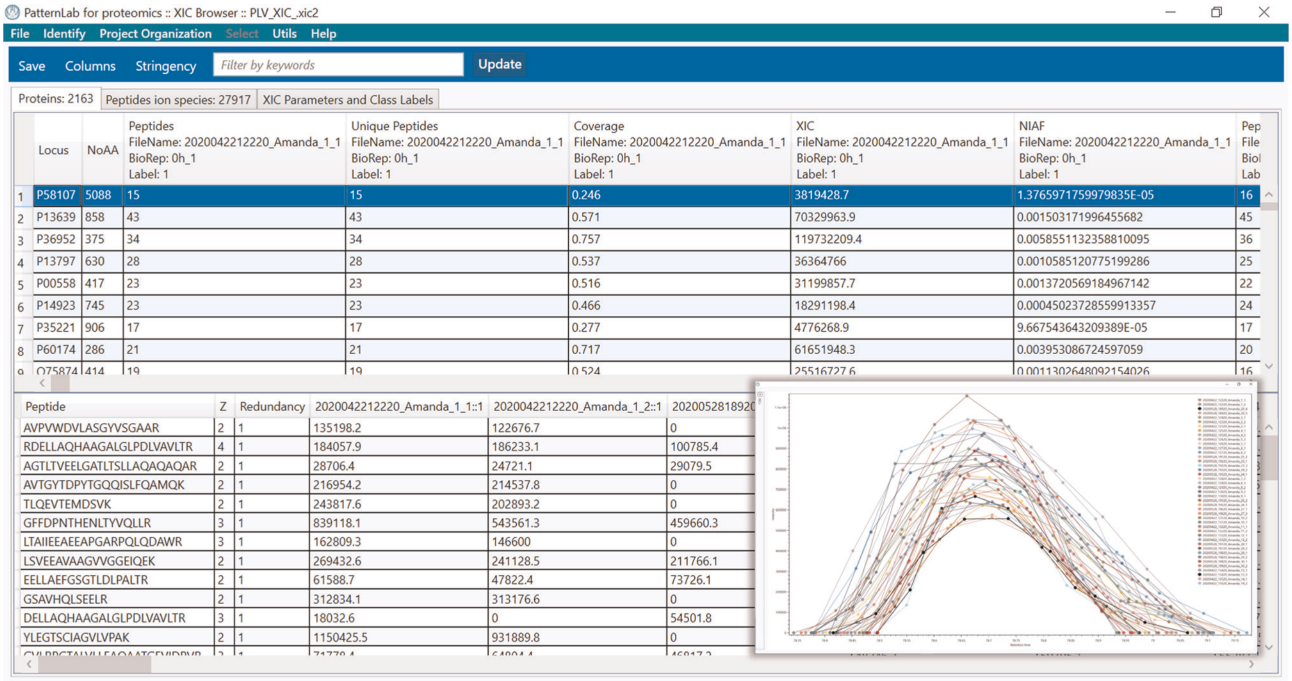
32 Once XIC extraction is finished, PLV will automatically open the XIC Browser (Fig. 10). The .xic2 files can also be opened by clicking on ‘Load’.

33 The ‘Stringency’ menu enables one to specify several parameters to filter the quantified data, as shown below. Once done, click on the ‘Update’ button to apply the modifications.

Maximum Parsimony	This parameter makes the software consider only a minimum set of proteins that explains all peptides
Only Unique Peptides	This parameter will remove all peptides that are not unique to a single protein sequence
Included Contaminants	Indicates whether to consider contaminant identifications in the results
Included Decoys	Indicates whether to include decoy identifications in the results
Consider Only Preferential Charge State	When enabled, PLV will consider, for each peptide sequence, only the charge state that yielded the highest average XIC throughout the experiment
Min. Number of Unique Peptides	The software will consider only proteins with at least the minimum number of unique peptides the parameter specifies
Min. Number of Peptides	The software will consider only proteins that have at least the minimum number of peptides the parameter specifies
Min. Number of MS1 Count	The software will consider only XICs that contain at least the minimum number of MS1 readings the parameter specifies

34 Click on ‘Columns’ to select the information to be displayed per protein:
- ‘Peptide’: total number of peptides quantitated for a protein
- ‘Unique Peptide’: total number of unique peptides quantitated per protein
- ‘Coverage’: the percentage of identified sequence residues
- ‘XIC’: the sum of peptide quantitation by XIC
- ‘NIAF’: the protein’s XIC is divided by the sum of all XICs in the mass spectrometry run

35 Click on a row displaying protein information for information on its peptides to be displayed in the panel below. By double-clicking on a quantitation value for a peptide, its XIC will be displayed (Fig. 10). All XICs related to that peptide will be displayed by double-clicking on a peptide sequence.


PatternLab for proteomics :: XIC Browser :: PLV_XIC_.xic2
File
Identify
Project Organization
Select
Utils
Help
Save
Columns
Stringency
Filter by keywords
Update
Proteins: 2163
Peptides ion species: 27917
XIC Parameters and Class Labels

	Locus	NoAA	Peptides FileName: 2020042212220_Amanda_1_1 BioRep: 0h_1 Label: 1	Unique Peptides FileName: 2020042212220_Amanda_1_1 BioRep: 0h_1 Label: 1	Coverage FileName: 2020042212220_Amanda_1_1 BioRep: 0h_1 Label: 1	XIC FileName: 2020042212220_Amanda_1_1 BioRep: 0h_1 Label: 1	NIAF FileName: 2020042212220_Amanda_1_1 BioRep: 0h_1 Label: 1	Pep File Biol Lab
1	P58107	5088	15	15	0.246	3819428.7	1.3765971759979835E-05	16
2	P13639	858	43	43	0.571	70329963.9	0.001503171996455682	45
3	P36952	375	34	34	0.757	119732209.4	0.0058551132358810095	36
4	P13797	630	28	28	0.537	36364766	0.0010585120775199286	25
5	P00558	417	23	23	0.516	31199857.7	0.0013720569184967142	22
6	P14923	745	23	23	0.466	18291198.4	0.00045023728559913357	24
7	P35221	906	17	17	0.277	4776268.9	9.667543643209389E-05	17
8	P60174	286	21	21	0.717	61651948.3	0.003953086724597059	20
9	O75874	414	19	19	0.524	25516727.6	0.0011302648092154026	16

Peptide	Z	Redundancy	2020042212220_Amanda_1_1::1	2020042212220_Amanda_1_2::1	202005281892
AVPVWDVLASGYVSGAAR	2	1	135198.2	122676.7	0
RDELLAQHAAGALGLPDLVAVLTR	4	1	184057.9	186233.1	100785.4
AGTLTVEELGATLTSLLAQAQAQAR	2	1	28706.4	24721.1	29079.5
AVTGYTDPYTGQQISLFQAMQK	2	1	216954.2	214537.8	0
TLQEVTEMDSVK	2	1	243817.6	202893.2	0
GFFDPNTHENLTYVQLLR	3	1	839118.1	543561.3	459660.3
LTAIIEEAEEAPGARPQLQDAWR	3	1	162809.3	146600	0
LSVEEAVAAGVVGGEIQEK	2	1	269432.6	241128.5	211766.1
EELLAEFGSGTLDLPALTR	2	1	61588.7	47822.4	73726.1
GSAVHQLSEELR	2	1	312834.1	313176.6	0
DELLAQHAAGALGLPDLVAVLTR	3	1	18032.6	0	54501.8
YLEGTSCIAGVLVPAK	2	1	1150425.5	931889.8	0
[illegible]	2	1	71770.4	64804.4	46817.2

Fig. 10 | XIC Browser. The upper panel lists protein identifications; clicking on a protein identification makes the lower panel display all identified peptides associated with it along with XIC values. Double-clicking on an XIC value will open an XIC plot, the *x* axis referring to minutes and the *y* axis to ion current. The area under the curve can then be used as a surrogate for a peptide's relative abundance in the mixture and provide a basis for comparison against the XIC of the same peptide in different mixtures.

36 Click on 'Save' to compile the XIC data for downstream bioinformatics analysis. The options for exporting the .plp file are shown below:

Export XIC, Export NIAF	'Export XIC' will cause the XIC values to be exported. 'Export NIAF' will normalize the XICs according to the NSAF procedure[51] before exporting. We recommend 'Export XIC' for most experiments
Normalize XIC by TIC	We recommend using this option when 'Export XIC' is chosen. Data will be normalized according to the TIC from the chromatogram. The normalization factors can be viewed by clicking on the 'XIC Parameters and Class Labels' tab
Proteins, Peptides	Specifies whether the information exported will be protein- or peptide-based
Independent, Technical Replicates, MudPIT	Choose 'Independent' if each mass spectrometry run is to be considered independently of the others. 'Technical Replicates', used most, will merge the information from all technical replicates in the same directory, as per Step 7, by averaging nonzero quantitation values; this minimizes the effect of undersampling. 'MudPIT' will likewise sum up the quantitation values
Excel	Save a spreadsheet with proteins, peptides, params and class labels
XML	Save the quantitation data in the XML format. This may be useful for other software programmed in C# to read the data, using the DataTable. ReadXML method

37 For experiments that comprise samples of mixtures of organisms, such as analyzing *Trypanosoma cruzi* in *Mus musculus* plasma[52], it might be necessary to obtain the proportions of XICs for each species in each analysis for normalization purposes. This is achievable by clicking on the 'Mixture Analysis' tab.

Shortlisting proteins of interest from the data ● Timing Variable; the process involves scientific interpretation that may require discussion and further research

38 Once proteins or peptides have been relatively quantitated, the next step is to draw biological conclusions by comparing proteomic profiles and by shortlisting proteins or peptides with a

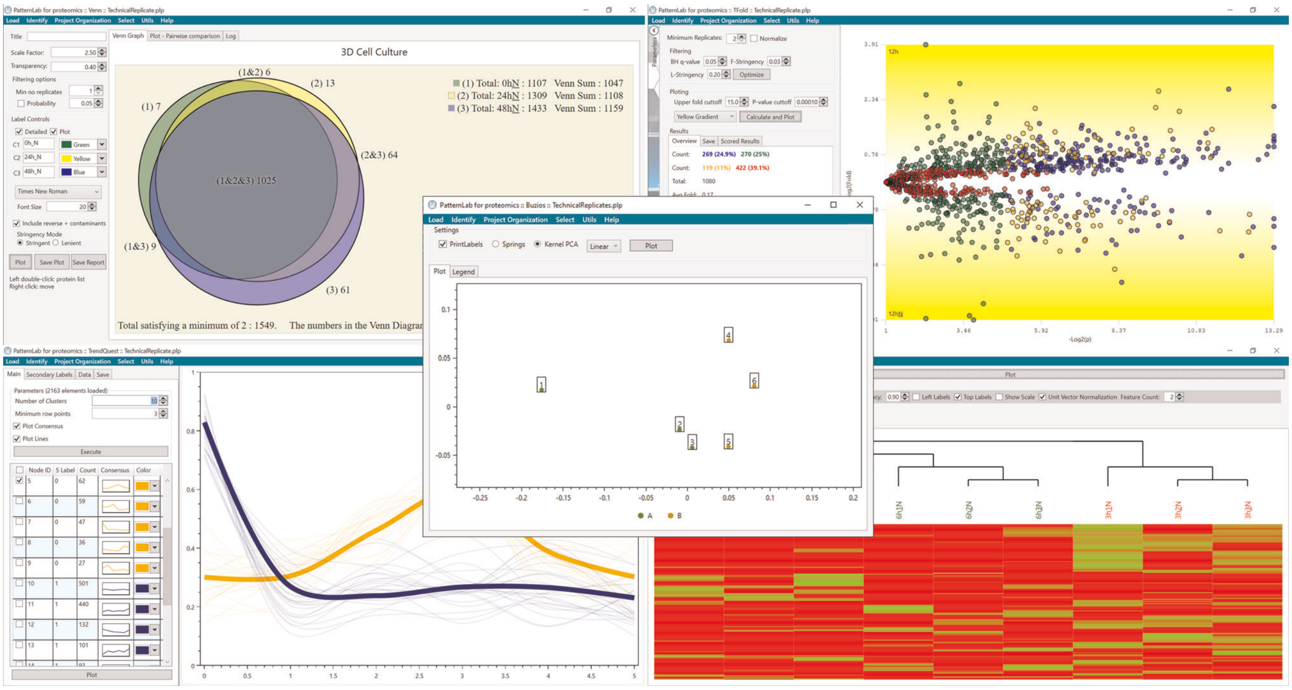

PatternLab for proteomics :: Venn :: TechnicalReplicate.plp
Load Identify Project Organization Select Utils Help
3D Cell Culture
(1&2) 6
(2) 13
(1) 7
(2&3) 64
(1&2&3) 1025
(1&3) 9
(3) 61
(1) Total: 0hN : 1107 Venn Sum : 1047
(2) Total: 24hN : 1309 Venn Sum : 1108
(3) Total: 48hN : 1433 Venn Sum : 1159
Total satisfying a minimum of 2 : 1549. The numbers in the Venn Diagram
PatternLab for proteomics :: TFold :: TechnicalReplicate.plp
Calculate and Plot
PatternLab for proteomics :: Buzios :: TechnicalReplicates.plp
PrintLabels
Springs
Kernel PCA
Linear
Plot
Legend
A
B
PatternLab for proteomics :: TrendQuest :: TechnicalReplicate.plp
Execute
Plot

Fig. 11 | Center panel (Buzios). Each numbered rectangle in the plot stands for a biological replicate, the dot within standing for the corresponding biological condition. The legend at the bottom ties dot colors to biological conditions. Upper left panel (Area-proportional Venn Diagram): GUI of the Area-proportional Venn Diagram module. Upper right panel (TFold): each dot in the plot represents a mapped protein according to its $-\log_2(P$ value), on the x axis, and $\log_2$(fold change), on the y axis. Red dots are proteins that satisfy neither the fold-change cutoff nor the q-value cutoff. Green dots are proteins that satisfy the fold-change cutoff but not the q-value cutoff. Orange dots are proteins that satisfy both the fold-change cutoff and q-value cutoff but have very low quantitation values; these results should be disregarded when using spectral counting. Finally, the blue dots are proteins that satisfy all statistical filters and are considered statistically differentially abundant. Lower right panel (Clustergram): a clustergram is generated by applying hierarchical clustering to the proteomic profiles extracted. Color encoding ranges from red, for a low-intensity signal, to green, for a high-intensity signal. Lower left panel (TrendQuest): proteins with similar quantitation profiles are grouped. The plot shows two of the clusters, each with a thick line representing the normalized average protein profile within it. The thin lines represent the protein profiles making up the two clusters.

statistical difference in their abundance means when comparing different biological conditions. Five modules (Fig. 11), namely Buzios, Area-proportional Venn Diagram, TFold, Clustergram and TrendQuest, have been tailored for this purpose. All have as input the .plp file generated by the options in the XIC Browser or Isobaric Analyzer. To unlock these modules, click ‘Load’ and load the .plp file; this will enable the ‘Select’ menu in the main GUI.

- The Buzios module aims to employ dimensional reduction strategies, such as PCA and multidimensional scaling (Springs), to represent each biological replicate in a two-dimensional plot and help provide a bird’s-eye view for determining global patterns within the samples
- The Area-proportional Venn Diagram module can list proteins identified only for a given biological condition
- On the other hand, the TFold module aims to pinpoint differentially abundant proteins and, therefore, complements the Area-proportional Venn Diagram module as it analyzes proteins common to both conditions but with statistical changes in their mean quantitation. Briefly, given an FDR bound specified by the user, the TFold module uses the Benjamini–Hochberg FDR estimator[53] to maximize the number of identifications that satisfy a fold-change cutoff that varies with the t-test P value as a power law and a stringency criterion that aims to detect lowly abundant proteins[21]
- The Clustergram module employs unsupervised clustering to group samples and provides a dendrogram with a heatmap; the user can verify whether samples clustered as expected or identify unanticipated patterns in the data
- Finally, the TrendQuest module is helpful particularly when analyzing proteomic experiments with more than two biological conditions. It helps group proteins that present similar expression profiles throughout the biological conditions

Obtaining a bird's-eye view with the Buzios module ● Timing ~5–10 min

▲CRITICAL The Buzios module (Fig. 11, center panel) is available in the 'Select' menu after clicking on 'Load' to load a .plp file. At least two biological replicates must be included in the .plp file.

39 Referring to Fig. 11 (center panel), make selections in 'Settings':
- 'Springs' for the multidimensional scaling algorithm
- 'Kernel PCA', followed by the Kernel function of choice (e.g., 'Gaussian', 'Linear', …, 'Spline'), for the PCA algorithm

40 Click on the 'Plot' button to generate the plot according to the settings.

Differential proteomics with the Area-proportional Venn Diagram module ● Timing ~5–10 min

▲CRITICAL The Area-proportional Venn Diagram module (Fig. 11, top left panel) is made available in the 'Select' menu after clicking on 'Load' to load a .plp file. At least two conditions must be included in the .plp file.

41 Filtering options:
- 'Min no replicates' defines the minimum number of biological replicates where the protein needs to have been detected
- 'Probability': refer to more detailed literature[22] for an in-depth explanation of how this filter works

  ▲CRITICAL STEP The choice of filtering options will affect behavior substantially. The user should specify a 'Min no replicates' to fix a minimum number of biological replicates in which a protein is to appear if it is to be considered as belonging to a biological condition.

42 Select 'Stringent' or 'Lenient' for 'Stringency Mode'.

▲CRITICAL STEP The 'Stringent' mode requires that all other conditions have zero replicates in which the protein appears. In this regard, the numbers in the Area-proportional Venn Diagram for a given condition may not add up to the number reported for that condition in the upper right-hand corner (Fig. 11, top left panel). On the other hand, 'Lenient' will allow proteins listed in other biological conditions to be considered even if appearing in fewer replicates than indicated by 'Min no replicates'. For example, if one specifies 3 for 'Min no replicates' and a given protein appears in three biological replicates for biological conditions A and B, but only in two biological replicates for condition C, this protein will be listed as belonging to the intersection of conditions A and B. Such a protein would be disregarded if 'Stringent' were chosen instead.

43 Click on the 'Plot' button.

▲CRITICAL STEP This module attempts to position each result label at the best possible location automatically, but sometimes the user might prefer repositioning labels. To reposition a label, right-click on it and drag it to the desired location.

44 Obtain the proteins for each section of the Area-proportional Venn Diagram by double-clicking on the corresponding numbers.

45 Generate an Excel report listing all biological conditions by clicking on the 'Save Report' button.

Differential proteomics with the TFold module for label-free data ● Timing ~5–10 min

▲CRITICAL The TFold module (Fig. 11, top right panel), used for pinpointing differentially abundant proteins, is made available in the 'Select' menu after clicking on 'Load' to load a .plp file. Briefly, for a user-specified FDR bound, the software uses the Benjamini–Hochberg FDR estimator to maximize the number of identifications that satisfy both a fold-change cutoff that varies with the *t*-test *P* value as a power law and a stringency criterion that aims to detect lowly abundant proteins. Further details are available in the module's publication[21]. 'Minimum Replicates' defines the minimum number of biological replicates in which the protein must appear to be considered in the analysis. At least two biological replicates are required. At least two conditions must be included in the .plp file.

46 Use the 'BH q-value' to determine the acceptable *P*-value cutoff after the Benjamini–Hochberg correction for the FDR.

47 Use 'F-Stringency' to define the variable fold-change cutoff. Higher values make for more stringent behavior.

▲CRITICAL STEP The optimal 'F-Stringency' can be obtained by clicking on the 'Optimize' button.

48 Use 'L-Stringency' to highlight (in orange) proteins that have low XICs or spectral counts and are more likely to yield false-positive differentially abundant proteins. Higher values make for more stringent behavior.

Box 2 | PatternLab Utils and Help Menus

Several modules are available under the Utils and Help Menus that are of broad applicability. The Utils menu allows one to access the 'Generate Search DB', 'Convert Thermo.Raw Files' and SEPro2 Fusion and Profiler options.

- The 'Generate Search DB' is optional, as the search module automatically generates the file with default settings. Here the user can, for example, modify the default list of contaminants or combine databases
- The 'Convert Thermo.Raw Files' module converts Thermo files to community standard formats such as MGF or MS2. The 'SEPro2' options allow one to merge results from various SEPro files into a single file, browse results from multiple SEPro files on-screen (profile mode) and even generate the required files for a complete submission to PRIDE
- The 'Help' menu provides information about the authors, links to this publication on the *Nature Protocols* website and links to our user forum where users can post questions

49 Use 'Upper fold cutoff' and 'P-value cutoff' to adjust the *y*-axis (fold change) and *x*-axis (*P*-value) cutoffs of the plot.

50 Click on 'Calculate and Plot' to refresh the plot.

51 Save the results to text files with the options made available in the 'Save' tab.

52 The 'Scored Results' tab provides a table that lists all proteins together with their *P* values, fold changes, color listing and other important information. Double-click on a protein of interest to circle it on the plot.

53 Hover the mouse over a dot for detailed information on the protein it represents.

Generating a clustergram with the Clustergram module ● Timing ~5–10 min

▲CRITICAL The Clustergram module (Fig. 11, bottom right panel) is made available in the 'Select' menu after clicking on 'Load' to load a .plp file. At least two conditions must be included in the .plp file.

54 To use this module, simply click on the 'Plot' button.

55 Use 'Feature Selection Stringency' to specify the percentage of proteins to be considered for generating the dendrogram. If 0.90 is selected, only 10% of the proteins will be considered.

Analyzing complex proteomic experiments with TrendQuest ● Timing ~25 min

▲CRITICAL The TrendQuest module (Fig. 11, bottom left panel) is made available in the 'Select' menu; click on 'Load' to load a .plp file. Several modules are available under the Utils and Help Menus; see Box 2 for more information.

56 Set the quality filters. 'Minimum row points' establishes a cutoff for eliminating identifications that do not appear in a minimum number of biological conditions. 'Number of Clusters' is the desired number of clusters and is passed on to the *k*-means algorithm.

57 Check the 'Plot Consensus' box to plot the average protein profile of each cluster.

58 Check the 'Plot Lines' box to plot the protein profiles as lines.

59 Click on 'Execute' to generate clusters.

60 In the table on the lower left-hand corner, with one line per cluster:

- 'Node ID' is the number of the cluster
- 'Count' is the number of proteins in the cluster
- 'Consensus' is the profile obtained by averaging the quantitation profiles of the proteins in the cluster
- 'Color' is a user-selected color to represent the cluster in the plot

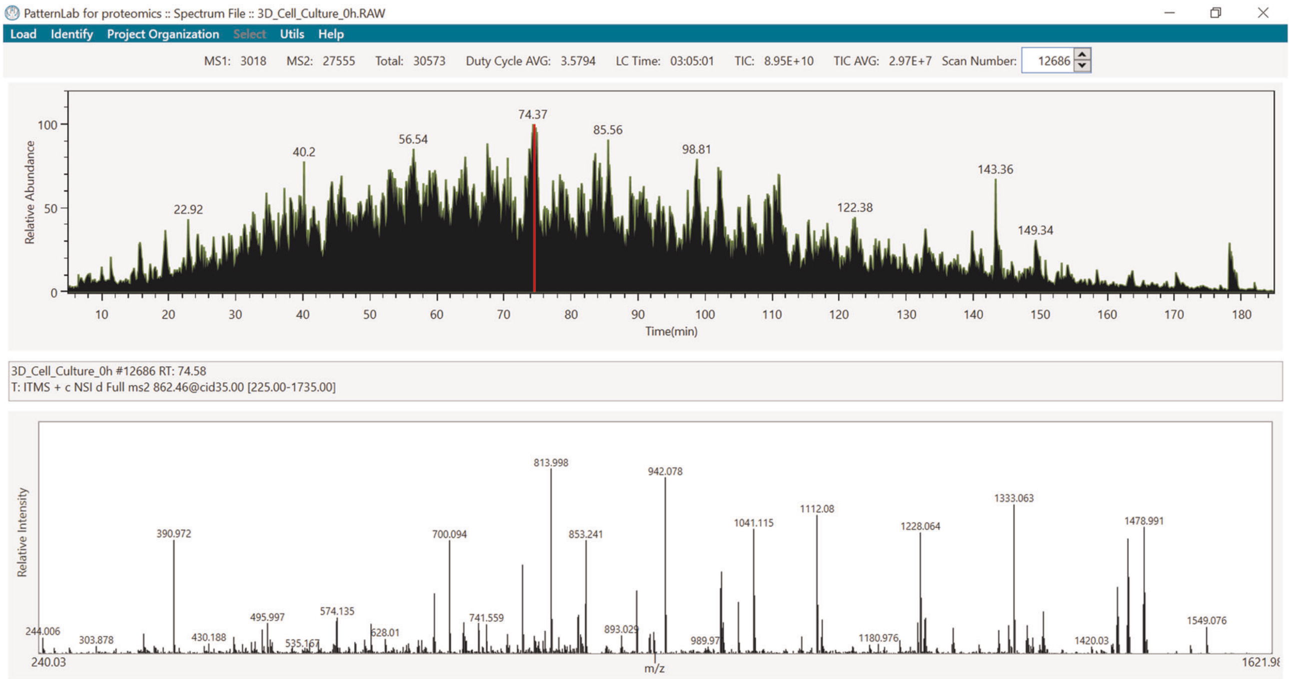
61 (Optional) In the 'Secondary Labels' tab, the user can label biological conditions to generate clusters according to labeling. An example of use for this feature is illustrated by Camillo-Andrade and collaborators[34]; the authors performed a time-course experiment on a cell culture exposed to quinoa bioester and analyzed results alongside those from an unexposed cell culture. 'Secondary Labels' were used to treat the two sets of results independently, even though they share timepoints.

62 The 'Data' tab can be used to list proteins belonging to a given cluster. Use 'Stringency' in the 'Data' tab to remove proteins from the cluster that are considered outliers.

63 Generate a report file in the 'Save' tab; this contains plots for all clusters.

Creating a custom target–decoy database ● Timing 5 min

▲CRITICAL The sequence database generator will simplify the locus names, unless the 'Preserve Header' is selected.



PatternLab for proteomics :: Spectrum File :: 3D_Cell_Culture_0h.RAW

MS1: 3018 MS2: 27555 Total: 30573 Duty Cycle AVG: 3.5794 LC Time: 03:05:01 TIC: 8.95E+10 TIC AVG: 2.97E+7 Scan Number: 12686

3D_Cell_Culture_0h #12686 RT: 74.58
T: ITMS + c NSI d Full ms2 862.46@cid35.00 [225.00-1735.00]

Fig. 12 | MS Browser. The upper and lower panels display the chromatogram and a selected spectrum, respectively. The acquisition time of the selected spectrum is indicated with a red line.

▲**CRITICAL** The ‘Nucleotide 2 Amino Acids’ tab provides options for converting a DNA sequence database into a protein sequence database.

64 Click on the ‘Utils’ menu, and then select ‘Generate Search DB’.

65 Choose whether to include the 123 common contaminants to mass spectrometry (e.g., keratins) by checking the ‘Include My Contaminant Library’ box.

66 (Optional) The default sequence contaminant library can be updated according to user preferences. To do this, click on the ‘My Contaminant Library’ tab and edit the entries available, then click on the ‘Update Library’ button. The original contaminant library can always be restored by clicking on the ‘Restore PatternLab’s Defaults’ button available on the same tab.

67 Check the ‘Include Reverse Decoys’ box unless you have already included a sequence database that includes decoys in it.

68 Remove subset sequences (i.e., sequences that are fully contained within another sequence) according to an identity percentage specified by the user. This is done using the ‘Eliminate subset sequence’ checkbox.

69 Click on the ‘Include Sequence Files’ button to select the files to be included, in FASTA format, in the Files dialog. Note that more than one file can be included by holding the Shift key while selecting the files.

70 Click on the ‘Save Database’ button to generate the new database.

Converting Thermo.raw files with MS Converter ● Timing ~1 min per raw file

71 Open the MS Converter from the ‘Utils’ menu.

72 Click on the ‘Browse’ button, and select the directory with the Thermo.raw files to be converted.

73 Select the MS levels to export (out of MS1, MS2 and MS3) and the desired output format (MGF or MS2).

74 Click on the ‘Go’ button. The new files will be placed in the same directory.

Browsing Thermo.raw files with MS Browser ● Timing ~1–2 min

75 Open the MS Browser (Fig. 12) by clicking on ‘Load’ to load the Thermo.raw file to visualize. The upper panel will display the chromatogram.

76 Click on the chromatogram or select a scan number to view the mass spectrum in the bottom panel.

The GUI will also display the raw file information listed below:

MS1	The number of MS1 spectra in the file
MS2	The number of MS2 spectra in the file
Total	The total number of spectra in the file
Duty Cycle time	Average duration, in seconds, between one MS1 and the next
LC Time	The total time of the chromatography run
Scan Number	Scan number of each spectrum

Generating mzIdentML for full PRIDE submissions ● Timing ~1–15 min

77 Click on 'Utils' from the main menu, then on 'SEPro2' and'Fusion'.
78 Click on the 'Browse' button and select the directory with SEPro and raw files to be converted.
79 Click on the 'Go' button.
80 Click on the 'Save mzIdentML' button to generate the files required to complete the full PRIDE submission.

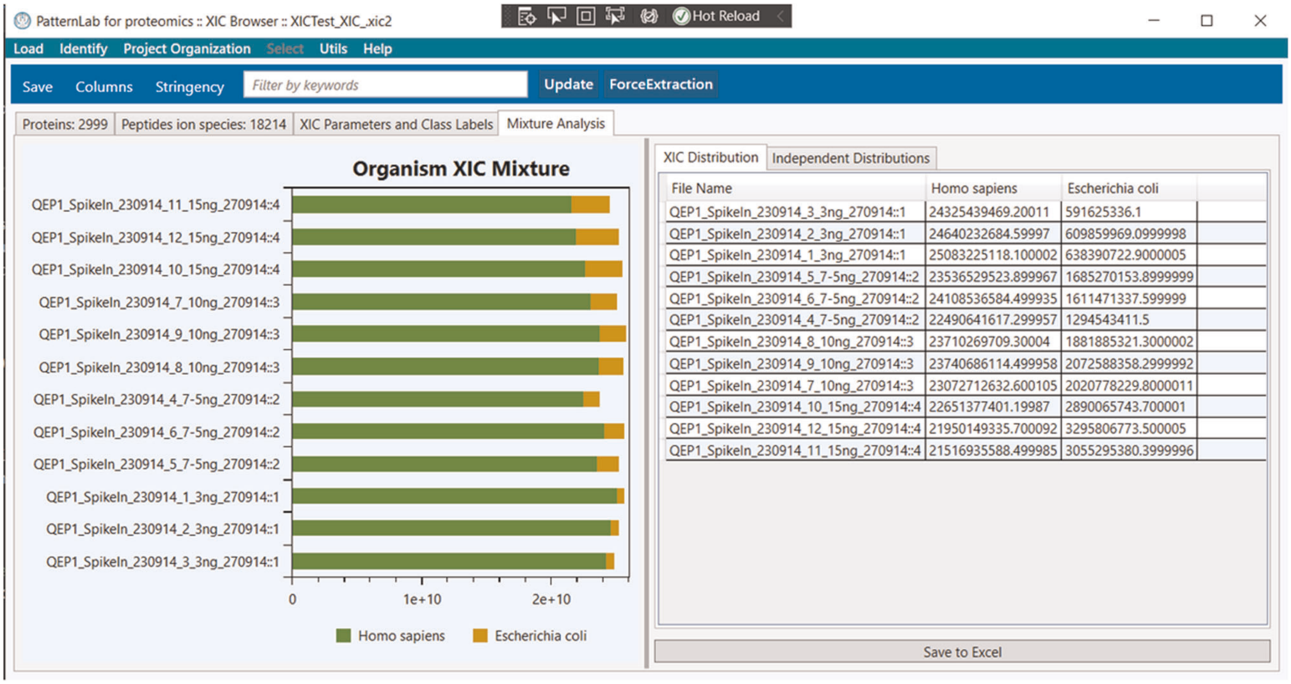
Timing

Steps 1–2, Installing the software: ~3–15 min
Steps 3–6, Downloading a sequence database: 5 min
Steps 7–15, Performing PSM with the integrated Comet search engine: can range from <1 h to even >1 d, depending on the number of samples analyzed and the equipment used
Steps 16–18, Specifying parameters for Comet and SEPro: ~2 min
Steps 19–26, Interpretation of protein identification results: this process involves scientific interpretation that may require discussion and further research
Steps 27–31, Project organization and quantitation: may vary depending on the number of raw files. It should take <20 min for quantitation to be performed in a typical proteomic experiment with ~20 raw files
Steps 32–37, Quantitation analysis with XIC: ~5–10 min as this is an exploratory module
Step 38, Shortlisting proteins of interest from the data: this process involves scientific interpretation that may require discussion and further research
Steps 39–40, Obtaining a bird's-eye view with the Buzios module: ~5–10 min
Steps 41–45, Differential proteomics with the Area-proportional Venn Diagram module: ~5–10 min
Steps 46–53, Differential proteomics with the TFold module for label-free data: ~5–10 min
Steps 54–55, Generating a clustergram with the Clustergram module: ~5–10 min
Steps 56–63, Analyzing complex proteomic experiments with TrendQuest: ~25 min
Steps 64–70, Creating a custom target–decoy database: 5 min
Steps 71–74, Converting Thermo.raw files with MS Converter: ~1 min per raw file
Steps 75–76, Browsing Thermo.raw files with MS Browser: ~1–2 min
Steps 77–80, Generating mzIdentML for full PRIDE submissions: ~1–15 min

Anticipated results

PLV is the product of joint efforts within our group's interaction with the scientific community since 2008. As such, it has been tested thousands of times over a broad range of scientific experiments with the aim of identifying and quantitating the proteins found in complex mixtures such as biological fluids, cell lysates and microorganisms. PLV has a completely rewritten GUI, built to improve user experience and thus reduce several tedious and laborious steps that the previous version required. Most algorithms have also been improved, rendering the software faster and more sensitive.

Throughout this protocol, we have shown examples to illustrate the outcomes of individual steps. In general, shotgun proteomic experiments will start with the identification step and most will proceed to quantitation, either by label-free (i.e., spectral counting or XICs) or labeled (e.g., iTRAQ or TMT) methods. In a typical proteomic experiment comprising complex protein mixtures, thousands of proteins will be identified. If this is not the case, we urge the investigator to refer to the mass spectrometry raw files using PatternLab's MSViewer to verify that the chromatograms are as expected. We also suggest referring to RawVegetable, a software for assessing experimental reproducibility and general quality control[54]. These issues settled, applicability is broad and anticipated


PatternLab for proteomics :: XIC Browser :: XICTest_XIC_.xic2
Hot Reload
Load
Identify
Project Organization
Select
Utils
Help
Save
Columns
Stringency
Filter by keywords
Update
ForceExtraction
Proteins: 2999
Peptides ion species: 18214
XIC Parameters and Class Labels
Mixture Analysis
Organism XIC Mixture
QEP1_SpikeIn_230914_11_15ng_270914::4
QEP1_SpikeIn_230914_12_15ng_270914::4
QEP1_SpikeIn_230914_10_15ng_270914::4
QEP1_SpikeIn_230914_7_10ng_270914::3
QEP1_SpikeIn_230914_9_10ng_270914::3
QEP1_SpikeIn_230914_8_10ng_270914::3
QEP1_SpikeIn_230914_4_7-5ng_270914::2
QEP1_SpikeIn_230914_6_7-5ng_270914::2
QEP1_SpikeIn_230914_5_7-5ng_270914::2
QEP1_SpikeIn_230914_1_3ng_270914::1
QEP1_SpikeIn_230914_2_3ng_270914::1
QEP1_SpikeIn_230914_3_3ng_270914::1
0
1e+10
2e+10
🟩 Homo sapiens
🟧 Escherichia coli
XIC Distribution
Independent Distributions
Save to Excel

File Name	Homo sapiens	Escherichia coli
QEP1_SpikeIn_230914_3_3ng_270914::1	24325439469.20011	591625336.1
QEP1_SpikeIn_230914_2_3ng_270914::1	24640232684.59997	609859969.0999998
QEP1_SpikeIn_230914_1_3ng_270914::1	25083225118.100002	638390722.9000005
QEP1_SpikeIn_230914_5_7-5ng_270914::2	23536529523.899967	1685270153.8999999
QEP1_SpikeIn_230914_6_7-5ng_270914::2	24108536584.499935	1611471337.599999
QEP1_SpikeIn_230914_4_7-5ng_270914::2	22490641617.299957	1294543411.5
QEP1_SpikeIn_230914_8_10ng_270914::3	23710269709.30004	1881885321.3000002
QEP1_SpikeIn_230914_9_10ng_270914::3	23740686114.499958	2072588358.2999992
QEP1_SpikeIn_230914_7_10ng_270914::3	23072712632.600105	2020778229.8000011
QEP1_SpikeIn_230914_10_15ng_270914::4	22651377401.19987	2890065743.700001
QEP1_SpikeIn_230914_12_15ng_270914::4	21950149335.700092	3295806773.500005
QEP1_SpikeIn_230914_11_15ng_270914::4	21516935588.499985	3055295380.3999996

Fig. 13 | XIC Browser's Mixture Analysis module. The 'Mixture Analysis' tab displays the sum of the ion current per organism. This feature is enabled when the sequences in the database are labeled with the OS=[Organism] tag in the description, as in UniProt. The analysis above was generated by spiking 3 ng, 7.5 ng, 10 ng and 15 ng of *E. coli* digest into 200 ng of HeLa digest to simulate protein expression differences in a background of an unchanged complex proteome (dataset ProteomeXchange identifier PXD001385)[57].

results are best illustrated by examples, as follows. First, Prieto and collaborators used PatternLab to interpret the mass spectrometric data derived from exosomes from chronic lymphocytic leukemia cells and showed that S100-A9 promoted NF-kB activity during disease progression[31]. Second, Horstmann et al. used PatternLab to generate protein sequence coverage maps when investigating the role of methylation in the flagellin of *Salmonella* in cell adhesion and invasion[33]. Third, Bonilauri et al. used PatternLab to compare proteomic and genomic data and shortlist proteins involved in the first step of adipogenesis in human adipose-derived stem cells[55].

PLV identifies differentially abundant proteins

Several of the screenshots used to illustrate this protocol were generated using datasets from recent publications; for example, the outer panels in Fig. 11 (screenshots from the Area-proportional Venn Diagram, TFold, Clustergram and TrendQuest modules, respectively, when read clockwise) are based on the dataset used by Camillo-Andrade and collaborators while investigating the effects of quinoa bioester on epidermal tissue through a time-course proteomic experiment[34]. The authors pinpointed differentially abundant proteins using the TFold module, appropriately clustered samples using the Clustergram module, and grouped proteins that presented similar expression profiles throughout several experimental timepoints using the TrendQuest module.

PLV provides confident identifications

Bioinformatics is a cornerstone for proteomics, so new algorithms or updates require rigorous testing[5,56]. Typically, peptide identifications are assessed by searching datasets of scrambled sequences or including unlabeled decoy sequences and quantitation extraction can be evaluated on datasets generated with spiked proteins of known amounts. Zhang and collaborators compared the trustworthiness of search engine identifications and reported PatternLab as the only algorithm to identify no decoy proteins when searching datasets containing only decoys[19].

PLV provides robust label-free quantitation

As for MS1-based label-free quantitation, Shalit and collaborators published a dataset for evaluating quantitation tools, generated using a Q-Exactive mass spectrometer to analyze an *E. coli* digest spiked into a HeLa digest in four concentrations to simulate protein expression differences in a background of an unchanged complex proteome (ProteomeXchange identifier PXD001385)[57]. In brief, groups of 3, 7.5, 10 and 15 ng of *E. coli* were added to 200 ng of Hela digest to simulate 5-, 2- and 1.5-fold changes relative to the 15 ng sample. Figure 8 of their manuscript finds the average quantitation ratios of Expressionist and MaxQuant to be 5.4 and 7.0 for the 5:1 test, 2.0 and 2.1 for the 2:1 test, and 1.6 and 1.8 for the 1.5:1 test, respectively. PatternLab's average ratios of ion currents for these results were 5.0, 2.0 and 1.5, respectively. The XIC used for obtaining these ratios is displayed in a screenshot of PLV's 'Mixture Analysis' tab from XICBrowser for this dataset (Fig. 13).

Where to find more example results and discussion

Importantly—although we mention this last—we have striven to support PatternLab users and are looking forward to implementing new features. In this regard, the PatternLab user forum continues to be for PLV the vital platform it has been so far, allowing the continual introduction of community-driven improvements to this freely available software. The project's website also provides a test dataset with expected results, allowing users to acquaint themselves with the system before tackling their data and committing to conclusions.

Data availability

All data associated with this protocol are provided within the paper or the supporting primary research papers, e.g., refs. [34,58].

Code availability

The software used in this protocol can be found at http://patternlabforproteomics.org

References

1. Washburn, M. P., Wolters, D. & Yates, J. R. III Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
2. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
3. Zhang, B., Chambers, M. C. & Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **6**, 3549–3557 (2007).
4. Elias, J. E. & Gygi, S. P. Target–decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
5. Yates, J. R. III et al. Toward objective evaluation of proteomic algorithms. *Nat. Methods* **9**, 455–456 (2012).
6. Barboza, R. et al. Can the false-discovery rate be misleading? *Proteomics* **11**, 4105–4108 (2011).
7. Carvalho, P. C. et al. Search engine processor: filtering and organizing peptide spectrum matches. *Proteomics* **12**, 944–949 (2012).
8. Moosa, J. M., Guan, S., Moran, M. F. & Ma, B. Repeat-preserving decoy database for false discovery rate estimation in peptide identification. *J. Proteome Res.* **19**, 1029–1036 (2020).
9. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
10. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
11. Keller, A., Eng, J., Zhang, N., Li, X. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 0017 (2005).
12. Kohlbacher, O. et al. TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **23**, e191–e197 (2007).
13. McDonald, W. H. et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom.* **18**, 2162–2168 (2004).
14. Xu, T. et al. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *J. Proteom.* **129**, 16–24 (2015).
15. Carvalho, P. C., Fischer, J. S. G., Chen, E. I., Yates, J. R. & Barbosa, V. C. PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinform.* **9**, 316 (2008).
16. Carvalho, P. C., Hewel, J., Barbosa, V. C. & Yates, J. R. III Identifying differences in protein expression levels by spectral counting and feature selection. *Genet. Mol. Res.* **7**, 342–356 (2008).
17. Liu, H., Sadygov, R. G. & Yates, J. R. III A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).
18. Carvalho, P. C., Yates Iii, J. R. & Barbosa, V. C. Analyzing shotgun proteomic data with PatternLab for proteomics. *Curr. Protoc. Bioinform*. Chapter 13, Unit 13.13.1–15 (2010).
19. Zhang, S.-R. et al. The Null-Test for peptide identification algorithm in Shotgun proteomics. *J. Proteom.* **163**, 118–125 (2017).
20. Carvalho, P. C., Fischer, J. S. G., Xu, T., Yates, J. R., III & Barbosa, V. C. PatternLab: from mass spectra to label-free differential shotgun proteomics. *Curr. Protoc. Bioinform*. Chapter 13, Unit13.19 (2012).

21. Carvalho, P. C., Yates, J. R. III & Barbosa, V. C. Improving the TFold test for differential shotgun proteomics. *Bioinformatics* **28**, 1652–1654 (2012).
22. Carvalho, P. C. et al. Analyzing marginal cases in differential shotgun proteomics. *Bioinformatics* **27**, 275–276 (2011).
23. de Saldanha da Gama Fischer, J. et al. Chemo-resistant protein expression pattern of glioblastoma cells (A172) to perillyl alcohol. *J. Proteome Res.* **10**, 153–160 (2011).
24. Leprevost, F. V. et al. PepExplorer: a similarity-driven tool for analyzing de novo sequencing results. *Mol. Cell Proteom.* https://doi.org/10.1074/mcp.M113.037002 (2014).
25. Fischer, J. et al. A scoring model for phosphopeptide site localization and its impact on the question of whether to use MSA. *J. Proteom.* https://doi.org/10.1016/j.jprot.2015.01.008 (2015).
26. Eng, J. K. et al. A deeper look into Comet–implementation and features. *J. Am. Soc. Mass Spectrom.* **26**, 1865–1874 (2015).
27. Carvalho, P. C. et al. Integrated analysis of shotgun proteomic data with PatternLab for proteomics 4.0. *Nat. Protoc.* **11**, 102–117 (2015).
28. Santos, M. D. M. et al. Mixed-data acquisition: next-generation quantitative proteomics data acquisition. *J. Proteom.* **222**, 103803 (2020).
29. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
30. Gatchalian, J. et al. A non-canonical BRD9-containing BAF chromatin remodeling complex regulates naive pluripotency in mouse embryonic stem cells. *Nat. Commun.* **9**, 5139 (2018).
31. Prieto, D. et al. S100-A9 protein in exosomes from chronic lymphocytic leukemia cells promotes NF-κB activity during disease progression. *Blood* **130**, 777–788 (2017).
32. Sogues, A. et al. Essential dynamic interdependence of FtsZ and SepF for Z-ring and septum formation in *Corynebacterium glutamicum*. *Nat. Commun.* **11**, 1641 (2020).
33. Horstmann, J. A. et al. Methylation of *Salmonella typhimurium* flagella promotes bacterial adhesion and host cell invasion. *Nat. Commun.* **11**, 2013 (2020).
34. Camillo-Andrade, A. C. et al. Proteomics reveals that quinoa bioester promotes replenishing effects in epidermal tissue. *Sci. Rep.* **10**, 19392 (2020).
35. Richards, A. L. et al. One-hour proteome analysis in yeast. *Nat. Protoc.* **10**, 701–714 (2015).
36. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41**, D43–D47 (2013).
37. Zahn-Zabal, M. et al. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.* **48**, D328–D334 (2020).
38. Li, H. et al. Evaluating the effect of database inflation in proteogenomic search on sensitive and reliable peptide identification. *BMC Genomics* **17**, 1031 (2016).
39. Ma, B. Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26**, 1885–1894 (2015).
40. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* **75**, 1895–1904 (2003).
41. Ong, S.-E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteom.* **1**, 376–386 (2002).
42. Santos, M. D. M. et al. A quantitation module for isotope-labeled peptides integrated into PatternLab for proteomics. *J. Proteom.* **202**, 103371 (2019).
43. Vizcaíno, J. A. et al. The mzIdentML data standard version 1.2, supporting advances in proteome informatics. *Mol. Cell Proteom.* **16**, 1275–1285 (2017).
44. Perez-Riverol, Y. et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
45. Martens, L. et al. mzML—a community standard for mass spectrometry data. *Mol. Cell Proteom.* **10**, R110.000133–R110.000133 (2011).
46. Eng, J. K., Searle, B. C., Clauser, K. R. & Tabb, D. L. A face in the crowd: recognizing peptides through database search. *Mol. Cell Proteom.* **10**, R111.009522 (2011).
47. Eng, J. K. & Deutsch, E. W. Extending Comet for global amino acid variant and post-translational modification analysis using the PSI extended FASTA format. *Proteomics* **20**, 1900362 (2020).
48. Wippel, H. H. et al. Comparing intestinal versus diffuse gastric cancer using a PEFF-oriented proteomic pipeline. *J. Proteom.* https://doi.org/10.1016/j.jprot.2017.10.005 (2017).
49. Punta, M. et al. The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
50. Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B. & Gough, J. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res.* **47**, D490–D494 (2019).
51. Zybailov, B. et al. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* **5**, 2339–2347 (2006).
52. Brunoro, G. V. F. et al. Reevaluating the *Trypanosoma cruzi* proteomic map: the shotgun description of bloodstream trypomastigotes. *J. Proteom.* **115**, 58–65 (2015).
53. Benjamini, Yoav & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
54. Kurt, L. U. et al. RawVegetable—a data assessment tool for proteomics and cross-linking mass spectrometry experiments. *J. Proteom.* **225**, 103864 (2020).

55. Bonilauri, B. et al. Proteogenomic analysis reveals proteins involved in the first step of adipogenesis in human adipose-derived stem cells. *Stem Cells Int.* **2021**, 1–14 (2021).
56. Leprevost, F. et al. On best practices in the development of bioinformatics software. *Front. Genet.* **5**, 199 (2014).
57. Shalit, T., Elinger, D., Savidor, A., Gabashvili, A. & Levin, Y. MS1-based label-free proteomics using a quadrupole orbitrap mass spectrometer. *J. Proteome Res.* **14**, 1979–1986 (2015).
58. Keshishian, H. et al. Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat. Protoc.* **12**, 1683–1701 (2017).

Acknowledgements

We thank W. Nagib, from Fiocruz, for creating the new PatternLab logo and entrance screen and J. Eng, from the University of Washington, for all the support and adaptations in the Comet search engine. We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Fiocruz, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support. R.H.V. (grant 304523/2019-4), V.C.B. (grant 300987/2019-6) and P.C.C. (grant 308930/2020-7) are CNPq research fellows. J.R.Y. acknowledges NIH P41 GM103533.

Author contributions

P.C.C., J.R.Y. and V.C.B. have participated since the initial version of PatternLab, published in 2008. M.D.M.S., D.B.L., M.A.C., L.U.K., L.C.M. and P.C.C. served as developers, implementing the many features that enabled the transition from PL4 to PLV. J.S.G.F., P.F.d.A., A.G.C.N.F., R.H.V., M.O.T., G.V.F.B., T.A.C.B.S., R.M.S., A.C.C.-A., M.B., F.C.G. and R.D. are all experts in proteomics and worked closely with the computational team in developing new features, improving user experience and performing in-depth testing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Valmir C. Barbosa or Paulo C. Carvalho.

Peer review information *Nature Protocols* thanks Annalisa Santucci, Yafeng Zhu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Received: 14 June 2021; Accepted: 8 February 2022;
Published online: 11 April 2022

Related links

Key references using this protocol

Gatchalian, J. et al. *Nat. Commun.* **9**, 5139 (2018): https://doi.org/10.1038/s41467-018-07528-9
Prieto, D. et al. *Blood* **130**, 777–788 (2017): https://doi.org/10.1182/blood-2017-02-769851
Sogues, A. et al. *Nat. Commun.* **11**, 1641 (2020): https://doi.org/10.1038/s41467-020-15490-8
Horstmann, J. A. et al. *Nat. Commun.* **11**, 2013 (2020): https://doi.org/10.1038/s41467-020-15738-3

Key data used in this protocol

Camillo-Andrade, A. C. et al. *Sci. Rep.* **10**, 19392 (2020): https://doi.org/10.1038/s41598-020-76325-6
Shalit, T. et al. *Proteome Res.* **14**, 1979–1986 (2015): https://doi.org/10.1021/pr501045t