Research • Technology
Communication • Education

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

Re: iPRG-2013: Proteome Informatics Research Group Study: *Using RNA-Seq Data to Refine Proteomic Data Analysis*

Dear iPRG 2013 Study Participant,

Thank you for participating in this year's Proteome Informatics Research Group study. This letter provides the instructions needed to get the data, complete your analysis, and submit your results. ***Results returned by Friday, February 8, 2013 will be included in the iPRG presentation at the 2013 ABRF conference (March 2 – 5, 2013 in Palm Springs, California).***

Overview of the Analysis Task

This collaborative LC-MS/MS data analysis study focuses on the contribution of different reference databases to the ability of proteomics laboratories to identify peptides present in a complex mixture.

This study requires you to perform the following bioinformatic analyses:

- Assign the CID spectra present in the sample with < 1% false discovery rate (FDR) at the unique peptide/sequence level for matches to multiple different target databases.
- Indicate in the results file any peptide identifications that were only (confidently) detected using a database derived from the RNA-Seq data.
- Complete a brief survey and attach a 1 – 2 page description of your methodology.

We are providing a common dataset (in several equivalent file formats), four types of reference protein databases, a DNA database and a complete RNA-Seq dataset. The mass spectrometry data were acquired on an LTQ-Orbitrap Velos mass spectrometer (Thermo), where precursor ions were measured in the orbitrap with high mass accuracy and resolution, and fragmentation was performed by HCD, also measured in the orbitrap. We ask you to provide a full list of *spectral* identifications despite requesting thresholding at the *unique peptide* level. We ask that you provide a comprehensive report of identifications that are both above and below your 1% FDR threshold at the unique peptide level and indicate matches to the decoy database. This will provide information about the number of correct answers that were found, but were not of high enough confidence to be above the acceptance threshold.

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

The data are derived from a whole cell lysate of human peripheral blood mononuclear cells. This is part of the proteomic data from an integrated personal omics profiling study [Cell 2012 148(6):1293-1307]; details of sample preparation can be found in the cited paper. Tryptic peptides were separated into 14 fractions by high pH reverse phase chromatography; each fraction was analyzed by LC-MS/MS using a 240-minute low pH reversed phase separation. Note: the 6plex tandem mass tag (TMT) reagents were employed for labeling these samples and cysteines were carbamidomethylated, so these modifications should be considered in the analysis.

Study Materials

Data formats: The dataset is available in several formats, enabling the use of many different software tools. You may use the format(s) of your choice. You are welcome to try multiple input types, although

we ask that you submit only one result set. The databases needed for analysis are being supplied, although you can create your own, providing that you submit your database(s) along with your results.

Materials checklist: The materials you will need to complete the analysis can be downloaded from: ftp.peptideatlas.org/ (in a web browser enter ftp://iprg_study:ABRF2929@ftp.peptideatlas.org/)
user: iprg_study
password: ABRF2929

a. Databases – Several databases are provided, as described below. You do not need to search against all of these databases. However, we ask that you do search against the "Homo_sapiens_non-redundant.GRCh37.68.pep.all_FPKM-cRAP_targetdecoy" database (or an equivalent version of this database that allows you to estimate FDR) and compare results from any other database to those achieved using this database. Each of the databases listed below is provided with and without decoy scrambled protein sequences appended.

1. **Homo_sapiens_non-redundant.GRCh37.68.pep.all_FPKM-cRAP**: Sequences of all non-redundant proteins derived from the Ensembl human gene database, plus those of common contaminant proteins.
2. **Homo_sapiens_non-redundant.GRCh37.68.pep.all_FPKMgt0-cRAP**: A subset of database 1 that only contains proteins for which there was evidence of expression (at any level) based on the RNA-Seq data.
3. **Homo_sapiens_non-redundant.GRCh37.68.pep.all_FPKM_SNV-cRAP**: All entries in database 1 plus additional sequences containing single amino acid substitutions compared to reference Ensembl gene sequences. These variants are derived from identifying single nucleotide variants in the RNA-Seq data.
4. **Homo_sapiens_non-redundant.GRCh37.68.pep.all_FPKM_NOVEL-cRAP**: All sequences in database 1 plus additional sequences corresponding to a six-frame translation of all sequences identified in the RNA-Seq data that could not be mapped to a gene sequence in Ensembl.
5. **possible_novel_transcript_dna**: DNA sequence database corresponding to all sequences identified in the RNA-Seq data that could not be mapped to a gene sequence in Ensembl (i.e., the sequences that underwent a six-frame translation to create database 4). *(Note: most database search engines will not be able to search this directly.)*
6. **SNV_peptide**: Protein sequence variants derived from detection of single nucleotide variants in the RNA-Seq data (i.e. the sequences that were added to database 3).
7. **BloodRNA_trimmed.fastq**: Compiled database of RNA-Seq data prior to any attempts to map it to gene sequences. *(Note: this cannot be used for database searching; it is only useful for compiling your own protein databases.)*

All databases with "FPKM" in their name contain information in the header for each protein about expression level observed in the RNA-Seq data. We have provided a database filtered to contain proteins with non-zero FPKM values, but participants could try to use different expression level thresholds to create different filtered versions of any of these databases. We are supplying a perl script "filterByFPKM.pl" that can be edited and used for this purpose.

In relevant databases, target and decoy sequences are distinguishable by their headers, as shown below:
Target:
>ENSP00000414185 ENSP00000414185
Decoy:
>DECOY_ ENSP00000414185 Decoy sequence

b. Data – The provided dataset is available in three different formats.
.raw – Raw data as acquired on the LTQ-Orbitrap Velos. Note these are very large files because all data were saved in profile mode.
.mzML – created from .raw file using ms-convert in ProteoWizard v2.2.
.mgf – created from .raw file using ms-convert in ProteoWizard v2.2.

c. Results template (Excel) – Use this template to submit your results. Other formats will not be accepted.

Results Submission

There are two steps needed to complete your contribution to the study that must be completed on or before Friday, February 1, 2013.

a. Send your results by email as an attachment using the required Excel template (preferably zipped) to anonymous.iPRG2013@gmail.com, naming the Excel file "iPRG2013submission#####.xls", using your identifier to replace the ##### section.

If you created your own database for searching, please submit a compressed version with your results.

b Please go to: http://www.surveymonkey.com/s/FKS56TK and complete the study survey. This is essential for the study; we estimate it should only take 15 minutes to complete. The survey tool will require you to create an identifier number. Please use the same identifier as you put in the name of your results 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).

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 url:

http://www.abrf.org/Other/Research%20Group%20and%20Committee%20Guidelines/Vendor_Support_of_RG_Studies_2012.pdf

Where to Send Questions

Please send questions to anonymous.iPRG2013@gmail.com. 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 your participation in this year's study.

Sincerely,

The ABRF Proteome Informatics Research Group (iPRG)

Robert Chalkley – UCSF (Chair)
Nuno Bandeira - UCSD
Matt Chambers – Vanderbilt University
John Cottrell – Matrix Science Ltd
Eric Deutsch - Institute for Systems Biology
Eugene A. Kapp - WEHI
Henry Lam - Hong Kong University of Science and Technology
Ruixang Sun – Chinese Academy of Sciences
Olga Vitek – Purdue University
Sue Weintraub – Univ. of Texas Health Science Center at San Antonio
Thomas Neubert (EB Liaison) - New York University