Association of Biomolecular Resource Facilities
Proteome Informatics Research Group (iPRG)

9650 Rockville Pike, Bethesda, MD 20814
Tel: 301-634-7306 ♦ Fax: 301-634-7455 ♦ Email: abrf@abrf.org

iPRG-2015 Proteome Informatics Research Group Study: Differential Abundance Analysis in Label-free Quantitative Proteomics

Dear iPRG 2015 Study Participant,

Thank you for participating in this year's Proteome Informatics Research Group (iPRG) study. This letter provides the instructions needed to access the data files, complete your analysis, and submit your results. Results returned by Saturday, January 31, 2015 will be included in the iPRG presentation at the 2015 ABRF conference (March 28 – 31, 2015 in St. Louis, Missouri).

Overview

This study of LC-MS/MS data analysis focuses on data processing and statistical assessment for relative protein-level quantification in label-free proteomics.

The data files were derived from analysis of four complex biological samples prepared from the same *S. cerevisiae* yeast cell lysate. Several purified proteins were spiked into these four samples at known, unequal concentrations. Three technical replicate runs for each of the four samples were acquired in random order.

The goal of the study is to detect the proteins that change in relative abundance among the four biological samples. Please note that this is *not* a study comparing database search engines or MS signal processing tools. Therefore, peptide identifications and their associated integrated peak intensities have been provided as starting points of the analyses. However, if you would like to start from the raw data and/or evaluate your own signal processing tools, the raw data files have also been made available.

Sufficient information has been provided for you to perform relative quantification based on spectral counting or signal intensities of the identified spectral peaks, or both. If you use multiple approaches, please report the results from each separately.

Protein inference (the process of inferring the set of proteins present in the sample based on the set of peptides detected) is not part of this study. Therefore, peptide identifications that can be assigned to multiple proteins should be ignored. Such identifications have already been removed from the provided files.

Deliverables

An Excel file template `iPRG2015_Submission_Template.xlsx` has been provided in the directory `/distro/submission/` at the FTP site (see Study Materials below) for submission of each participant's results. (Please delete the example entries before submission.)

The file contains worksheets described below: "Quantified protein list" and six other sheets (one for each sample pair comparison) for reporting differential abundances. Instructions are provided below and at the top of each worksheet to facilitate our processing of the results.

1. *Quantified protein list*

Prepare a list of the computed quantitative measures that were used as input for the downstream statistical analyses reported on the subsequent worksheets. Provide the protein identifier followed by the corresponding quantitative measure to complete the 12 columns of values for each protein—one for each run. You can use a single measure (*e.g.*, summarized spectral counts or intensity-based value for each protein) or multiple measures that will be aggregated by the method of your choice prior to statistical analysis (*e.g.*, a combination of spectral counts and intensity associated with each protein). Use a separate row for each quantitative measure and define the quantitative measure(s) you use as specifically as you can in the submitted Excel file and in the survey (Step 3 below). In all cases, missing values should be indicated with "NA."

2. *Protein ratios worksheets* (worksheets labeled as 1vs2, 1vs3, 1vs4, 2vs3, 2vs4, 3vs4—one for each sample pair comparison)

Report the following for each of the identified proteins in each pair of samples:

Column B [$\log_2$(Ratio)]: Logarithm (base 2) of protein abundance ratio.

Column C [Estimate of confidence associated with the $\log_2$(Ratio)]: The associated statistical measure of confidence of your choice for the reported value of $\log_2$(Ratio) in Column B. Examples of statistical measures of confidence include: standard error, confidence interval, predictive interval.

Column D [Confidence measure for differential abundance]: The measure of strength of evidence for differential abundance for this pair of conditions. Examples of these measures are t-statistics, p-values, or posterior probabilities. Please sort the proteins in the worksheet with respect to this column, with the most confident (e.g., smallest p-value) at the top.

Column E [Decision regarding the differential abundance, at FDR 5%?]: Indicate "YES" if the protein is judged differentially abundant for this pair of conditions, or "NO" otherwise, according the procedure of your choice, while controlling the false discovery rate (FDR) at 5%. In other words, for every 100 proteins marked with YES in this worksheet, on average 5 of them are *not* differentially abundant between the pair of samples in question.

3. *Survey*

Please complete the brief survey at https://www.surveymonkey.com/s/NM82MCN and provide a detailed description of your methodology in the indicated textbox.

Optional Deliverables (if starting from the raw data)

In addition to the Deliverables above, we request that participants who start from the raw data submit intermediate files containing peptide identifications and/or extracted peak intensities. This is to enable us to separately assess the impact of identification and/or peak detection and integration on the results. Please submit an Excel file (or use plain-text, tab-delimited table format) that is clearly annotated.

Description of Sample Preparation and LC-MS/MS Data Acquisition

Four samples, each containing a constant background of tryptic digests of 200 ng *S. cerevisiae* (ATCC strain 204508/S288c) were separately spiked with different concentrations of several protein digests. All proteins had been reduced and alkylated with iodoacetamide prior to digestion with trypsin. The four samples were analyzed in triplicate by LC-MS/MS (total of 12 analyses) in random order. The digests were loaded directly onto a 15 cm X 75 µm PicoFrit column (New Objective) self-packed with 3 µm Reprosil-Pur C18-AQ beads (Dr. Maisch HPLC GmbH). Samples were separated using a Thermo Scientific Easy nLC 1000 system with a 110-min linear gradient of 0 – 40% acetonitrile in 0.1% formic acid at 250 nL/min directly connected to a Thermo Scientific Q-Exactive mass spectrometer. Data were acquired in data-dependent (DDA) mode, with each MS survey scan followed by 10 MS/MS HCD scans (AGC target 10E6, max fill time 60 msec), with 30-sec dynamic exclusion. Both MS and MS/MS data were acquired in profile mode in the Orbitrap, with resolution 70,000 for MS and 17,500 for MS/MS.

Study Materials

All data and relevant files can be downloaded by FTP from `ftp.peptideatlas.org`. The username is `iprg_study` and the password is `ABRF329`. If you access the site via a web browser, include the username in the URL, (*i.e.*, `ftp://iprg_study@ftp.peptideatlas.org`). The provided files are:

a. *Peptide identifications*

Peptide identifications are provided in two formats: pepXML and a tab-delimited table.

pepXML: directory `/distro/id/pepXML/`; files `ipro3-1A.pep.xml` to `ipro3-4C.pep.xml`

Tab-delimited tables: directory `/distro/id/tsv/`; files `ID_1A.tsv` to `ID_4C.tsv`

There are 12 files, each corresponding to one LC-MS/MS run, distinguished by the last two characters of the file name: 1, 2, 3, 4 refers to the sample; A, B, C refers to the replicate. For details of how the peptide identifications were performed, please refer to Appendix A of this document.

For participants who use a spectral counting approach starting from these files, typical steps may include filtering the less confident IDs, compiling the spectral counts for each protein, applying normalization or scaling, aggregating the replicates, computing fold changes, and performing statistical analyses to assess the significance of observed differences in relative quantity (see Deliverable items 1 and 2 above).

b. *MS1 Peak intensities with associated peptide identifications*

The MS1 peak intensities (integrated extracted ion chromatograms) with associated peptide identifications are provided in a tab-delimited table in directory `/distro/intensity/`. The file is `SkylineIntensities.tsv`.

The results from all 12 LC-MS/MS runs were merged into one large file and sorted by peptide ion precursor. Three isotope peaks (M, M+1, M+2) are reported separately for each precursor. For details of how the peak intensities were extracted, please refer to Appendix B of this document.

For participants who use an intensity-based relative quantification approach starting from the MS1 peak intensity file, typical steps may include filtering the less confident IDs and/or peaks less reliable for quantification (if any), applying normalization or scaling, compiling the various peak intensities into protein-level metrics, computing fold changes, and performing statistical analyses to assess the significance of observed differences in relative quantity (see Deliverable items 1 and 2 above).

c. *Raw data*

The raw data of the 12 LC-MS/MS runs are provided in two formats: Thermo native "`.raw`" format and HUPO-PSI open "`.mzML`" format.

Thermo raw files: directory `/distro/raw/`; files `JD_06232014_sample1-A.raw` to `JD_06232014_sample4-C.raw`

mzML: directory `/distro/mzML/`; files `JD_06232014_sample1-A.mzML` to `JD_06232014_sample4-C.mzML`

d. *Protein sequence database*

A suitable protein sequence database for analyzing this dataset is provided in the directory `/distro/fasta/`. The file is named `iPRG2015.TargDecoy.fasta`.

To facilitate comparison of the results, all participants who start with the raw data files are asked to use the provided sequence database without modification. For the purpose of statistical validation, for each target (yeast) protein, one equal-length decoy protein sequence was generated by shuffling amino acids randomly between tryptic sites. The decoy protein names are prefixed with `DECOY_`.

The database is in the FASTA format, with protein accession numbers and names in the UniProt style, which is compatible with most if not all software for peptide identification. *Hint:* Since some protein names contain commas, it is generally recommended to use tabs as delimiters in tables containing peptide or protein identification results.

The same database without decoys added is also provided in the file `iPRG2015.fasta`, but we strongly prefer you use our target-decoy database for peptide identification instead of generating your own.

Submission of results

As indicated above, results must be returned by **Saturday, January 31, 2015** in order to be included in the iPRG presentation at the 2015 ABRF conference. Submission of your results consists of two steps:

a. Send your completed Excel file(s) by email as an attachment to anonymous.iprg2015@my.abrf.org. Use a five-digit number of your choice as an identifier, and name the Excel file "iPRG2015submission#####.xlsx" using your identifier to replace the #####. If you started from the raw data instead of the intermediate IDs and peak intensities that we provided, please submit a separate Excel file named "iPRG2015intermediate#####.xlsx" containing your identifications and/or extracted peak intensities.

b. Please go to https://www.surveymonkey.com/s/NM82MCN and complete the study survey. This is essential for the study; we estimate it should only take 15 minutes to complete. The survey will require you to provide an identifier number. Please use the same five-digit identifier as the one you put in the name of your submitted Excel file.

The iPRG has enlisted the services of an "anonymizer" (*i.e.*, an individual who is not involved in the study and not a member of the iPRG) to collect the submitted results from your emails in order to ensure the anonymity of participants. (The anonymization process will clear the properties dialog in the Excel file, if you have not already removed identifying information before submission.)

Special note to vendors and commercial laboratories: ABRF imposes strict guidelines on the use of study results for marketing purposes. These guidelines are described at the following site:

http://www.abrf.org/Other/RG_Comm_Guidelines/Research_Group_Study_Participation_Guidelines_2010.pdf

Questions?

Please send questions to anonymous.iPRG2015@my.abrf.org. All identifying information will be removed prior to forwarding the question to the iPRG group members.

We thank you for your support of the ABRF and look forward to receiving your results for the study.

Sincerely,

The ABRF Proteome Informatics Research Group (iPRG)

Henry Lam - Hong Kong University of Science and Technology (Co-chair)
Eugene Kapp - Walter and Eliza Hall Institute of Medical Research (Co-chair)
Brett Phinney - University of California at Davis (ABRF Executive Board Liaison)
John Cottrell - Matrix Science Ltd
Michael Hoopmann - Institute for Systems Biology
Sangtae Kim - Pacific Northwest National Laboratory
Thomas Neubert - New York University School of Medicine
Magnus Palmblad - Leiden University Medical Center
Olga Vitek - Northeastern University
Sue Weintraub - University of Texas Health Science Center at San Antonio

Appendix A. Methodology for peptide identification

The MS2 spectra were searched against the provided target-decoy protein database using three sequence search engines: OMSSA [1], MS-GF+ [2] and Comet [3]. The search parameters were as follows:

OMSSA: precursor tolerance, 100 ppm; product tolerance, 0.01 Th; semi-tryptic specificity; two missed cleavages allowed; fixed modification, carbamidomethyl (CAM) on C; variable modifications, oxidation on M, cyclization on Q, E and C(CAM) at peptide N-terminus, acetylation on protein N-terminus.

MS-GF+: precursor tolerance, 10 ppm (-1, +1, +2 Da isotope error allowed); fragmentation, HCD; instrument, Q-Exactive; semi-tryptic specificity; fixed modification, carbamidomethyl (CAM) on C; variable modifications, oxidation on M, cyclization on Q and E at peptide N-terminus, acetylation on protein N-terminus.

Comet: precursor tolerance, 100 ppm (-1, +1, +2, +3 Da isotope error allowed); product tolerance 0.02 Th; semi-tryptic specificity; two missed cleavages allowed; fixed modification: carbamidomethyl (CAM) on C; variable modifications: oxidation on M, acetylation on protein N-terminus.

The results were first validated at the peptide-spectrum match (PSM) level by PeptideProphet [4], employing decoy-assisted semi-parametric modeling [5]. Accurate mass modeling was turned on to adjust probabilities based on the precursor mass deviations. Then, iProphet [6] was used to combine the results from the three search engines. The resulting confidence metric is a PSM-specific probability equivalent to (1 – posterior error probability). Note that although decoys were used to help fit the mixture distribution at the PeptideProphet stage, decoys were not used in subsequent probability adjustments. Consequently, the error estimate may not correspond to those obtained by simple decoy counting.

To provide an additional error estimate, a PSM-level q-value was also computed by decoy counting and is listed in the ID tables. Specifically, the PSMs were first sorted by descending iProphet probability, and the q-value of a particular PSM was taken to be the fraction of decoys among all IDs at or above its iProphet probability.

Decoy PSMs are retained in the ID tables in case the participant wishes to compute error estimates in some other manner.

Peptide IDs that can be mapped to multiple proteins in the database were removed from the tables.

References

1. Geer, L. Y., et al. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* 3, 958-64.
2. Kim, S., et al. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* 5, 5277.
3. Eng, J. K., et al. (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics* 13, 22-4.
4. Keller, A., et al. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 74, 5383-92.
5. Choi, H., et al. (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. *J. Proteome Res.* 7, 286-92.
6. Shteynberg, D., et al. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* 10, M111.007690.

Appendix B. Methodology used for creating the table with peak intensities

MS1 peak intensities were calculated using Skyline's MS1 Full-Scan Filtering method [1]. The data were imported from the provided mzML, pepXML, and FASTA files. Skyline was run using default parameters, except for the following: Cut-off, 0.15; Select "all" modifications; Precursor charges, 2, 3, 4; Precursor mass analyzer, Orbitrap; Use only scans within "1" minutes of MS/MS IDs; and Max missed cleavages, 2. After the data import, peptide ions and corresponding MS1 peak areas were exported using the Transition Results report in Skyline. Finally, the iProphet probability and q-value were appended as extra columns. For information about use of Skyline, refer to the following link:

https://skyline.gs.washington.edu/labkey/wiki/home/software/Skyline/page.view?name=tutorial_ms1_filtering

Peptide IDs that can be mapped to multiple proteins in the database were removed from the tables.

Reference

1. Schilling, B. et al. (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Mol. Cell. Proteomics* 11, 202-14.

Appendix C. Useful information and links for statistical analysis.

a. Generic statistical analysis methods

Generic method, such as Student's t-test, ANOVA, or Mann-Whitney (equivalently, Wilcoxon) test are implemented in a variety of software tools. Examples are:
http://www.excel-easy.com/data-analysis/analysis-toolpak.html (Excel)
http://www.statmethods.net/stats/index.html (R)

b. Specialized methods based on counts of MS/MS spectra

A perspective on working with spectral counts is discussed in this article:
http://www.ncbi.nlm.nih.gov/pubmed/20121475

Examples of statistical analysis tools specific for proteomics are:
(QPROT and QSPEC) http://www.nesvilab.org/software.html
(Quasitel) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2920032/

Tools for finding differentially expressed genes in RNA-seq experiments may also be useful:
(DESeq2) http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html

c. Specialized methods based on peak intensities

A review of statistical methods based on peak intensities:
http://www.ncbi.nlm.nih.gov/pubmed/23019139

Examples of statistical analysis tools specific for proteomics are:
(MSstats) http://www.msstats.org/
(DanteR) http://omics.pnl.gov/software/danter

Tools for finding differentially expressed genes in microarray experiments may also be useful:
(Limma) http://www.bioconductor.org/packages/release/bioc/html/limma.html

d. Tools for statistical analysis starting from raw data, based on peak-intensity

These tools can also perform signal processing and provide data visualization. Examples include:
(MaxQuant/Perseus) http://www.maxquant.org/
(OpenMS) http://open-ms.sourceforge.net/
(Skyline) https://skyline.gs.washington.edu/labkey/project/home/software/Skyline/begin.view
(IDPicker 3) http://fenchurch.mc.vanderbilt.edu/software.php