The Association of Biomolecular Resource Facilities

Research · Technology · Communication · Education

The aims of the Metabolomics Research Group are a) to educate research scientists and resource facilities in the analytical approaches and management of data resulting from comprehensive metabolite studies and b) to promote the science and standardization of metabolomic analyses for a variety of applications. Group efforts will also include conducting surveys and organizing sessions at the annual ABRF meeting to explore the current state of the art in the field and the organization of research studies.

ABRF-MRG2016 Metabolomics Research Group Data Analysis Study

Dear Colleagues:

Thank you for participating in the ABRF-MRG2016 Metabolomics Research Group (MRG) study. All data and relevant files can be downloaded by FTP from:

http://bioshare.bioinformatics.ucdavis.edu/bioshare/view/tthvputgo2am4yo/

Downloadable files are listed under ***Study Materials***.

IMPORTANT: All results are submitted anonymously, identified only by a 5-digit identification number that you select, by emailing to:

anonymousmrg2016@gmail.com

In addition we ask you to complete the online MRG2016 Survey Questionnaire available at:

https://www.surveymonkey.com/r/abrfmet2016

The primary goal of this study is to examine reproducibility and optimal data analysis strategies for metabolomics studies, by comparing a consortium of analyses using the same dataset. In addition, this study will provide participants with an opportunity to evaluate their approaches with regard to the following:

- Determining relative quantitative metabolite differences across two sample types.
- The effects of different computational techniques on the determination of significantly altered metabolites in the two groups.
- Assessing the level of confidence and consistency in the results obtained from unique computational and chemometric approaches.
- The ability of software to determine differences across samples or help analyze data from metabolomics experiments.
- Databases used for assigning metabolite ID.

Overall Study Design

Metabolomics is an evolving field. One of the major bottlenecks in the field is the varied application of bioinformatics and statistical approaches for pre- and post-processing of global metabolomic profiling data sets collected using high resolution mass spectrometry platforms. Several publications now recognize that data analysis outcome variability is caused by different data treatment approaches [1-3]. Yet, there is a lack of inter-laboratory reproducibility studies that have looked at the contribution of data analysis techniques toward variability/overlap of results.

Thus our study design recapitulates a typical metabolomics experiment where the goal is difference detection of features between two groups. Specifically, for this study we have used urine samples (a commonly used matrix for metabolomics-based biomarker studies) from mice exposed to 5 Gray of external X-ray and those exposed to sham irradiation (control group). We have used five biological replicates for each group. The urine extracts were separated on an H-class Acquity UPLC and the data acquired on a quadrupole time-of-flight mass spectrometer operating in positive or negative ionization modes. The sample queue was randomized to remove injection bias. A mixture of standard compounds (referred to as metmix composed of 1ug/mL acetominophen, val-tyr-val, sulfaguanidine, sulfadimethoxine, leucine-enkephalin, terfenadine) was injected at the beginning and end of the sequence to monitor mass accuracy while pooled quality control samples were used to monitor data reproducibility. The presentation ***20151010_MRG_ABRF_study_TOFMS_vf1.pptx*** (accessible from the ftp site https://bioshare.bioinformatics.ucdavis.edu/bioshare/view/tthvputgo2am4yo in the folder labeled MRG Sample Info) demonstrates that the mass accuracy was within 5 ppm while the solvent blanks indicated minimal carryover. The MRG has also collected fragmentation data using MS^E (elevated collision energy) with the QC samples and these files are available upon request for metabolite identification based on fragmentation patterns. However, we expect that only those participants who use *Waters* software would be able to take advantage of these data.

As described in detail below, the data are made available in raw as well as pre-processed formats (netCDF). We ask the participants to report on the **top 50 features** that show statistically significant differences in relative abundances between the two study groups using data analysis tools/workflows that are routinely used in the laboratory.

At the conclusion of this study we expect to inform the scientific community about the current status of metabolomics data analysis strategies, the possible causes of variability in data pre-processing, difference detection, and metabolite identification.

Deliverables

An Excel file template ***MRG2016_Submission_Template.xlsx*** has been provided in the directory at the FTP site (see ***Study Materials*** below) for submission of each participant's results. Please delete the example entries before submission. There are two reporting sheets on this excel file. The first sheet is the peak detection sheet (if you have pre-processed the data using the net CFD files) for reporting the number of m/z detected and the relative abundance of each peak across every sample.

The second sheet is meant for reporting difference detection for top 50 significantly changes metabolites. You can use additional sheets for reporting the visualizations you used (for ex. PCA or volcano plots)

1. Sheet 1: Peak Detection Worksheet

This sheet is meant for reporting the pre-processed output (if you have used net CDF files). Prepare a list of the reported quantitative measures for each feature detected in each sample that were used as input for the downstream statistical analyses reported on the subsequent worksheets (reporting template has been provided in an Excel format). For each row indicate the feature identifier followed by the complete quantitative measure, or values for each metabolite corresponding to each sample — one per sample run. Provide a single quantitative value for the spectral abundance of each feature that was normalized based on the internal standards detected for each run, as follows:

a. 4-nitrobenzoic acid ($O_2NC_6H_4CO_2H$; MW: 167.12), m/z 166.0141 $(M-H)^-$

b. debrisoquine ($C_{10}H_{13}N_3$; MW: 175.23), m/z 176.1187 $(M+H)^+$

2. **Sheet 2:** ***Difference Detection metabolite list - feature ratios worksheet***

This sheet is meant for reporting the top 50 features/peaks/metabolites that show significant change when comparing sham and irradiated groups. You can use the XCMS output (.csv files in the NEG and POS modes) or pre-processed data file tha you generated using the netCDF files.

Report the following for each of the identified features in the two study groups (and label worksheets as one for the pair comparison. (Control vs 5 Gy):

- Column A, spectral feature ID written as m/z
- Column B, retention time (in minutes)
- Column C, ESI mode
- Column D, putative ID (based on accurate mass based database search)
- Column E, data base used
- Column F, ppm error (deviation from expected m/z)
- Column G, putative adduct
- Column H, test of significance value (ex. p-value)
- Column I, Irradiated/Sham (Relative Ratio)
- Column J, [Decision regarding the differential abundance, at FDR 5%?]: Indicate "YES" if the metabolite is judged differentially abundant for this pair of conditions, or "NO" otherwise, according to the procedure of your choice, while controlling the false discovery rate (FDR) at 5%. In other words, for every 100 metabolites marked with YES in this worksheet, on average 5 of them are *not* differentially abundant between the pair of samples in question.

3. Sheet 3: ***Description of Data Analysis techniques for files or figures generated***

Please provide visual representation of the results. For example, this could be a PCA or a heatmap.

4. *Survey*
Please complete the brief online survey at

https://www.surveymonkey.com/r/abrfmet2016

and provide a detailed description of your methodology in the appropriate textboxes.

5. *Summary*
In addition to completing the online survey, the MRG requests that each participant prepare a short write-up that summarizes the approach that was taken, the methods that were used, and the key findings that were obtained. These anonymous write-ups will be posted online and linked to each participant's results. These write-ups provide each participant the opportunity to communicate their results in their own words and share important details about the analysis that may not have been captured in the online survey.

Please e-mail your anonymous write-up as a pdf file to

anonymousmrg2016@gmail.com

The file name should match the **5-digit code** that you entered at the beginning of the online survey.

Optional Deliverables (if starting from the raw data i.e. .raw files)
In addition to the deliverables above, we request those participants who start from the raw data to submit intermediate files containing data pre-processing parameters and small molecule identifications and/or quantitative values (i.e., abundance) for extracted peaks. This information will enable us to separately assess the impact of identification and/or peak detection from alternative sources in tandem with the integration on the results. Please submit an Excel file (or use plain-text, comma-delimited table *.csv format) that is clearly annotated with m/z, corresponding retention time and abundance values across samples.

Description of Sample Preparation and ESI-QTOF-MS Based Data Acquisition
The 11 samples include 5 controls, 5 irradiated samples and 1 quality control (QC). The control group provided a baseline measurement for sham-irradiation (control) profile. The irradiated sample group was collected from mice that were exposed to 5 Gray (Gy) of external X-ray. The pooled QC was prepared by combining 5 µL from each extract of each sample from the two groups. Metabolite extraction and protein precipitation were performed as described on slide 10 of presentation entitled ***20151010_MRG_ABRF_study_TOFMS_vf1.pptx*** located in the folder labeled **MRG Sample Info** at

http://bioshare.bioinformatics.ucdavis.edu/bioshare/view/tthvputqo2am4yo/

The analytes were separated using a Waters Xevo G2 QTOFMS system with an 11 min. binary gradient of water with 0.1% formic acid (Solvent A) and acetonitrile with 0.1% formic acid (Solvent B) as detailed on slide 9, ***20151010_MRG_ABRF_study_TOFMS_vf1***. Data were acquired using untargeted profiling with each survey performed at scan speed of 0.3s, capillary voltage of 2.5kV (ES+) and 0.8kV (ES-), cone voltage of 15V (ES+) or 30V (ES-), source temperature of 120^{o}C, desolvation temperature of 500^{o}C, and desolvation gas flow 1000L/h. An external standard of 2ng/uL leucine-enkephalin reference was infused at a rate of 10 µL/min

from separate Lock Mass Correction channel at the ion source to maintain in-run mass correction. The QC pooled sample was analyzed with 6 replicates over the course of the experiment. A total of 16 UPLC-QTOFMS centroid surveys were performed in both ESI positive and negative modes.

Analysis
The MRG requests that participants A) pre-process the raw data using a pre-processing software of their choice and provide a data matrix consisting of m/z, retention time, and ion intensity (e.g., peak area); B) post-process the data using statistical tools and determine the **top 50** most perturbed (statistical significance e.g., p-value) urinary spectral features post exposure to 5 Gy external beam irradiation in mice; and, C) assign putative identification to these urinary spectral features using various online databases. The results should be submitted using the data submission template named as MRG2016_Submission_Template.xlsx. Additional information about the analysis details can be submitted while completing the survey. Features are present in amounts suitable for the use of mass spectrometry to determine the relative abundance ratios of features across the two groups indicating significance of change in the urinary excretion levels as determined by statistical tests (e.g., t-test). We recommend that the analysis of the two groups be performed using different statistical methods to get an idea of the variability of the methods used.

Study Materials

You can access and download the data using the following link:

http://bioshare.bioinformatics.ucdavis.edu/bioshare/view/tthvputgo2am4yo/

MRG Data Folder contents:

1. MRG Sample Info folder, which contains:
    a. *20151010_MRG_ABRF_study_TOFMS xv.xlsx* file includes sample annotation for biological samples, technical replicates and pooled qc samples filenames as labeled in netcdf and RAW data formats.
    b. Two pre-processed xcms files
    20151015_XEVO_MRG_URN_NEG_XCMS_report vf.csv
    20151015_XEVO_MRG_URN_POS_XCMS_report v1.csv
    c. 20151010_MRG_ABRF_study_TOFMS_vf1.pptx (QC presentation)

2. MRG folder, which contains:
    a. 25 netcdf files for positive mode and 25 netcdf files for negative mode
    b. RAW Masslynx Data folder with zipped files including the negative mode and positive mode datasets
    c. *Pooled Sample Fragmentation Data* folder, zipped RAW files containing fragmentation data for pooled qc samples from each mode.

Returning Data

As with past MRG studies, result submissions will be coded to ensure **anonymity** of the participating laboratories. A summary of the study outline will be available at the ABRF2016 meeting and will be subsequently posted on the MRG homepage:

https://abrf.org/research-group/metabolomics-research-group-mrg

Please return your results no later than October 01, 2016.

Please note: if your computer accepts cookies, you can return to the form and make changes later as long as you enter the data from the same computer. Otherwise, you can fill out the entire survey again using the same identification code so that we can recognize the duplication and ignore your previous entry. **Most important**, in order for you to obtain feedback about your analyses you need to choose a **unique 5-digit identification number** that only you will know. The number should be entered into the online data analysis form. The data analysis results will be identified by the number that you pick. **We would like to emphasize that information about unsuccessful analyses is as vital to this study as are successful analyses.** All results are compiled in a completely **anonymous** manner, so there is absolutely no need to feel shy about submitting negative results.

Please fill out the online data analysis form and e-mail your data to

anonymousmrg2016@gmail.com

regardless of your results. This study is not a contest!!!

If necessary, we will post updates about the study on the MRG homepage:

https://abrf.org/research-group/metabolomics-research-group-mrg

If you have any questions about filling out the online data analysis form, please e-mail

Amrita Cheema at *akc27@georgetown.edu* or Chris Turck at *turck@psych.mpg.de*

We thank you for your support of the ABRF and we look forward to your participation in this study!

The ABRF Metabolomics Research Group

Amrita Cheema	Georgetown University
Allis Chien (EB Liaison)	Stanford University
Maryam Goudarzi	Georgetown University
Tytus Mak	NIST
Andrew Patterson	Pennsylvania State University
Chris Turck (Chair)	Max Planck Institute

References

1. Mahieu, N.G., J.L. Spalding, and G.J. Patti, *Warpgroup: increased precision of metabolomic data processing by consensus integration bound analysis.* Bioinformatics, 2016. **32**(2): p. 268-75.
2. Klupczynska, A., P. Derezinski, and Z.J. Kokot, *Metabolomics in Medical Sciences--Trends, Challenges and Perspectives.* Acta Pol Pharm, 2015. **72**(4): p. 629-41.
3. Kessler, N., et al., *ALLocator: an interactive web platform for the analysis of metabolomic LC-ESI-MS datasets, enabling semi-automated, user-revised compound annotation and mass isotopomer ratio analysis.* PLoS One, 2014. **9**(11): p. e113909.