HHS Public Access

Author manuscript

*Curr Protoc Bioinformatics*. Author manuscript; available in PMC 2017 March 24.

Published in final edited form as:

*Curr Protoc Bioinformatics*. ; 53: 13.28.1–13.2816. doi:10.1002/0471250953.bi1328s53.

# Using PSEA-Quant for Protein Set Enrichment Analysis of Quantitative Mass Spectrometry-Based Proteomics

**Dr. Mathieu Lavallée-Adam** and

Department of Chemical Physiology, The Scripps Research Institute, 10550 North Torrey Pines Road, SR302, La Jolla, California, 92037, USA, Phone: 1-858-784-1000 ext. 43078, Fax: 1-858-784-8883

**Dr. John R. Yates III**

Department of Chemical Physiology and Molecular and Cellular Neurobiology The Scripps Research Institute, 10550 North Torrey Pines Road, SR302, La Jolla, California, 92037, USA, Phone: 1-858-784-8862, Fax: 1-858-784-8883

## Abstract

PSEA-Quant analyzes quantitative mass spectrometry-based proteomics datasets to identify enrichments of annotations contained in repositories such as the Gene Ontology and Molecular Signature databases. It allows users to identify the annotations that are significantly enriched for reproducibly quantified high abundance proteins. PSEA-Quant is available on the web and as a command-line tool. It is compatible with all label-free and isotopic labeling-based quantitative proteomics methods. This protocol describes how to use PSEA-Quant and interpret its output. The importance of each parameter as well as troubleshooting approaches are also discussed.

### Keywords

Gene Set Enrichment Analysis; Gene Ontology; Quantitative Proteomics; Functional Enrichment Analysis; Mass Spectrometry

## Introduction

PSEA-Quant is a web-based software package that performs protein set enrichment analyses on quantitative mass spectrometry (MS)-based proteomics datasets containing two replicates or more (Lavallée-Adam et al., 2015, 2014). Currently, PSEA-Quant allows the analysis of protein sets derived from the Gene Ontology repository (Ashburner et al., 2000), the Molecular Signature database (Liberzon et al., 2011), and the Cardiac Organellar Protein Atlas Knowledgebase (Zong et al., 2013). It supports the analysis of quantitative proteomics datasets of organisms including *Homo sapiens*, *Mouse musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*.

Correspondence to: John R. Yates, III.

## Strategic Planning

There are two main types of quantitative proteomics datasets that can be analyzed with PSEA-Quant: datasets involving a single experimental condition (Basic Protocol 1) and datasets derived from the comparison of two experimental conditions (Basic Protocol 2). The former may consist of the quantitative characterization of the proteins of a given sample (e.g. cell line, tissue, or biological fluid) or of the quantitative analysis of protein-protein interactions with techniques such as affinity purification coupled to MS (AP-MS). We refer to these datasets as absolute quantification datasets. The latter corresponds to the quantitative proteomics analysis of datasets where one experimental condition is compared to another. This includes the analysis of a sample originating from a disease state versus a healthy state or of a sample treated with a given drug being compared to a control sample. We refer to these datasets as relative quantification datasets, since protein quantification values from one condition are relative to another condition. In both cases, PSEA-Quant will identify the protein sets that are statistically significantly enriched for abundant proteins with reproducible quantification measurements. For an absolute quantification dataset, such proteins are associated with high absolute quantification values (spectral counts (Liu et al., 2004) or intensity values from extracted ion chromatograms (Chelius et al., 2003; Radulovic et al., 2004)). For relative quantification datasets, PSEA-Quant analyzes the enrichment of protein sets among proteins that are significantly more abundant in one condition versus the other (high abundance ratios).

## Basic Protocol 1

### Using Psea-Quant To Analyse A Quantitative Proteomics Dataset Involving A Single Experimental Condition (Absolute Quantification)

Basic Protocol 1 describes the PSEA-Quant analysis of a quantitative proteomics dataset containing at least two replicates for which absolute quantification values are provided for each protein in the dataset. These quantification values can be spectral counts, normalized spectral counts (e.g. NSAF (Zybailov et al., 2006), dNSAF (Zhang et al., 2010)) or intensity values obtained from extracted ion chromatograms. In this protocol, PSEA-Quant will identify the proteins sets (e.g. biological processes, molecular functions, pathways, cellular components, and disease associations) that are statistically significantly enriched for proteins with high absolute abundance that were reproducibly measured. Upon completion of its analysis, PSEA-Quant will output a text file listing the protein sets along with their estimated *q*-values, which represent the false discovery rates (FDRs). The following protocol guides you through all the steps to fill out the PSEA-Quant web form and submit a dataset to be analyzed.

#### Necessary Resources

**Hardware:** Any system that can run an Internet browser is sufficient to execute PSEA-Quant. For instance, the system requirement to execute Google Chrome is a computer with an Intel Pentium 4 processor or later with 512 MB or more of RAM.

**Software:** PSEA-Quant can be accessed at http://sealion.scripps.edu:18080/PSEA-Quant/ using any web browser, such as Google Chrome, Mozilla Firefox, Microsoft Internet Explorer, Safari, or Opera.

**File:** Input file: a tab-separated text file (.txt) in which each line corresponds to the absolute quantification values of a protein identified in a set of replicate proteomics experiments performed under the same experimental condition (at least two replicate experiments). All proteins that have been quantified in the set of replicates have to be listed in this input file. The first column contains the gene names of the identified proteins in the dataset. In this paper, gene names refer to gene symbols or gene acronyms (e.g. CFTR or CDK9). We use this terminology since it is compatible with that of UniProt. When analyzing a *Homo sapiens* dataset, you also have the option to provide UniProt accessions (Uniprot Consortium, 2014) as protein identifiers. If your dataset identifies proteins with a different identifier than gene names, follow Support Protocol 1 to convert these identifiers into gene names. The input file then contains at least two other columns (one column per replicate) listing the absolute quantification values (e.g. spectral counts or intensity values) of the proteins in each replicate. PSEA-Quant can handle any scale of abundance values, even very large intensity values, as long as the scale of the values remains similar across all replicates within a dataset. Replicates can be technical or biological replicates. There is no descriptive header at the top of the columns in the input file. It is expected that a number of proteins will not be associated with a quantification value in a given replicate simply because they were not quantified or identified in the corresponding experiment. In this case, you can input a quantification value of 0 in the corresponding column, a string of characters such as “NA”, or leave the field empty. Character strings and empty fields are interpreted as the value 0 by PSEA-Quant. See Appendix 1 for an example input file. File format examples are also accessible at this URL: http://sealion.scripps.edu:18080/PSEA-Quant/files/

1. Select “Absolute quantification” as quantification type.
2. Select the organism from which the dataset originates.

   At the time of publication, PSEA-Quant supported the following organisms:

   - Homo Sapiens
   - Mus musculus
   - Rattus norvegicus
   - Saccharomyces cerevisiae
   - Drosophila melanogaster

   More organisms will be made available over time.
3. Select the input file on which you want to perform the PSEA-Quant analysis. See the above instructions for input file format.
4. Indicate the number of samplings to be performed by the Monte Carlo procedure.

   This parameter affects the accuracy of the statistical significance p-values of the protein set enrichments and ultimately their q-values. The higher this number is,

the more accurate the p-value estimations will be. However, the running time of PSEA-Quant also increases with the number of samplings. On the other hand, a smaller number of samplings will allow a fast execution of PSEA-Quant to the expense of p-value estimation accuracy. The default number of samplings is set to 10,000 per protein sets. This number ensures a reasonable running time and p-value estimation accuracy. With 10,000 samplings the lowest non-zero p-value a protein set can be associated with is 0.0001 (1/10,000). The maximum number of samplings allowed is 100,000. See the Critical Parameters section to learn more about the adjustments of this parameter.

**5.** Select the annotation database to be used to construct the protein sets analyzed by PSEA-Quant.

The available databases at the time of publication are:

- Gene Ontology (GO) (Ashburner et al., 2000; see UNIT 7.2)
- Molecular Signature database (MSigDB) (Liberzon et al., 2011)
- Cardiac Organellar Protein Atlas Knowledgebase (COPaKB) (Zong et al., 2013)

The GO database contains the GO terms from all three subset hierarchies: cellular component, molecular function, and biological process. Upon selection of the GO database, all three subset hierarchies will be analyzed by PSEA-Quant.

MSigDB contains the following categories of protein sets: positional, curated, motif, computational, oncogenic, and immunologic. The positional protein sets group proteins based on the chromosomal position of their encoding genes. The curated protein sets are produced from online pathway databases, publications listed in PubMed, and domain experts. The motif protein sets contain proteins for which the genes share a conserved cis-regulatory motif. The computational protein sets were built by mining cancer-oriented microarray data. The oncogenic protein sets are defined from microarray gene expression data obtained in the context of cancer gene perturbations. The immunologic protein sets are built from microarray data of immunologic studies. Upon selection of MSigDB, all of these categories will be analyzed simultaneously by PSEA-Quant.

COPaKB contains a collection of protein sets of protein associations with cardiac diseases retrieved from the Online Mendelian Inheritance in Man (OMIM; see UNIT 1.2) database (Hamosh et al., 2005) and curated from peer-reviewed publications.

Both the MSigDB and COPaKB can only be analyzed with PSEA-Quant when a Homo sapiens dataset is provided.

**6.** Select the Monte Carlo sampling procedure.

There are two Monte Carlo sampling procedures that are available. The two procedures go as follows:

- Assumes protein abundance independence in the dataset.
- Assumes protein abundance dependence in the dataset (recommended).

The first option assumes that the abundance levels of proteins are independent of each other when evaluating the statistical significance of protein set enrichments. This option is not recommended unless you explicitly want to apply it based on the context of your dataset. It is discouraged since in most cases this assumption is wrong and it may yield inaccurate p-value estimations. For instance, proteins that are members of the same protein complex are typically more likely to have similar abundances than proteins that are not. The second option, which is also the default option, models protein abundance dependencies in the Monte Carlo sampling procedure to yield more accurate p-value estimations. If you select this option, you must enter a coefficient of variation tolerance factor, which models the amount of abundance dependence between proteins. This tolerance factor can be within the range of 0.1 to 1.0, where a low value signifies a higher dependence but a longer running time and a high value represents a lower dependence and a shorter running time. The default recommended value is 0.5. In cases where you suspect that there would be a greater dependency than usual for protein abundances (e.g. quantification of protein-protein interactions using AP-MS), you may want to lower this coefficient below 0.5.

**7.** Select whether the Monte Carlo sampling procedure should take into account literature annotation biases.

There are two options for this step:

- Assumes no protein annotation biases.
- Assumes protein annotations biases (recommended).

The first option assumes that each protein has the same probability to be part of a protein set when performing the Monte Carlo sampling procedure. This option is not recommended unless you explicitly want to apply it based on the context of your dataset. Some proteins are associated to more protein sets than others simply because they perform more functions, are part of more biological processes, or are studied more than others in the literature. For this reason, we recommend to use the second option where these annotation biases are considered by PSEA-Quant, which performs in this context a weighted Monte Carlo sampling procedure to produce accurate p-value and q-value estimations.

**8.** Enter the email address at which you want the PSEA-Quant analysis results to be sent.

Based on the parameters selected and the computational load on our server, PSEA-Quant runs can take a few minutes to several hours. An email will be automatically sent to the email address you have entered upon completion of the analysis.

**9.** Click the submit button to send the job to the PSEA-Quant server.

### PSEA-Quant Output

10. The page appearing after a job submission to PSEA-Quant contains the parameters that have been selected for the analysis as well as the URL to the location of the output file that will be generated by PSEA-Quant.

This URL is the same as the one that will be sent by email upon completion of the PSEA-Quant analysis.

11. Upon reception of an email from the PSEA-Quant server at the email address provided in step 8, follow the URL in the body of the email to access the PSEA-Quant output file.

You can save the text output file on your local machine and can open it using spreadsheet processing software to analyze its content. While PSEA-Quant output files typically remain available on our server for several months, they are still deleted periodically for disk space purposes. See Commentary section for more details on how to analyze PSEA-Quant output files.

## Support Protocol 1

### Converting Protein Identifiers To Gene Names

MS-based proteomics data analysis software packages used upstream of PSEA-Quant may report protein identifiers that are not gene names. Since gene names are one of the protein identifiers accepted as input by PSEA-Quant, we propose the following protocol to convert a large variety of protein identifiers to gene names.

**Hardware—**Any system that can run an Internet browser is sufficient to perform this support protocol. For instance, the system requirement to execute Google Chrome is a computer with an Intel Pentium 4 processor or later with 512 MB or more of RAM.

**Software—**Any web browser, such as: Google Chrome, Mozilla Firefox, Microsoft Internet Explorer, Safari, or Opera

A text editor, such as Vim, NotePad, or Emacs

1. Access the UniProt Retrieve/ID Mapping tool (Consortium, 2014) at this URL: http://www.uniprot.org/uploadlists/
2. Copy and paste the list of protein identifiers you want to convert into gene names in the text box under “Provide your identifiers”.

   UniProt allows for a wide variety of identifiers to be provided as input. The identifiers must either be separated by a space or by a new line. Alternatively, you can upload a text file following the same format that contains the identifiers you want to convert.
3. Select the source identifier type and choose UniProtKB as the target identifier type under “Select options”.

If your source identifiers are UniProt Accessions or UniProt IDs, the web tool will allow you to select gene names as output. In this case jump to step 13 of this support protocol.

4. Click on the “Go” button.
5. Click on the “Download” button of the result page.
6. Select “Download all”.
7. Select “Target List” as output format.
8. Select an “Uncompressed” output file.

   You may select a “compressed” output file to accelerate download speed. However, you will need to decompress it with a third-party software package in order to complete this support protocol.
9. Click on the “Go” button.
10. Open the downloaded target list file using a text editor of your choice.
11. Access the UniProt Retrieve/ID Mapping tool at this URL: http://www.uniprot.org/uploadlists/
12. Copy the list of UniProt Accessions from the target list file into the text box under “Provide your identifiers”.
13. Select “UniProtKB AC/ID” as the source identifier type and select “Gene name” as the target identifier type under “Select options”.
14. Click on the “Go” button.
15. Click on the “Download” button of the result page.
16. Select “Target List” from the view format options.
17. Associate the resulting list of gene names to their corresponding quantification values in your dataset to produce a valid PSEA-Quant input file.

## Basic Protocol 2

### Using Psea-Quant To Analyse A Quantitative Proteomics Dataset Involving Two Experimental Conditions (Relative Quantification)

Basic Protocol 2 describes the PSEA-Quant analysis of a quantitative proteomics dataset containing two replicates or more for which relative quantification values are provided for each protein in the dataset. These relative quantification values take the shape of ratios of the abundance measurements of the proteins in one condition (condition *A*) versus another (condition *B*), and can be ratios of spectral counts or intensity values obtained from extracted ion chromatograms (label-free quantification (Chelius et al., 2003; Radulovic et al., 2004), $^{15}N$ (Washburn et al., 2002), SILAC (Ong et al., 2002), $^{18}O$ (Yao et al., 2001), DiMethyl (Boersema et al., 2009; Hsu et al., 2003)) or from reporter ions (TMT (Thompson et al., 2003) and iTRAQ (Ross et al., 2004)), but must not be log-transformed. In this protocol, PSEA-Quant will identify the proteins sets (e.g. biological processes, molecular

functions, pathways, cellular components, and disease associations) that are statistically significantly enriched for proteins with high relative abundances (high A/B ratio) that were reproducibly measured. Upon completion of its analysis, PSEA-Quant will output a text file listing the protein sets along with their estimated *q*-values, which represent the FDRs. The following protocol guides you through all the steps to fill out the PSEA-Quant web form and submit a dataset to be analyzed.

## Necessary Resources

**Hardware:** Any system that can run an Internet browser is sufficient to execute PSEA-Quant. For instance, the system requirement to execute Google Chrome is a computer with an Intel Pentium 4 processor or later with 512 MB or more of RAM.

**Software:** PSEA-Quant can be accessed at http://sealion.scripps.edu:18080/PSEA-Quant/ using any web browser, such as Google Chrome, Mozilla Firefox, Microsoft Internet Explorer, Safari, or Opera.

**File:** Input file: a tab-separated text file (.txt) in which each line corresponds to the relative quantification values (abundance ratios) of a protein identified in a set of replicate proteomics experiments performed under the same experimental conditions (at least two replicate experiments). All proteins that have been quantified in the set of replicates have to be listed in this input file. The first column contains the gene names of the identified proteins in the dataset. In this paper, we refer to gene symbols or gene acronyms (e.g. CFTR or CDK9) as gene names. We use this terminology since it is compatible with that of UniProt. When analyzing a *Homo sapiens* dataset, you also have the option to provide UniProt accessions as protein identifiers. If your dataset identifies proteins with a different identifier than gene names, follow Support Protocol 1 to convert these identifiers into gene names. The input file then contains at least two other columns (one column per replicate) listing the relative quantification values (i.e. abundance ratios of a given condition *A* over a given condition *B*) of the proteins in each replicate. Replicates can be technical or biological replicates. There is no descriptive header at the top of the columns in the input file. It is expected that a number of proteins will not be associated with a quantification value in a given replicate simply because they were not quantified or identified in the corresponding experiment. In this case, you can input a quantification value of 0 in the corresponding column, a string of characters such as “NA”, or leave the field empty. Character strings and empty fields are interpreted as the value 0 by PSEA-Quant. Character strings and empty fields are interpreted as the value 0 by PSEA-Quant. See Appendix 2 for an example input file. File format examples are also accessible at this URL: http://sealion.scripps.edu:18080/PSEA-Quant/files/

1. Select “Abundance ratios” or “Abundance ratios (+ inverse ratios)” as quantification type.

   If you want to analyze the enrichment of protein sets among proteins with high abundance ratios for a condition A over a condition B then you should select “Abundance ratios”. However, if you also want to analyze the enrichment of protein sets among proteins with high abundance ratios for condition B over

condition A, then you should select “Abundance ratios (+ inverse ratios)”. When selecting this option, PSEA-Quant will compute the inverse of the inputted abundance ratios and will perform two independent PSEA-Quant analyses (one on the provided abundance ratios, and one on their inverse).

2. Perform steps 2 to 10 from Basic Protocol 1.

3. Upon receipt of an email from the PSEA-Quant server at the email address provided in Basic Protocol 1 step 8, follow the URL in the body of the email to access the PSEA-Quant output file.

   You can save the text output file on your local machine and can open it using spreadsheet processing software to analyze its content. While PSEA-Quant output files typically remain available on our server for several months, they are deleted periodically for disk space purposes. See Commentary section for more details on how to analyze PSEA-Quant output files. Of note, if “Abundance ratios (+ inverse ratios)” was selected in step 1, you will receive an email containing two URLs to the two output files produced by PSEA-Quant. The first one represents the analysis of condition A over condition B and the second one corresponds to the analysis of condition B over condition A.

## Alternate Protocol 1

### Using Psea-Quant On The Command-Line

We recommend using the PSEA-Quant web server to perform functional enrichment analyses for an up-to-date functionality. However, you may prefer to deploy your own version of PSEA-Quant on a local server. However, you should be warned that the PSEA-Quant package available for download and local deployment is not maintained as regularly as the web server.

#### Necessary Resources

**Hardware:** A Unix based system with at least 4 CPUs and 32 Gb of RAM.

**Software:** Any Linux distributions, such as Ubuntu, CentOS, or Red Hat

A Java runtime environment (version 7 and above) and a Java compiler (version 7 and above)

**File:** A PSEA-Quant input file as described in Basic Protocol 1 and 2

1. Download the PSEA-Quant source code at this URL: http://sealion.scripps.edu:18080/PSEA-Quant/files/PSEA-Quant.zip

2. Decompress the directory with the following command:

   ```
   unzip <PSEA-Quant_zip_file>
   ```

   Where `PSEA-Quant_zip_file` is the file path of the downloaded PSEA-Quant.zip file.

3. Move the current working directory to the location of the java files (src/ directory).

4. Compile all java files in the directory using the following command:

```
javac *.java
```

5. To run PSEA-Quant use the following command:

```
java PSEAQuant <Quantification_type Organism Input_file
Number_samplings Annotation_type Gene_independence
[CV_tolerance_factor] Literature_bias>
```

All parameters are described below and must be entered in the specified order. The options for each parameter refers to the same options as those that can be selected on the PSEA-Quant web form (Basic Protocol 1 and Basic Protocol 2).

`Quantification_type`: This parameter specifies the quantification type. The options are: absolute for “ `absolute` quantification”, `ratio` for “Abundance ratios”, and `ratio_inv` for “Abundance ratios (+ inverse ratios)” (Basic Protocol 1 Step 1 and Basic Protocol 2 Step 1).

`Organism`: This parameter specifies the organism from which the dataset originates. The options are: `human` for “Homo Sapiens”, `mouse` for “Mus musculus”, `rat` for “Rattus norvegicus”, `yeast` for “Saccharomyces cerevisiae”, and `fly` for “Drosophila melanogaster” (Basic Protocol 1 Step 2).

`Input_file`: This is the file name of your PSEA-Quant input file (Basic Protocol 1 Step 3).

`Number_samplings`: An integer between 1,000 and 100,000 specifying the number of samplings for the Monte Carlo procedure (Basic Protocol 1 Step 4).

`Annotation_type`: This parameter specifies the type of protein sets that should be investigated for enrichment. The options are: `go` for “Gene Ontology”, `msigdb` for “Molecular Signature Database”, and `copakb` for “Cardiac Organellar Protein Atlas Knowledgebase” (Basic Protocol 1 Step 5).

`Gene_independence`: This parameter specifies the type of Monte Carlo sampling procedure. The options are: `true` to assume protein abundance independence in the dataset or `false` to assume protein abundance dependence in the dataset (Basic Protocol 1 Step 6).

[ `CV_tolerance_factor`]: if `gene_independence` is set to `false`, a coefficient of variation tolerance factor between 0.1 and 1.0 must be entered. This parameter should not be entered if `Gene_independence` is set to `true` (Basic Protocol 1 Step 6).

`Literature_bias`: This parameter specifies if the Monte Carlo sampling procedure should consider literature annotation biases when performing its sampling. The options are: `true` to assume literature annotation biases or `false` to not assume literature annotation biases (Basic Protocol 1 Step 7).

For example, to analyze a dataset with PSEA-Quant the command line could take the following form:

```
java PSEAQuant absolute human input/CFBE_Sample_Input.txt 10000 GO
true 0.5 true
```

when `Gene_independence` is set to `true`.

Or the following when `Gene_independence` is set to `false`:

```
java PSEAQuant absolute human input/CFBE_Sample_Input.txt 10000 GO
false true
```

6. The PSEA-Quant output files will be found in the `src/output/` directory under the name of the input file appended with the following string of characters: "_Output.txt".

   If you entered `ratio_inv` for the `Quantification_type` then a second output file with the suffix "_Output_inv.txt" will be created for the analysis of the inverse ratios.

## Guidelines For Understanding Results

### Output file

Each PSEA-Quant analysis produces one output text file or two identically formatted files when "Abundance ratios (+ inverse ratios)" is selected (Basic Protocol 2 step 1). To ease visualization, the text output file can be opened in a spreadsheet processing software package such as Microsoft Excel or OpenOffice Calc. An example of an output file is shown in Figure 1. The first two lines of the output file are header lines. Each following line describes a protein set and its associated information and statistical values. The lines in the output files produced with the analysis of protein sets built from MSigDB and COPaKB are divided into seven columns listed in the following order: "Annotation description", "PES", "*p*-value", "*q*-value (FDR)", "Number of proteins with annotation in dataset", "Total number of proteins with annotation", and "Core". Each column is described below:

"Annotation description": Name of the protein set.

"PES": Protein Enrichment Score of the protein set. This internal score is used by PSEA-Quant to compute the *p*-value of the protein set. The PES is used to derive *p*-values and should not be considered to assess the significance of the enrichment of a protein set since it considerably varies based on the size of a protein set.

"*p*-value": statistical significance enrichment *p*-value of the protein set. However, this *p*-value generally overestimates the significance of protein sets. This is why a *q*-value (FDR) representing the actual significance of a protein set is provided.

"*q*-value (FDR)": false discovery rate associate to the protein set.

"Number of proteins with annotation in dataset": the number of proteins belonging to the protein set that are present in the input dataset.

"Total number of proteins with annotation": the number of proteins belonging to the protein set in the entire annotation database. This number is purely informative and is not used by PSEA-Quant in any calculations. This helps you to quickly understand if the annotation is associated to a small or to a large number of proteins.

"Core": the proteins that contributed the most to the significance of the enrichment of the protein set. These proteins are therefore the most abundant proteins and those for which the abundance was measured with the most reproducibility among all of the proteins in the protein set. PSEA-Quant reports the core of a protein set if its *p*-value < 0.001. These cores are not reported for all protein sets since their computation is time consuming. If the *p*-value is higher, the output file will label with the string "Core not computed". Otherwise, a core is reported by listing all protein identifiers composing the core in a comma-separated vector.

The output file of a PSEA-Quant analysis of Gene Ontology annotation protein sets contains an additional column providing the "GO term name" at the beginning of each line. It provides the GO term complete name of each GO term numerical identifier, which is listed under "Annotation description".

### Identifying statistically significant protein sets

Protein sets are listed in the output file in no particular order. All protein sets with reasonably small *p*-values (typically smaller than 0.1) are listed in the output file. However, as mentioned previously, the *p*-values may overestimate the significance of protein set enrichments. Assessing the statistical significance of the enrichment of a protein set is the role of the *q*-value. Typically, a *q*-value < 0.1 is considered moderately significant, while a *q*-value < 0.05 is deemed significant, and a *q*-value < 0.01 is viewed as highly significant. A protein set meeting such *q*-value thresholds is therefore statistically significantly enriched for proteins of high abundance for which the abundance levels were reproducibly measured. Only the protein sets meeting such *q*-value thresholds should be considered and reported.

When analyzing quantitative protein sets from multiple experimental conditions with multiple PSEA-Quant analyses, comparing the significant protein sets identified in each PSEA-Quant analyses of the different conditions can reveal much about the underlying cellular mechanisms and biological processes involved in the experimental conditions analyzed. For instance, take two sets of replicate absolute quantification experiments performed under two different experimental conditions and analyzed by two runs of PSEA-Quant. You can compare the protein sets and identify those that are significantly enriched in one condition for abundant proteins with reproducible quantification measurements and that

are not in the other condition. PSEA can therefore provide insights on the differences between the two experimental conditions studied.

## Commentary

### Background Information

**Gene Ontology enrichment analysis and PSEA-Quant—**Functional enrichment analysis of proteomics datasets is a very common step of most proteomics pipeline. Gene Ontology enrichment analyses that are performed by tools such as DAVID (Dennis Jr et al., 2003; see UNIT 13.11), Ontologizer (Bauer et al., 2008), GOrilla (Eden et al., 2009), and GOStat (Beissbarth and Speed, 2004) identify the over-representation of proteins involved in a particular biological process or performing a certain molecular function in a given proteomics dataset. These tools facilitate the characterization of the datasets and provide much information about the underlying mechanisms playing a role in the studied experimental conditions by examining the occurrences of the GO terms among the identified proteins. However, when applied to all identified proteins in a dataset, such analyses do not consider the protein abundance measurements nor their reproducibility in order to assess GO term enrichments. To take into account protein quantification values, GO analyses can be applied to a subset of proteins in a dataset that are either of high abundance or that are significantly differentially expressed in between two conditions. The main drawback of this kind of analysis is that it requires the establishment of arbitrary abundance or significance thresholds. Indeed, GO analyses treat two proteins with similar abundance values very differently if one is above the abundance threshold and the other is not.

By contrast, PSEA-Quant performs an enrichment analysis based on protein quantitative measurements and their reproducibility. It does so without requiring the filtering of either quantitative proteomics datasets or the use of abundance or significance thresholds. It is important to note that traditional GO analyses and PSEA-Quant ask and answer two different statistical questions about the data. Protein sets or GO terms containing several proteins in a given dataset may be significantly enriched according to a GO analysis. Nevertheless, if these proteins have a low abundance they will not be deemed significant by PSEA-Quant. Similarly, a protein set may be associated to very few proteins in a dataset and not be viewed as significant using a GO analysis, but still be considered significant by PSEA-Quant if these proteins are surprisingly abundant. Hence GO and PSEA-Quant analyses are complementary and both can be performed to analyze quantitative proteomics datasets.

**Supplementary accessibility—**PSEA-Quant is also accessible in a web-based comprehensive environment, called Integrated Proteomics Pipeline (IP2) (Integrated Proteomics Applications) that allows the analysis of large-scale MS-based proteomics datasets. The IP2 environment facilitates the use of PSEA-Quant by implementing the same interface as the PSEA-Quant web form, while allowing any quantitative datasets stored in its system to be sent automatically to PSEA-Quant without requiring the user to build the necessary input file. PSEA-Quant is also integrated in the same fashion in the Proteomics INTegrator (PINT) software package (Pankow et al., in press). PINT is a repository that

compiles proteomics results and analyses, and allows the comparison and querying of multiple proteomics datasets. You can therefore follow Basic Protocol 1 and Basic Protocol 2 and skip Basic Protocol 1 step 3 when using PSEA-Quant in the IP2 and PINT environments.

## Critical Parameters

**Number of samplings—**It is quite possible that a PSEA-Quant analysis shows that no protein sets are significantly enriched ($q$-value < 10%) among abundant proteins in a given dataset. However, it is also possible that for this dataset, PSEA-Quant requires greater $p$-value estimation accuracy in order to provide more reliable false discovery rate estimation. You should consider increasing the number of samplings for the Monte-Carlo procedure whenever at least one protein set is associated to a $p$-value of 0 but to a high $q$-value (> 10%). In such a case, you may want to perform the PSEA-Quant analysis again with 100,000 samplings. This will provide more precise $p$-values and $q$-values and may show that some protein sets are significantly enriched for abundant proteins that were reproducibly quantified. The $p$-value of a protein set is never actually 0, but they are reported as 0 for ease of visualization. A $p$-value of 0 is simply smaller than $1/s$ where $s$ is the number of samplings.

**Protein identifiers mapping to the same protein—**Some gene names may be synonyms of each other and map to the same protein. You have to be especially careful to avoid listing a given protein with the same quantification values multiple times using different protein identifiers in the PSEA-Quant input file. These can artificially lower the enrichment $p$-value of a protein set and yield under-estimated $q$-values. This may lead you to believe that such a protein set is significant when it is actually not the case.

## Troubleshooting

**PSEA-Quant yielding different results from run to run—**PSEA-Quant is based on a Monte-Carlo sampling algorithm, which involves a stochastic process. It is therefore expected that the enrichment $p$-values and $q$-values of protein sets vary in between two different PSEA-Quant analyses, even if the same set of parameters was selected. Typically, when using the recommended set of default parameters, the protein set $p$-values and $q$-values will not dramatically change. However, this characteristic remains dataset dependent. If you find that the significance values differ by an unreasonable margin or that protein sets deemed significant in one PSEA-Quant analysis are not in another, then you should increase the number of samplings of the Monte Carlo procedure. This will increase the $p$-value and $q$-value estimation accuracy and will limit the variation of the results from one run to the other.

**PSEA-Quant did not send an email with a URL to the output file—**Most PSEA-Quant analyses terminate and produce an output file within a few hours, or in some cases, a few minutes. An email is then sent to the provided address. However, some analyses may take longer based on the set of input parameters, the size of the dataset, and the server load at the time of submission. PSEA-Quant is designed to run gracefully on the vast majority of inputs, but server troubles and input format problems that were not anticipated may occur. If you have not received an email within twenty four hours of your PSEA-Quant submission, a

problem may have occurred with your analysis and you should contact our team at the email address provided at this URL: http://sealion.scripps.edu:18080/PSEA-Quant/.

### Suggestions for Further Analysis

**Representation and dissemination of PSEA-Quant's output—**PSEA-Quant results can be represented using different strategies. A table providing the protein set name (annotation description), its *q*-value and the number of proteins in the dataset associated to the protein set reports the relevant information about the results of a PSEA-Quant analysis (Table 1). A horizontal bar graph A can provide a more visually appealing representation of the results (Figure 2). Annotation descriptions can be labeled on the y-axis, while the number of proteins in the dataset associated to the annotations can be displayed on the x-axis. The *q*-value associated with each annotation can be displayed at the end of each bar. Finally, if you want to compare multiple PSEA-Quant analyses of datasets performed under different experimental conditions, you can combine multiple PSEA-Quant outputs into a heatmap with color-coded *q*-values by including all proteins sets that were deemed significant in at least one PSEA-Quant run (Figure 3). This visual representation allows the reader to quickly grasp the differences in enrichment between each experimental condition.

**GO term redundancy—**The GO database structure may cause some GO terms to be redundant when mapping them to the proteins quantified in a dataset. PSEA-Quant may therefore sometimes output a list of significant GO terms, where some of them are actually associated with the same or a very similar set of proteins in the dataset. This behavior causes the output of PSEA-Quant to be more difficult to interpret. The REVIGO web tool (Supek et al., 2011) can be used to remove redundant GO terms and make the PSEA-Quant output list of significant GO terms more intelligible. REVIGO uses a clustering algorithm to identify the representative subset of GO terms and also provides visualization tools, which help data analysis.

**Study of unsupported organisms—**We expect that more organisms will be made available in the near future for PSEA-Quant analyses using Gene Ontology protein sets. However, if you would like to perform a PSEA-Quant analysis on a dataset from an organism that is not available on the PSEA-Quant website, you can build an input dataset by mapping the protein identifiers of your organism of interest to human ortholog identifiers. This can be done using the BioMart tool of Ensembl (Cunningham et al., 2015), which is available at this URL: http://www.ensembl.org/biomart/martview. You can then submit the generated input file to PSEA-Quant selecting *Homo sapiens* as the organism analyzed. This PSEA-Quant analysis of human orthologs can shed light on the biological processes and molecular mechanisms at work in your sample.

## Acknowledgments

The authors are grateful to Claire M. Delahunty, Sung Kyu Robin Park, and Salvador Martínez-Bartolomé for helpful discussions and comments. They acknowledge funding from the following National Institute of Health grants: P41 GM103533, R01 MH067880, R01 MH100175, UCLA/NHLBI Proteomics Centers (HHSN268201000035C), and U54 GM114833. M.L.A. holds a postdoctoral fellowship from the Fonds de recherche du Québec – nature et technologies (FRQNT).

## Appendix 1

Example of a PSEA-Quant input file to analyze a quantitative proteomics dataset involving a single experimental condition (Basic Protocol 1).

## Appendix 2

Example of a PSEA-Quant input file to analyze a quantitative proteomics dataset involving two experimental conditions (Basic Protocol 2).

## Literature Cited


Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]

Bauer S, Grossmann S, Vingron M, Robinson PN. Ontologizer 2.0-a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics. 2008; 24:1650–1651. [PubMed: 18511468]

Beissbarth T, Speed T. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics. 2004; 881

Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat Protoc. 2009; 4:484–494. [PubMed: 19300442]

Chelius D, Zhang T, Wang G, Shen RF. Global protein identification and quantification technology using two-dimensional liquid chromatography nanospray mass spectrometry. Anal Chem. 2003; 75:6658–6665. [PubMed: 14640742]

UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2014 gku989.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2015. Nucleic Acids Res. 2015; 43:D662–D669. [PubMed: 25352552]

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome biol. 2003; 4:P3. [PubMed: 12734009]

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009; 10:48. [PubMed: 19192299]

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005; 33:D514–D517. [PubMed: 15608251]

Hsu JL, Huang SY, Chow NH, Chen SH. Stable-isotope dimethyl labeling for quantitative proteomics. Anal Chem. 2003; 75:6843–6852. [PubMed: 14670044]

Lavallée-Adam M, Park SKR, Martínez-Bartolomé S, He L, Yates JR III. From Raw Data to Biological Discoveries: A Computational Analysis Pipeline for Mass Spectrometry-Based Proteomics. J Am Soc Mass Spectrom. 2015:1–7.

Lavallée-Adam M, Rauniyar N, McClatchy DB, Yates JR III. PSEA-Quant: A protein set enrichment analysis on label-free and label-based protein quantification data. J Proteome Res. 2014; 13:5496–5509. [PubMed: 25177766]

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011; 27:1739–1740. [PubMed: 21546393]

Liu H, Sadygov RG, Yates JR. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem. 2004; 76:4193–4201. [PubMed: 15253663]

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell proteomics. 2002; 1:376–386. [PubMed: 12118079]

Pankow S, Bamberger C, Calzolari D, Martínez-Bartolomé S, Lavallée-Adam M, Balch WE, Yates JR III. ΔF508 CFTR interactome remodeling promotes rescue of Cystic Fibrosis. Nature. In press.

Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol Cell proteomics. 2004; 3:984–997. [PubMed: 15269249]

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell proteomics. 2004; 3:1154–1169. [PubMed: 15385600]

Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011; 6:e21800. [PubMed: 21789182]

Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Hamon C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal Chem. 2003; 75:1895–1904. [PubMed: 12713048]

Washburn MP, Ulaszek R, Deciu C, Schieltz DM, Yates JR. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. Anal Chem. 2002; 74:1650–1657. [PubMed: 12043600]

Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. Anal Chem. 2001; 73:2836–2842. [PubMed: 11467524]

Zhang Y, Wen Z, Washburn MP, Florens L. Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. Anal Chem. 2010; 82:2272–2281. [PubMed: 20166708]

Zong NC, Li H, Li H, Lam MPY, Jimenez RC, Kim CS, Deng N, Kim AK, Choi JH, Zelaya I, et al. Integration of cardiac proteome biology and medicine by a specialized knowledgebase. Circ Res. 2013; 113:1043–1053. [PubMed: 23965338]

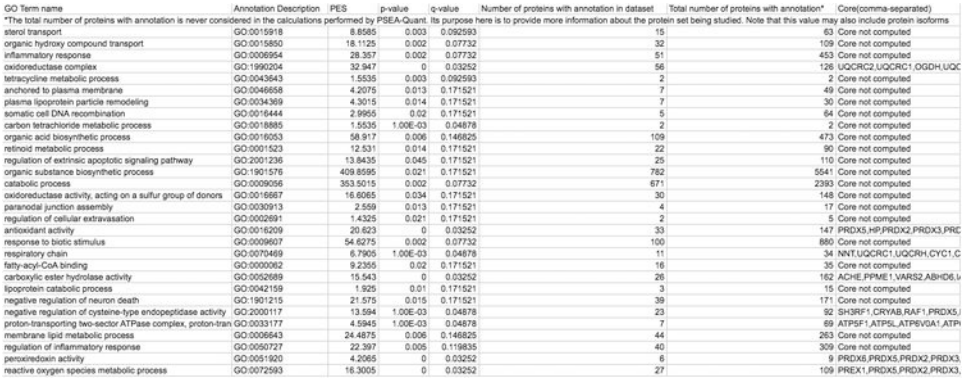
Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces c erevisiae. J Proteome Res. 2006; 5:2339–2347. [PubMed: 16944946]

## Key References

Description of PSEA-Quant's algorithm:Lavallée-Adam, M., Rauniyar, N., McClatchy, D.B., Yates III, J.R., 2014. PSEA-Quant: A protein set enrichment analysis on label-free and label-based protein quantification data. *J. Proteome Res.* 13, 5496–5509.

## Internet Resources

PSEA-Quant web server: http://sealion.scripps.edu:18080/PSEA-Quant/PSEA-Quant example input and output files: http://sealion.scripps.edu:18080/PSEA-Quant/files/PSEA-Quant source code: http://sealion.scripps.edu:18080/PSEA-Quant/files/PSEA-Quant.zipUniProt protein identifier conversion tool: http://www.uniprot.org/uploadlists/Ensembl tool for protein ortholog mapping: http://www.ensembl.org/biomart/martviewREVIGO, GO term summarizing tool: http://revigo.irb.hr/



| GO Term name | Annotation Description | PES | p-value | q-value | Number of proteins with annotation in dataset | Total number of proteins with annotation* | Core(comma-separated) |
|---|---|---|---|---|---|---|---|
| *The total number of proteins with annotation is never considered in the calculations performed by PSEA-Quant. Its purpose here is to provide more information about the protein set being studied. Note that this value may also include protein isoforms | | | | | | | |
| sterol transport | GO:0015918 | 8.8585 | 0.003 | 0.092593 | 15 | 63 | Core not computed |
| organic hydroxy compound transport | GO:0015850 | 18.1125 | 0.002 | 0.07732 | 32 | 109 | Core not computed |
| inflammatory response | GO:0006954 | 28.357 | 0.002 | 0.07732 | 51 | 453 | Core not computed |
| oxidoreductase complex | GO:1990204 | 32.947 | 0 | 0.03252 | 56 | 126 | UQCRC2,UQCRC1,OGDH,UQC |
| tetracycline metabolic process | GO:0043643 | 1.5535 | 0.003 | 0.092593 | 2 | 2 | Core not computed |
| anchored to plasma membrane | GO:0046658 | 4.2075 | 0.013 | 0.171521 | 7 | 49 | Core not computed |
| plasma lipoprotein particle remodeling | GO:0034369 | 4.3015 | 0.014 | 0.171521 | 7 | 30 | Core not computed |
| somatic cell DNA recombination | GO:0016444 | 2.9955 | 0.02 | 0.171521 | 5 | 64 | Core not computed |
| carbon tetrachloride metabolic process | GO:0018885 | 1.5535 | 1.00E-03 | 0.04878 | 2 | 2 | Core not computed |
| organic acid biosynthetic process | GO:0016053 | 58.917 | 0.006 | 0.146825 | 109 | 473 | Core not computed |
| retinoid metabolic process | GO:0001523 | 12.531 | 0.014 | 0.171521 | 22 | 90 | Core not computed |
| regulation of extrinsic apoptotic signaling pathway | GO:2001236 | 13.8435 | 0.045 | 0.171521 | 25 | 110 | Core not computed |
| organic substance biosynthetic process | GO:1901576 | 409.8595 | 0.021 | 0.171521 | 782 | 5641 | Core not computed |
| catabolic process | GO:0009056 | 353.5015 | 0.002 | 0.07732 | 671 | 2393 | Core not computed |
| oxidoreductase activity, acting on a sulfur group of donors | GO:0016667 | 16.6065 | 0.034 | 0.171521 | 30 | 148 | Core not computed |
| paranodal junction assembly | GO:0030913 | 2.559 | 0.013 | 0.171521 | 4 | 17 | Core not computed |
| regulation of cellular extravasation | GO:0002691 | 1.4325 | 0.021 | 0.171521 | 2 | 5 | Core not computed |
| antioxidant activity | GO:0016209 | 20.623 | 0 | 0.03252 | 33 | 147 | PRDX5,HP,PRDX2,PRDX3,PRD |
| response to biotic stimulus | GO:0009607 | 54.6275 | 0.002 | 0.07732 | 100 | 880 | Core not computed |
| respiratory chain | GO:0070469 | 6.7905 | 1.00E-03 | 0.04878 | 11 | 34 | NNT,UQCRC1,UQCRH,CYC1,C |
| fatty-acyl-CoA binding | GO:0000062 | 9.2355 | 0.02 | 0.171521 | 16 | 35 | Core not computed |
| carboxylic ester hydrolase activity | GO:0052689 | 15.543 | 0 | 0.03252 | 26 | 162 | ACHE,PPME1,VARS2,ABHD6,\ |
| lipoprotein catabolic process | GO:0042159 | 1.925 | 0.01 | 0.171521 | 3 | 15 | Core not computed |
| negative regulation of neuron death | GO:1901215 | 21.575 | 0.015 | 0.171521 | 39 | 171 | Core not computed |
| negative regulation of cysteine-type endopeptidase activity | GO:2000117 | 13.594 | 1.00E-03 | 0.04878 | 23 | 92 | SH3RF1,CRYAB,RAF1,PRDX5,I |
| proton-transporting two-sector ATPase complex, proton-tran | GO:0033177 | 4.5945 | 1.00E-03 | 0.04878 | 7 | 69 | ATP5F1,ATP5L,ATP6V0A1,ATPI |
| membrane lipid metabolic process | GO:0006643 | 24.4875 | 0.006 | 0.146825 | 44 | 263 | Core not computed |
| regulation of inflammatory response | GO:0050727 | 22.397 | 0.005 | 0.119835 | 40 | 309 | Core not computed |
| peroxiredoxin activity | GO:0051920 | 4.2065 | 0 | 0.03252 | 6 | 9 | PRDX6,PRDX5,PRDX2,PRDX3, |
| reactive oxygen species metabolic process | GO:0072593 | 16.3005 | 0 | 0.03252 | 27 | 109 | PREX1,PRDX5,PRDX2,PRDX3, |

**Figure 1.**

View from a spreadsheet-processing program of the top of an example of a PSEA-Quant output file.

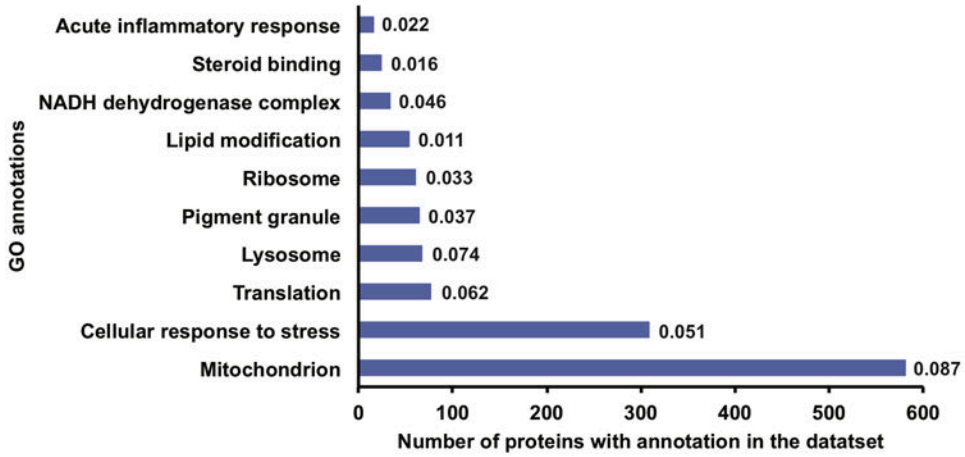



**Figure 2.**

Example of a bar graph representation of the statistically significantly enriched protein sets in a given quantitative proteomics datasets. The *q*-values for all protein sets are provided at the end of each bar.

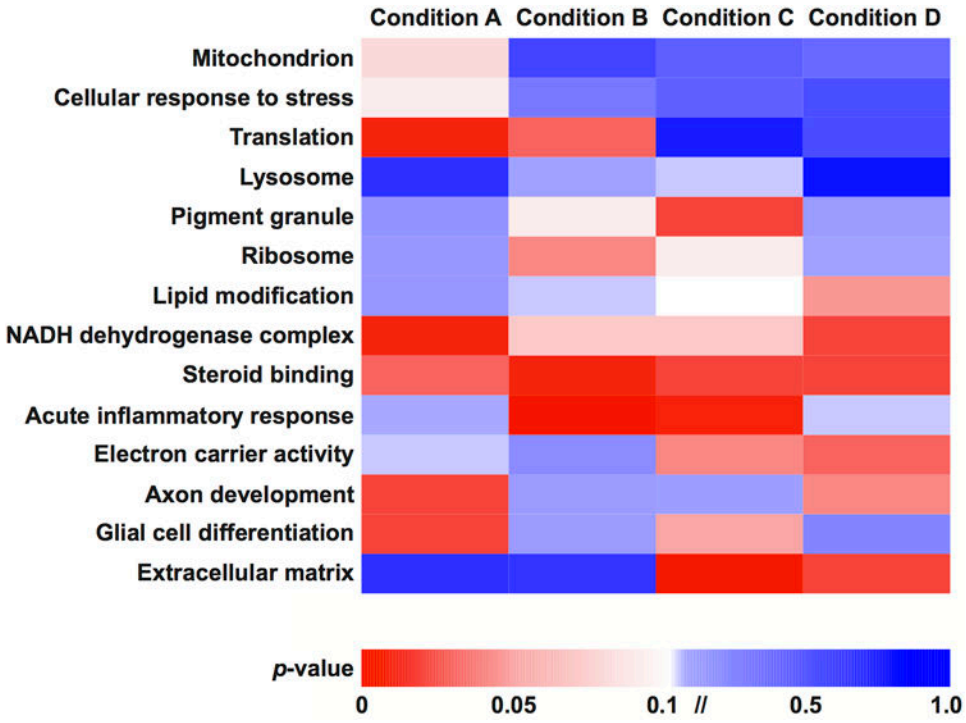



**Figure 3.**

Example of a heatmap representation of the PSEA-Quant analyses of multiple quantitative proteomics datasets obtained under different experimental conditions. The *q*-values for all protein sets are color-coded.

**Table 1**

Example of a table representation of the results of a PSEA-Quant analysis.

| Annotation Description | *q*-value | Number of proteins with annotation in dataset |
|---|---|---|
| Lipid modification | 0.011 | 54 |
| Steroid binding | 0.016 | 24 |
| Acute inflammatory response | 0.022 | 16 |
| Ribosome | 0.033 | 61 |
| Pigment granule | 0.037 | 64 |
| NADH dehydrogenase complex | 0.046 | 33 |
| Cellular response to stress | 0.051 | 309 |
| Translation | 0.062 | 76 |
| Lysosome | 0.074 | 67 |
| Mitochondrion | 0.087 | 581 |