



# Quantitative Proteomics in Yeast: From bSLIM and Proteome Discoverer Outputs to Graphical Assessment of the Significance of Protein Quantification Scores

Nicolas Sénécaut, Pierre Poulain, Laurent Lignières, Samuel Terrier,  
Véronique Legros, Guillaume Chevreux, Gaëlle Lelandais, Jean-Michel  
Camadro

## ► To cite this version:

Nicolas Sénécaut, Pierre Poulain, Laurent Lignières, Samuel Terrier, Véronique Legros, et al.. Quantitative Proteomics in Yeast: From bSLIM and Proteome Discoverer Outputs to Graphical Assessment of the Significance of Protein Quantification Scores. Frédéric Devaux. Yeast Functional Genomics. Methods and Protocols, 2477, Springer US, pp.275-292, 2022, Methods in Molecular Biology, 978-1-0716-2256-8. 10.1007/978-1-0716-2257-5\_16 . hal-03814950

**HAL Id: hal-03814950**

**<https://hal.science/hal-03814950>**

Submitted on 14 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Quantitative Proteomics in Yeast: From bSLIM and Proteome Discoverer Outputs to Graphical Assessment of the Significance of Protein Quantification Scores

Nicolas Sénécaut, Pierre Poulain, Laurent Lignières, Samuel Terrier, Véronique Legros, Guillaume Chevreux, Gaëlle Lelandais, and Jean-Michel Camadro

### Abstract

Simple light isotope metabolic labeling (bSLIM) is an innovative method to accurately quantify differences in protein abundance at the proteome level in standard bottom-up experiments. The quantification process requires computation of the ratio of intensity of several isotopologs in the isotopic cluster of every identified peptide. Thus, appropriate bioinformatic workflows are required to extract the signals from the instrument files and calculate the required ratio to infer peptide/protein abundance. In a previous study (Sénécaut et al., J Proteome Res 20:1476–1487, 2021), we developed original open-source workflows based on OpenMS nodes implemented in a KNIME working environment. Here, we extend the use of the bSLIM labeling strategy in quantitative proteomics by presenting an alternative procedure to extract isotopolog intensities and process them by taking advantage of new functionalities integrated into the Minora node of Proteome Discoverer 2.4 software. We also present a graphical strategy to evaluate the statistical robustness of protein quantification scores and calculate the associated false discovery rates (FDR). We validated these approaches in a case study in which we compared the differences between the proteomes of two closely related yeast strains.

**Key words** Mass spectrometry, Proteomics, Quantification, Isotopic labeling, Yeast, Metabolism, bSLIM, Proteome Discoverer, Minora, KNIME

---

## 1 Introduction

The yeast *Saccharomyces cerevisiae* is a biological model that is widely used for the development and validation of global analytical methods in functional genomics and genetics. Yeast has been extensively studied for many years, resulting in a solid understanding of its physiology and metabolism. Yeast is the first eukaryotic

---

Gaëlle Lelandais and Jean-Michel Camadro are co-senior authors.

organism for which the genome was fully sequenced [1]. This has opened up new avenues for the exploration of living organisms, notably through the analysis of gene and protein expression using the amazing recent technical developments in transcriptomics and proteomics. In-depth knowledge of the yeast genome has enabled the construction of complete collections of haploid or diploid strains carrying modified alleles, for example, disruptions, deletions, and ORF or promoter fusions with a large number of reporter genes for use as probes to assess gene function and associated regulatory networks, laying the foundation for systems biology.

One specific aspect of yeast is its ability to grow in the presence (aerobiosis) or absence (anaerobiosis) of molecular oxygen. This is made possible by a metabolic switch that allows passage from respiratory to fermentative metabolism, provided that the carbon source available to the yeast can be metabolized by fermentation. This requires processes in which mitochondrial functions are essential, making yeast a critical organism for deciphering the genetics, biochemistry, and physiology of this energy producing organelle.

Yeast cells are capable of growing on synthetic media in which the vitamins and essential trace elements are provided or on complex media containing yeast hydrolysates or peptones. In wild-type yeast grown on synthetic media, cell metabolism is based on the assimilation of organic nitrogen (usually ammonium sulphate or chloride) and the catabolism of a single carbon source, such as glucose, glycerol, or acetate. The genetics of yeast has been brought to light through the selection and use of mutants affected in the synthesis of certain nucleotides (e.g., Ura3) or amino acids (e.g., Lys, His, Leu, Met, Arg, Trp). Such auxotrophic mutants require the addition of the defective bases and/or amino acids in the synthetic growth media.

Beyond the identification of proteins in complex extracts, mass spectrometry-based proteomic analysis allows the quantification of differences in the proteome between several biological states. Several bottom-up quantitative proteomics approaches have been reported [2], providing critical information in yeast biology. They are based either on *in vitro* labeling of peptides by isobaric chemical probes, releasing fragments in MS/MS, of which the measured intensity reflects the abundance of the protein in the initial extract (e.g., ICAT or TMT labeling), or on differential metabolic labeling *in vivo*, in which the cells are cultured in the presence of “light” (unlabeled) or “heavy” (labeled) amino acids that will be incorporated into the proteins, allowing their quantification after tryptic digestion and the measurement of a heavy–light ratio for each peptide/protein (e.g., SILAC and derived methods) (for a review, *see* [3]). Despite the fact that TMT- and SILAC-based quantitative proteomics allow multiplexing multiple samples in a single run, one of the most widely used proteomics approaches is

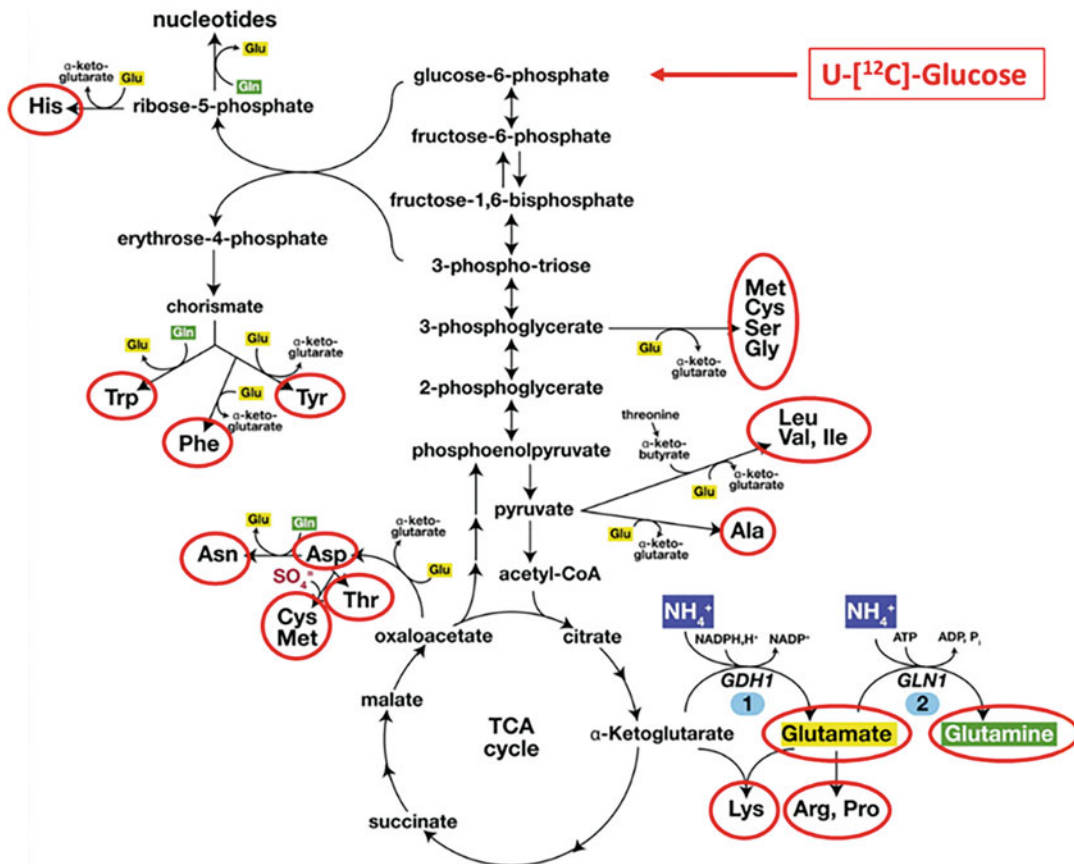
**Table 1**  
**Relative abundance of the stable isotopes of the elements found in proteins**

$^{12}\text{C} = 98.93\%$	$^{13}\text{C} = 1.07\%$		
$^1\text{H} = 99.9885\%$	$^2\text{D} = 0.0115\%$		
$^{16}\text{O} = 99.757\%$	$^{17}\text{O} = 0.0373\%$	$^{18}\text{O} = 0.2057\%$	
$^{14}\text{N} = 99.632\%$	$^{15}\text{N} = 0.368\%$		
$^{32}\text{S} = 94.93\%$	$^{33}\text{S} = 0.76\%$	$^{34}\text{S} = 4.29\%$	$^{36}\text{S} = 0.02\%$
$^{31}\text{P} = 100\%$			

still the label-free approach, in which individual LC-MS/MS runs are compared and the intensity of the peptide ions in MS1 are measured [4] to determine differences in protein abundance.

Recently we presented an innovative quantification method, called simple light-isotope metabolic (SLIM) labeling [5]. The SLIM-labeling strategy uses the fundamental property of the living matter in which all the biomolecules are basically composed of carbon, nitrogen, hydrogen, oxygen, and sulfur, with several additional elements, such as phosphorus, selenium, or iodine. Most of these elements, except phosphorus, are naturally present in the form of several stable isotopes for which the abundance is fixed (Table 1). It is thus possible to infer their isotopic abundance in biomolecules, such as amino acids by solely taking into account their average elemental composition:  $\text{C}_{4.9384} \text{H}_{7.7583} \text{N}_{1.3577} \text{O}_{1.4773} \text{S}_{0.0417}$  [6].

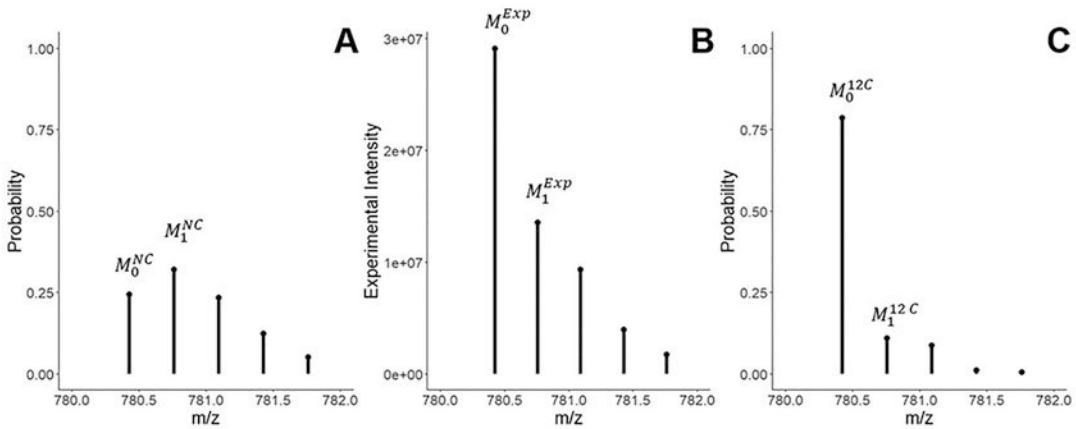
This has important consequences in terms of high-resolution MS-based peptide/protein analysis. Every peptide is measured as a series of ions ( $m/z$ ) in an isotope cluster of similar charge ( $z$ ) but with the mass ranging from the monoisotopic mass  $m_0$ , containing only the lightest isotopes of each element, to higher masses resulting from the statistical distribution of additional neutrons present in the stable isotopes (isotopologs). The intensities of the various isotopologs within an isotope cluster therefore depend on the elemental composition of the peptide and follows a Poisson distribution that can be accurately modeled using dedicated software, such as MIDAS [7]. The basic principle of SLIM labeling is to manipulate the elemental composition of proteins in vivo from the natural abundance of the isotopes of the atoms present in proteins (C, H, N, O, S, P), defining the “NC” (natural carbon) condition as the condition named “ $^{12}\text{C}$ ” in which the proteins are enriched in the light isotope of carbon ( $^{12}\text{C}$ ) and, eventually, nitrogen ( $^{14}\text{N}$ ) (referred as to “ $^{12}\text{C}^{14}\text{N}$ ” condition). Considering the main routes for amino-acid biosynthesis in yeast (Fig. 1) [8], we hypothesized that providing yeast cells with U- $^{12}\text{C}$ -glucose as the sole carbon source would result in the rapid synthesis of U- $^{12}\text{C}$ -



**Fig. 1** Yeast Metabolism (adapted from Liungdahl PO et al. YeastBook 2012 [5])

amino acids and their incorporation into newly synthesized proteins. Applying this labeling method allowed us to experimentally evaluate the half-life of the proteome in *Candida albicans*, and measure the effect of the proteasome inhibitor MG132 and a broad-specificity serine-protease inhibitor, PMSF, on the dynamics of the proteome in this organism [5].

Increasing the amino-acid content in  $^{12}\text{C}$  (and to a lesser extent in  $^{14}\text{N}$ ) results in a different and simpler isotopic cluster that always remains within the boundaries of that observed with a natural isotopic composition, but with the intensity of the monoisotopic ion greatly enhanced. This has significant impact on downstream analyses, that is, allowing better signal-to-noise discrimination, more precise mass determination, and better MS/MS fragmentation spectra. As a result, higher scores for peptide identification and protein sequence coverage are obtained (*see* characteristic mass spectra in Fig. 2a–c). We took advantage of these characteristics to develop a new quantitative proteomics method in which peptides originating from the NC condition are mixed in equimolar amounts with peptides from the  $^{12}\text{C}$  condition. The intensity of



**Fig. 2** Theoretical spectra of the peptide DQPILFWGGATAVGQMLIQLAK from the protein YNL134c under NC (a) and 12C (c) conditions and the corresponding experimental spectrum (b) of the same peptide from a 1:1 mixture of total extract from the 12C and NC conditions

every isotopolog in any isotope cluster is thus the sum of the intensity of the isotopologs from each condition (Fig. 2b). Therefore, measuring the ratio between the experimental values of the monoisotopic ion ( $M_0$ ) and the next ion containing one more neutron ( $M_1$ ), *modulo* the values of their theoretical intensity, expressed as the probability of occurrence, in each condition allows calculation of the molar fraction of the peptide originating from the NC- and 12C conditions. We recently described the full formalism for the quantification of  $^{12}\text{C}$  incorporation into proteins/peptides and its use in quantitative proteomics, and we developed the data processing tools required to smoothly run SLIM labeling experiments [9].

One critical step in the SLIM-labeling quantification procedure is the accurate extraction of the intensities for all isotopologs in each isotope cluster from the experimental spectra. In our initial study, we used commercial software, Progenesis QI for metabolomics, but it does not provide the possibility to automatically link the quantification files with the identification files [5]. This prompted us to develop another workflow, referred to as bSLIM [9], in which only the intensities of the identified peptides are used to extract the data using an OpenMS node, FeatureFinderIdentification [9], which was modified to fit every mass trace in the isotope cluster. This approach only required us to install and run the KNIME (Konstanz Information Miner) environment for computation [9], together with the latest versions of OpenMS [10–12], and hence is fully independent of any commercial software.

Here, we present an alternative integrated procedure that takes advantage of the tools available in the proprietary software suite Thermo Scientific Proteome Discoverer. Proteome Discoverer (PD v2.4) is a popular program for the analysis of peptide-centric

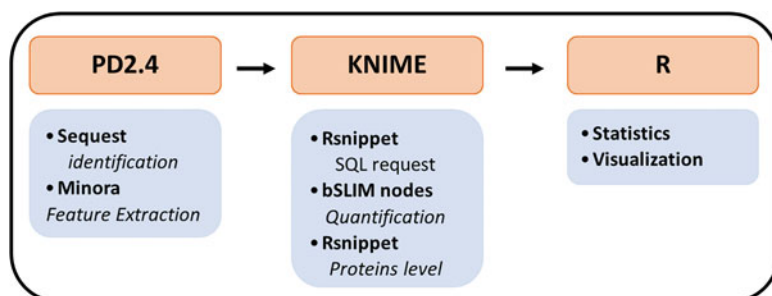
proteomics data, with a high level of integration with Thermo Fisher Scientific high-resolution, high-sensitivity Orbitrap instruments. This analytical platform includes many algorithms developed by Thermo or third parties, such as the IMP Protein Chemistry facility (<https://pd-nodes.org/>) [13, 14] and others. Proteome Discoverer is therefore widely used on a routine basis in many MS-based proteomics laboratories. It associates and integrates both raw spectra processing and filtering, peptide identification through database searches in data-dependent analyses, diverse quantification routines, and convenient spectra viewers. The output of the Proteome Discoverer analyses is written in “.msf” or “.pdresult” files. A key feature of “.pdresult” files is that they are SQLite relational databases (<https://www.sqlite.org>) that can be queried using SQL. The possibility to visualize individual annotated spectra for peptides up to the level of their isotope clusters with their associated intensity prompted us to develop the appropriate tools to extract this data and use it as input for our bSLIM labeling quantitative proteomics strategy. We accessed the individual mass trace intensities by taking advantage of the capabilities of the newly developed node, Minora, initially designed for label-free quantification.

We also present a solution to assess the robustness of the protein quantification scores calculated using bSLIM which was missing from our previous data analyses workflow. Derived from the SAM (Significance Analysis of Microarrays [15]) method, the general idea is to randomize the original bSLIM output data sets multiple times and calculate the associated “random scores.” These scores are graphically compared to the “real scores” obtained from the original bSLIM data. Proteins for which the real scores vary the most from the random scores are thus easily detected and worth considering for further analyses. In this chapter, we present a case study to illustrate the different outputs from the various workflows that we developed. We compared differences between the proteomes of two “wild-type” *Saccharomyces cerevisiae* strains with the same genetic background, but with one strain (BY4742) harboring the deletion of four genes (Ura3, His3, Leu2, and Lys2) relative to the reference strain S288c (see **Note 1** for data availability). The proteomes of the two strains are expected to be very similar and therefore represent a challenging test to assess the sensitivity and specificity of our quantification methods.

Overall, we expect that these alternative solutions implemented in the bSLIM data analysis workflow will be useful for proteomics laboratories running Orbitrap-based mass spectrometers, which are very familiar with Proteome Discoverer. This is an original way to combine the completeness and reproducibility of routine proprietary software with the power of open-source tools.

## 2 Materials

1. Reagents for yeast synthetic growth media and appropriate supplements.
2. SLIM-labeling specific reagent: U- $^{12}\text{C}$ ]-glucose (e.g., Cambridge Isotope Laboratories).
3. Lysis buffer: 40 mM HEPES–KOH, pH 7.5, 350 mM NaCl, 10% glycerol, 0.1% Tween-20.
4. Acid-washed silica beads (0.45–0.5 mm Ø).
5. 200 µg/mL trypsin solution, prepared by dissolving 20 µg trypsin (Proteomic grade) in 100 µL of 1 mM HCl.
6. Cold Acetone.
7. 50 mM ammonium carbonate ( $\text{NH}_4\text{HCO}_3$ ).
8. 0.1% formic acid (MS grade).
9. Dry incubator at 37 °C.
10. Vacuum dryer (Speed Vac).
11. Low-binding microcentrifuge tubes.
12. 4–12% polyacrylamide gradient gels.
13. Coomassie blue (MS friendly, such as SimplyBlue SafeStain, Invitrogen).
14. Bradford protein assay reagent.
15. An instrument setup for LC-MS/MSMS data acquisition (*see Note 2*).
16. Appropriate software suites for quantification and identification of the peptide/protein content of the samples analyzed (Fig. 3).



**Fig. 3** General organization of the data processing workflows



### 3 Methods

#### 3.1 Cell Growth and Preparation of Protein Extracts

1. Grow the cells to be compared in a synthetic medium with either regular glucose (NC-condition), or U-[<sup>12</sup>C]-glucose (12C-condition) as the sole carbon source (*see Note 3*).
2. At the appropriate cell density (mid-exponential phase of growth), collect the cells by centrifugation for 10 min at  $4000 \times g$  at 4 °C.
3. Wash the cell pellet with cold water, resuspend the cells in lysis buffer at a cell density of 0.6 g/mL, and lyse the cells by adding 0.32 mL acid-washed, heat-sterilized silica glass beads (0.45–0.5 mmØ) to 0.6 mL cell suspension and vortexing the resulting suspension three times for 5 min, leaving the tubes on ice for 5 min between each vortexing.
4. Centrifuge the lysed cells for 5 min at  $3000 \times g$  and collect the supernatant, referred to as the cell homogenate.
5. Carefully measure the protein concentration using the Bradford Protein microassay and validate the protein measurement by running an aliquot on an SDS-PAGE gel and staining with Coomassie Blue.
6. Precipitate a 50-µg protein aliquot using 6 vol. cold acetone for 2.5 h at –20 °C.
7. Resuspend the dry protein pellet in 50 mM ammonium carbonate buffer by heating for 15 min at 95 °C.
8. Add 5 µL of 200 µg/mL trypsin stock solution and incubate for 12 h at 37 °C in a dry incubator.
9. Remove all solvents by vacuum drying.
10. Resuspend the peptides in 0.1% formic acid.
11. Carefully mix an equal amount of peptides from the NC- and 12C-conditions.
12. Inject the samples, typically 5 µg in  $\leq 5$  µL, into the LC-MSMS instrumental setup (*see Note 4*).
13. Ensure that your instrumental setup allows the isotopic resolution of all the peptides analyzed (*see Note 5*).
14. Save the “.raw” files for data processing and signal extraction.
15. Create a folder to gather all the “.raw” files from one project together.

#### 3.2 Data Processing Workflows

1. Install the appropriate computational resources.
  - (a) Proteome Discoverer 2.4 or higher with a valid activation key.

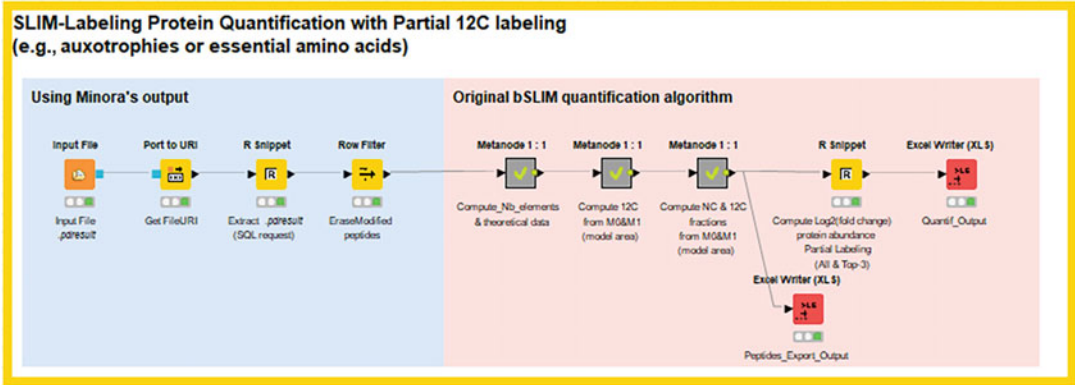
- (b) KNIME (v4.2.3) with all available extensions (<https://www.knime.com/downloads>): the OpenMS nodes (v2.6.0) are part of the “community nodes.”
- (c) R (v4.0.2), including the dplyr, dbplyr, RSQLite, sqldf, readr, raster, RMySQL packages and libraries (<https://cran.r-project.org/bin/windows/>).

### 3.2.1 Proteome Discoverer Analysis

1. Open Proteome Discoverer 2.4.
2. Create a new study and add your Thermo Fisher Scientific mass spectrometry *.raw* files.
3. Select the appropriate “processing.” The basic processing workflow is composed of the following nodes: spectrum files, spectrum selector, sequestHT (1.1.0.189), Percolator (3.02.1), and IMP-ptmRS. To this Processing workflow, add the “Minora Feature Detector” node linked to the “Spectrum Files” node and set the correct advanced parameters) (*see Note 6*).
4. Select the appropriate “consensus” workflow composed of the following nodes: MSF Files, PSM Grouper, Peptide Validator, Peptide and Protein Filter with a link to Protein annotation, Protein Scorer with a link to Protein FDR Validator, and Protein Grouping (*see Note 7*).
5. Enable the postprocessing node “Display Settings.”
6. Run the analysis, **one file at the time**, by giving nonambiguous names to the output files. The produced results files from PD2.4 have the extension *.pdresult* and are used as input in our KNIME workflow.

### 3.2.2 Isotopolog Intensity Extraction and Peptide/ Protein Quantification Using a Dedicated bSLIM KNIME Workflow (Fig. 4)

1. Open KNIME 4.2.3.
2. Import from <https://zenodo.org> (DOI 10.5281/zenodo.4467829), the bSLIM quantification workflow “*File > Import KNIME workflow*” (file extension is “.knwf”). The workflow is a modification of the original workflow presented in our previous study [9]. The adaption is very simple (disconnection between metanode 1.2 “FFiDData filtering” and connection with the Row filter “erase ModifiedPeptides”). This workflow contains three main parts.
  - (a) The “.pdresult” file, which extracts all data concerning every peptide, including their identification and the intensity of the isotopologues in the isotope clusters. This procedure uses an R script written specifically for this study. It is embedded in a dedicated RSnippet as a part of the KNIME quantification workflow (*see Note 8* on SQL request formalism).



**Fig. 4** KNIME workflow of the integrated “.pdresult” processing node connected to the quantification nodes

- (b) The peptide/protein quantification workflow based on our previous study [9]. As previously described, two biological conditions are considered: complete labeling of the proteins (true wild-type strain) or partial labeling of the proteins (strains that are auxotrophic for specific amino acids).
  - (c) The procedure to compute the statistics on the identified peptides/proteins. These two procedures use R scripts embedded in specific RSnippets.
3. Check that your installation of Rserver allows all RSnippets to run smoothly.
4. There are two possible cases:
- (a) The samples come from an autotrophic yeast strain and the SLIM labeling is complete: follow “*cases of total labeling experiment.*”
  - (b) The samples come from auxotrophic cells for which certain amino acids are not labeled. The SLIM labeling is thus incomplete: follow “*case of incomplete labeling experiment.*” In the latter case, the essential amino acids are defined as containing “Carbon B,” corresponding to carbon atoms of natural isotopic abundance. It is therefore necessary to open the meta-node “Compute\_Nb\_elements& theoretical data” and then the meta-node “Compute Elemental Composition” and, finally, the carbon B calculus node (Compute Nb\_Carbon-B (Ex: HKL)) to add the correct total number of carbons to each exogenous amino acid by typing the regular expression in the form “( $\$Nb\_H\$*6$ )+ ( $\$Nb\_K\$*6$ )+ ( $\$Nb\_L\$*6$ )”, as exemplified for the BY4742 strain auxotrophic for histidine (H), lysine (K), and leucine (L). The number of carbon atoms for each amino acid is shown in Table 2.

**Table 2**  
**Number of carbon atoms per amino acid**

Amino acid	A	C	D	E	F	G	H	I	K	L
Nb carbon atoms	3	3	4	5	9	2	6	6	6	6
Amino acid	M	N	P	Q	R	S	T	V	W	Y
Nb carbon atoms	5	4	5	5	6	3	4	5	11	9

5. Set the Excel exporter nodes with an appropriate name for file output (scores for quantifications at proteins or peptide levels) (*see Note 9*).

### 3.2.3 Statistics and Graphical Assessment of Score Significance

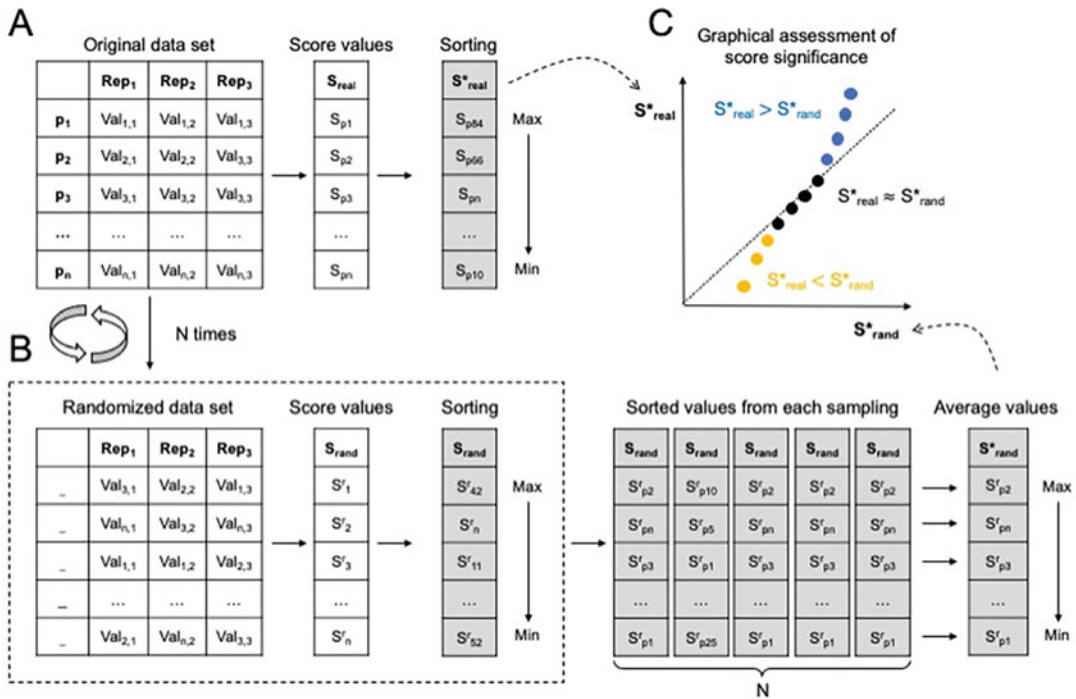
For statistical analysis of differential expression, we reproduced the SAM methodology, adapting it to the specifics of the quantitative measurements of protein abundance that are obtained at the end of the bSLIM workflow (Fig. 5) (*see Note 10*). Workflows available at <https://zenodo.org> (DOI 10.5281/zenodo.4467882).

1. Aggregate the protein quantification results from individual experiments (replicates) into a single “.tsv” table: Accession/ “name of column-2”/ “name of column-3”/ ... / “name of column-n.” Typically, column-2 to -n represents the  $\log_2$ (Fold change) of protein abundance per experimental condition.
2. Load the R scripts developed in this study to compute the scoring functions, and save the analysis.
3. Use the graphical package ggplot2, within the script, to produce the figures showing differentially expressed proteins.
4. Retrieve the table produced by the script containing the proteins for which the over- or underexpression is statistically significant between the different experimental conditions.

### 3.3 Case Study: BY4742 Vs S288c Proteome Comparison

To test and illustrate the different outputs from the various workflows, we compared the differences between the proteomes of two “wild-type” *Saccharomyces cerevisiae* strains with the same genetic background but with one strain (BY4742) harboring the deletion of four genes (Ura3, His3, Leu2, and Lys2) versus the reference strain S288c. The proteomes of the two strains are expected to be very similar (*see Note 3* for experimental details).

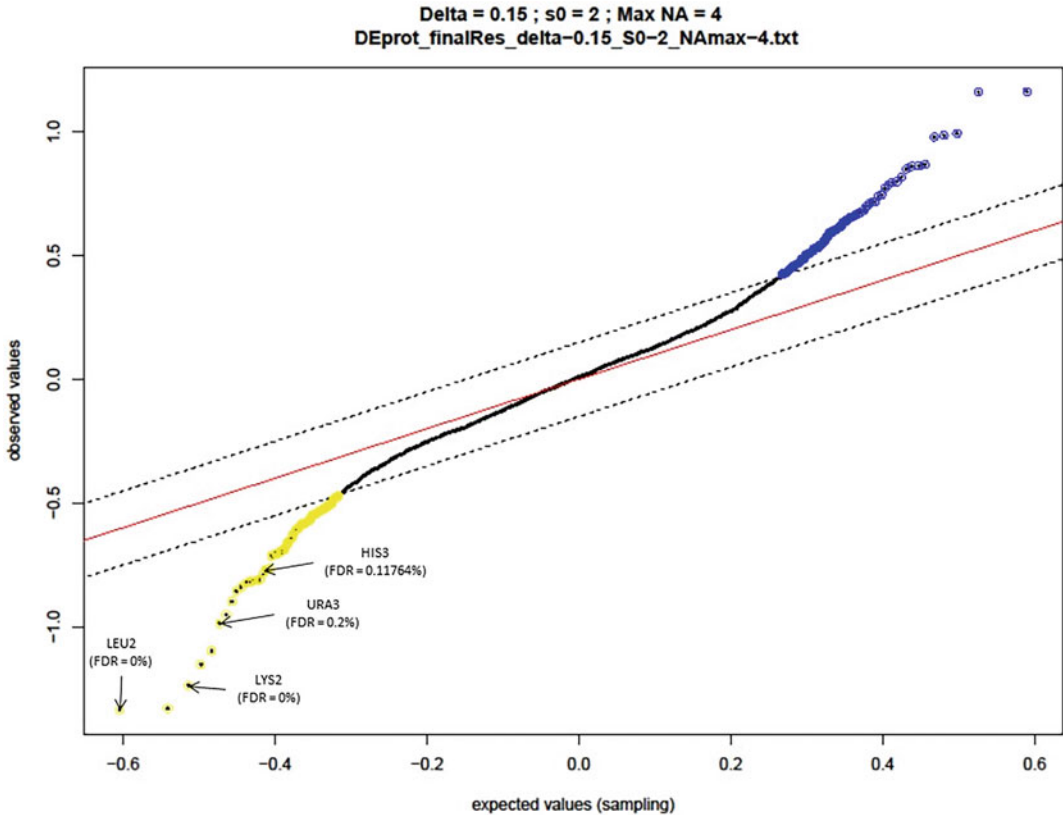
As shown in Fig. 6, the graphical representation of the distribution of quantification quality scores shows the efficiency of the workflows to identify proteins that are underexpressed (yellow) or overexpressed (blue) in the laboratory wild-type strain BY4742 relative to the reference strain S288c. All the proteins encoded by the genes that are deleted in BY4742 appear as the most significantly diminished (indeed absent) in BY4742, showing the sensitivity and specificity of the proposed quantification methods.



**Fig. 5** Schematic representation of the methodology used to assess the relevance of the bSLIM results exported for each protein. **(a)** The original bSLIM output data set is organized as a table in which the proteins are presented in rows and the repetitions of the experiments in columns. For all proteins, score values are calculated and are next sorted from the highest to the lowest. **(b)** Random data set is created from the original dataset, randomly sampling the values in each column. New scores are next calculated and sorted from the highest to the lowest. This process is repeated N times, resulting in a final table with N columns, comprised of the sorted values from each sampling. The average values of all sorted scores are finally calculated. Note that the maximal average value is derived from calculation of the mean between the maximal score values obtained in each sampling. **(c)** The significance of the scores obtained from the original data set are graphically assessed by plotting the S<sup>\*</sup><sub>real</sub> values **(a)** against the S<sup>\*</sup><sub>rand</sub> values **(b)**. Significant scores are those that are higher (colored in blue) or lower (colored in yellow) than the random scores. False discovery rates are finally calculated, comparing for each protein score, the average number of other scores in the table of sorted values (from each sampling, see **(b)**) that are higher (respectively lower) than the number of scores that are higher (respectively lower) in the original dataset. This is the same method as detailed in [15]

## 4 Notes

1. The original data sets are publicly available in the ProteomeX-change platform under the Pride submission number PXD021329.
2. The instrumental setup in our laboratory consists of Orbitrap Fusion Tribrid ETD and Orbitrap Q-Exactive Plus mass spectrometers, equipped with Easy-Spray nanoelectrospray ion sources. The LC setup consists of Easy nano-LC Proxeon 1000 or 1200 systems equipped with an Acclaim PepMap100



**Fig. 6** Experimental distribution of the statistical distribution of the protein quantification quality scores in the characterization of the differences between the BY4742 and S288c proteomes

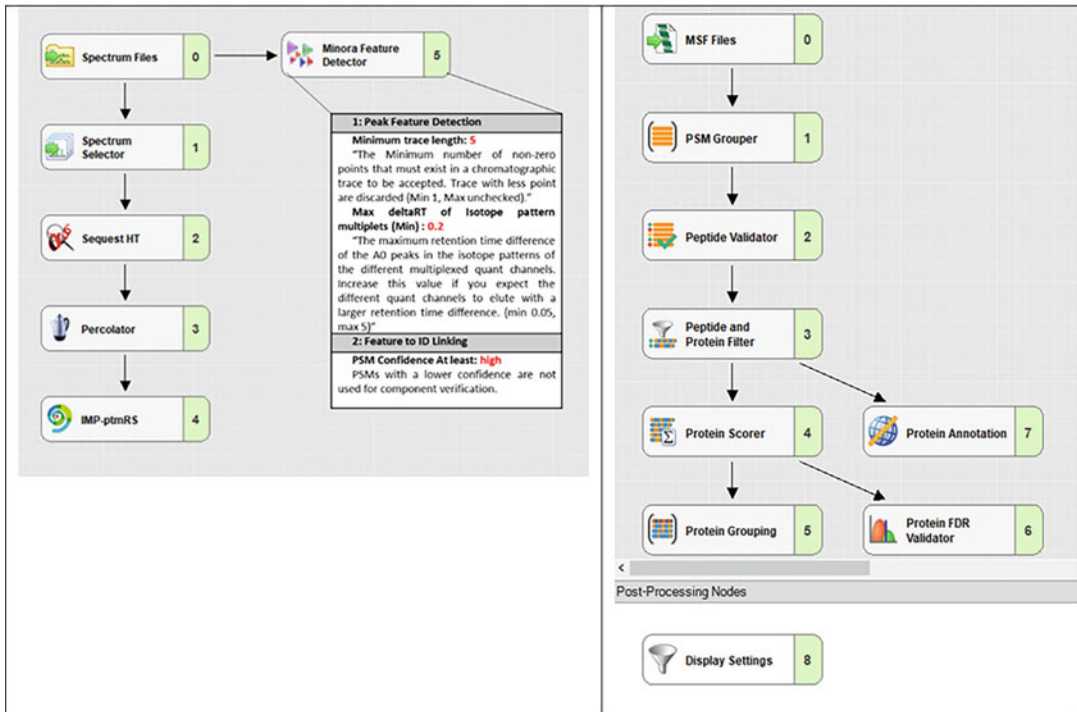
C18 precolumn and a Pepmap-RSLC Proxeon C18 column. These devices are all from Thermo Fisher Scientific (Bremen, Germany and San Jose, CA, USA).

3. In the case study presented here, the *S. cerevisiae* strains S288c (MAT $\alpha$  SUC2 gal2 mal2 mel flo1 flo8-1 hap1 ho bio1 bio6) [16] and its isogenic derivative BY4742 (MAT $\alpha$  his3 $\Delta$ 1 leu2 $\Delta$ 0 lys2 $\Delta$ 0 ura3 $\Delta$ 0) are grown on a synthetic medium made of 6.7 g/L Yeast Nitrogen Base (YNB) with ammonium sulfate, without amino acids, with 0.5% glucose as the sole carbon source. The auxotrophy of the BY4742 strain is complemented with uracil (20 mg/L), histidine (20 mg/L), and leucine and lysine (both 30 mg/L). The carbon source was either regular D(+)-glucose anhydrous, defining the normal condition (NC condition), or U-[ $^{12}\text{C}$ ]-glucose (Cambridge Isotope Laboratories), defining the  $^{12}\text{C}$  condition. The 10% glucose stock solutions were filter-sterilized.
4. Liquid chromatography coupled to mass spectrometry data acquisition:

In the case study presented here, the chromatographic separation of peptides was performed using the following parameters: Acclaim PepMap100 C18 precolumn (2 cm, 75  $\mu\text{m}$  i.d., 3  $\mu\text{m}$ , 100 Å), Pepmap-RSLC Proxeon C18 column (75 cm, 75  $\mu\text{m}$  i.d., 2  $\mu\text{m}$ , 100 Å), and 300 nL/min flow. The chromatographic separation of peptides was obtained with a gradient consisting of 95% solvent A (water, 0.1% formic acid) to 35% solvent B (99.9% acetonitrile, 0.1% formic acid) in 90 min, followed by column regeneration for 15 min, giving a total run time of 1 h and 45 min.

5. Peptides masses were analyzed in the Orbitrap cell in full ion scan mode at a resolution of 70,000 with a mass range of  $m/z$  375–1500 and an AGC target of  $3.10^6$ . MS/MS were performed in a Top 20 DDA mode. Peptides were selected for fragmentation by Higher-energy C-trap Dissociation (HCD) with a Normalized Collisional Energy of 27%, and a dynamic exclusion of 30 s. Fragment masses were measured in the Orbitrap cell at a resolution of 17,500, with an AGC target of  $2.10^5$ . Monocharged peptides and unassigned charge states were excluded from the MS/MS acquisition. The maximum ion accumulation times were set to 50 ms for MS and 45 ms for MS/MS acquisitions respectively.
6. All MS/MS data are processed using the SequestHT (v1.1.0.189) node. The mass tolerance is set to 6 ppm for precursor ions and 0.02 Da for fragments when using an Orbitrap Q-Exactive Plus mass spectrometer. The following alterations are used for various modifications: carbamidomethylation (C), if the sample is reduced and alkylated, and oxidation (M). Phosphorylation (STY) and acetylation (K, N-term) are generally added for additional analyses of trypsin digests. The maximum number of missed cleavages by trypsin is limited to two. MS/MS data are searched against the Uniprot *Saccharomyces cerevisiae* reference proteome UP000002311 (<https://www.uniprot.org/proteomes/UP000002311>, 6049 protein counts).
7. The Consensus workflow is very basic, because using the Minora node, as presented here, strictly requires that only one results file is processed per run (Fig. 7: Proteome Discoverer 2.4 consensus workflow for presentation).
  - (a) The R snippet uses SQL query to link the table of identification with the isotopic intensities. The data are then incorporated in the KNIME workflow.
  - (b) The computation is rapid and can be performed as a side analysis during the bSLIM experiment. In cases of auxotrophy, only the amino acids synthesized by the yeast are labeled, whereas the exogenous amino acids that need to





**Fig. 7** Details of Proteome Discoverer 2.4 processing (left panel) and consensus (right panel), including the specific Minora node parameters to extract the isotopolog intensities for every isotope cluster

be added to the media are not, resulting in mixed labeling. Quantification is possible with the introduction of a new calculation to accommodate this type of analysis. Experimental data were analyzed using BY4742 auxotrophic yeast.

8. For each identified peptide, we extracted the following items, which are combined in a single output table: FeatureId / MassOverCharge / ParentProteinAccessions / ParentProtein-Descriptions / MasterScanNumbers / RetentionTime / Charge / Sequence / Modifications / MonoisotopicMassOverCharge / Area / Intensity / NumberOfIsotopes / MassOverChargeIsotope / PeakHeight ( $M_0$ ) / PeakArea ( $M_0$ ) / PeakHeight ( $M_1$ ) / PeakArea ( $M_1$ ) / ... / PeakHeight ( $M_4$ ) / PeakArea ( $M_4$ ).
  - (a) The ".pdresult" file is an SQLite relational database and the information can be accessed using SQL queries. The database contains all PD search parameters, variables, and results. In our quantification workflow, we only need to access three tables that contain the relevant information:
    - *TargetPsm*s, containing the peptide IDs.
    - *LcmsFeatures*, with the description of the MS1 cluster used for the identification.



- *LcmsPeaks*, which is the most important in this study, because it contains the abundance of each individual isotopolog contained in each independent identified isotopic cluster.
- (b) To extract and produce the final output table from the database (“.pdresult”), we created an original workflow to:
- Import the “.pdresults” file.
  - Create explicit path names to access the requested data, using the Uniform Resource Identifier (URI) node of KNIME.
  - Embed the SQL commands into an RSnippet to retrieve the expected data and create two tables (Feature\_Data and Peak\_Data in the R code). These tables are further joined using a connection link between them as an intermediate table used as a “dictionary” of ID equivalents.
  - Define the ordered rank of the isotopologs extracted by sequential numbering of the lines related to each PSM (Protein Spectrum Match).
- (c) Within the R script, the complete SQL request for generating Table Feature\_Data is as follows:

```
select TargetPsms.MassOverCharge, TargetPsms.ParentProteinAccessions, TargetPsms.ParentProteinDescriptions, TargetPsms.MasterScanNumbers, TargetPsms.RetentionTime, TargetPsms.Charge, TargetPsms.Sequence, TargetPsms.Modifications, LcmsFeatures.MonoisotopicMassOverCharge, LcmsFeatures.Area, LcmsFeatures.Intensity, LcmsFeatures.NumberOfIsotopes, LcmsFeatures.Id as FeatureId
from TargetPsms, TargetPsmsLcmsFeatures, LcmsFeatures
where TargetPsms.PeptideID = TargetPsmsLcmsFeatures.TargetPsmsPeptideID
and TargetPsmsLcmsFeatures.LcmsFeaturesId = LcmsFeatures.Id
```

- (d) Within the R script, the complete SQL request for generating Table Peak\_Data is as follows:

```
select LcmsFeatures.Id as FeatureId, LcmsPeaks.MassOverCharge, LcmsPeaks.PeakHeight, LcmsPeaks.PeakArea
from LcmsFeatures, LcmsFeaturesLcmsPeaks, LcmsPeaks
where LcmsFeaturesLcmsPeaks.LcmsFeaturesId = LcmsFeatures.Id
and LcmsFeaturesLcmsPeaks.LcmsPeaksId = LcmsPeaks.Id
```

- (e) The intensity of the isotopolog ions is defined by the peak area.

- (f) We restrict the number of isotopologs during extraction in the final dataset to five, as we only require  $M_0$  and  $M_1$  for further quantification calculations.
9. The produced table is used in the bSLIM workflow for complete or incomplete labeling with the correct exogenous (non-labeled) amino acids given in parameters. The code proceeds to the ratio of  $M_1$  over  $M_0$  to quantify the molar fraction, the key variable for the quantification. The ratio of the molar fraction/(1-molar fraction) is calculated. For protein levels, all top N peptides  $\log_2(\text{Ratio})$  are grouped together to obtain classical fold changes for biological interpretations.
  10. A key question in proteomics data analysis is the distinction between noteworthy (or significant) results from other observations, which are false positives, that is, acquired by random chance. Indeed, the large amount of data arising from proteomics technologies is associated with an increase in the possibility to observe atypical “by chance” values in the dataset. In this context, statistical methodologies generally assume that all variations in the data are due to random fluctuations and, accordingly, derive a probability to observe variations that are greater than those present in the data. Random fluctuations can be modelled in two different ways. In the first, a mathematical function is chosen (often normal or student laws) and a statistical hypothesis is used to discriminate “significant” from “nonsignificant” observations, based on a predefined error rate (generally 5%). In the second, random permutations of the original dataset are performed to define empirical distributions, which will be used to assess potential random fluctuations. It is a remarkably interesting approach, especially when the theoretical probability distributions of the studied parameters are not demonstrated, as is the case with the bSLIM output dataset.

---

## Acknowledgments

We would like to thank Dr. Bernard Delanghe (Thermo Fisher Scientific) for providing us with the table structure of the “.pdresult” file that allowed us to develop the present workflow. NS received a thesis grant from CRI-Paris. This work was supported by the ARN, grant ANR-18-CE44-0014. The English text was edited by Alex Edelman & Associates (<http://www.alexedelman.com/>).

## References

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al (1996) Life with 6000 genes. *Science* 274:546. 563–567
2. Wilm M (2009) Quantitative proteomics in biological research. *Proteomics* 9:4590–4605
3. Chahrour O, Cobice D, Malone J (2015) Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *J Pharm Biomed Anal* 113:2–20
4. Leger T, Garcia C, Videlier M, Camadro JM (2016) Label-free quantitative proteomics in yeast. *Methods Mol Biol* 1361:289–307
5. Leger T, Garcia C, Collomb L, Camadro JM (2017) A simple light isotope metabolic labeling (SLIM-labeling) strategy: a powerful tool to address the dynamics of proteome variations in vivo. *Mol Cell Proteomics* 16:2017–2031
6. Senko MW, Beu SC, McLafferty FW (1995) Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *J Am Soc Mass Spectrom* 6:52–56
7. Alves G, Ogurtsov AY, Yu YK (2014) Molecular isotopic distribution analysis (MIDAs) with adjustable mass accuracy. *J Am Soc Mass Spectrom* 25:57–70
8. Ljungdahl PO, Daignan-Fornier B (2012) Regulation of amino acid, nucleotide, and phosphate metabolism in *Saccharomyces cerevisiae*. *Genetics* 190:885–929
9. Sénécaut N, Alves G, Weisser H, Lignieres L, Terrier S, Yang-Crosson L, Poulain P, Lelandais G, Yu YK, Camadro JM (2021) Novel insights into quantitative proteomics from an innovative bottom-up simple light isotope metabolic (bSLIM) Labeling data processing strategy. *J Proteome Res* 20:1476–1487
10. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K et al (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinform* 9:163
11. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Andreotti S, Ehrlich HC, Gutenbrunner P, Kenar E et al (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* 13:741–748
12. Pfeuffer J, Sachsenberg T, Alka O, Walzer M, Fillbrunn A, Nilse L, Schilling O, Reinert K, Kohlbacher O (2017) OpenMS—a platform for reproducible analysis of mass spectrometry data. *J Biotechnol* 261:142–148
13. Doblmann J, Dusberger F, Imre R, Hudecz O, Stanek F, Mechtler K, Durnberger G (2019) apQuant: accurate label-free quantification by quality filtering. *J Proteome Res* 18:535–541
14. Griss J, Stanek F, Hudecz O, Durnberger G, Perez-Riverol Y, Vizcaino JA, Mechtler K (2019) Spectral clustering improves label-free quantification of low-abundant proteins. *J Proteome Res* 18:1477–1485
15. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98:5116–5121
16. Mortimer RK, Johnston JR (1986) Genealogy of principal strains of the yeast genetic stock center. *Genetics* 113:35–43

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

