Recommendations for Enhancing the Use and Reuse of Cancer Metabolomics Data

Prepared by the NMR Interest Group within the Metabolomics Association of North America

Funding agencies and the metabolomics community have initiated and built repositories for cancer metabolomics data. Notable examples of these data repositories include the Metabolomics Workbench [1], MetaboLights [2], and MassIVE [3], among others. This is an important step towards enabling the use and re-use of diverse types of metabolomics data including cancer metabolomics data. Researchers in the Nuclear Magnetic Resonance (NMR) Interest Group within the Metabolomics Association of North America (MANA) have extensive collective experience in generating, using, and reusing metabolomics data sets. We provide herein our recommendations for best practices based on our personal experiences that we hope will maximize the utility of these data and enhance the overall scientific impact of metabolomics.

1. Experience in generating, using, and reusing metabolomics data sets:

- **Experience in generating and depositing metabolomics data sets**
  Members of the MANA NMR Interest group and the NMR community in general have extensive metabolomics experience, generating numerous NMR and mass spectrometry (MS) metabolomics data sets. While some of our metabolomics data sets have been deposited in public repositories such as Metabolomics Workbench or MetaboLights, a majority of these data sets remain with individual laboratories or research groups. The final destination of data created by scientists outside of the metabolomics community is unclear.

  There are several practical barriers to depositing metabolomics data sets into public repositories. The fact that these data repositories are a relatively new resource is a prime reason for the limited number of currently deposited data sets. Next, the process of depositing a metabolomics data set into a repository tends to be time-consuming, cumbersome, and frustrating without any perceived benefit to the investigator. There may be an unwillingness or inability to deposit data due to the substantial time requirement, intellectual property conerns, institutional restrictions or other regulations. Simply, there is not a strong community incentive since most journals or granting agencies do not require the deposition of metabolomics data into repositories to complete the publication of a manuscript or to adhere to funding requirements.

- **Experiences in using metabolomics data sets**
  Most NMR metabolomics experts typically only have experience in using NMR metabolomics datasets with limited experience with MS metabolomics data. Similarly, MS metabolomics experts typically have limited to no experience with NMR data sets. In effect, there is a small (but slowly growing) group of investigators with experience utilizing varied platforms including NMR and MS data sets (*i.e.,* multi-platform metabolomics). Thus, the use of MS data by NMR experts (and vice versa) is not always straightforward and may likely require assistance from an expert. Due to the complementary nature of NMR and MS metabolomics data, it is also important to have the necessary tools and information to integrate these two types of data and to train researchers to proficiently use these resources.

- **Experiences in reusing metabolomics data sets in data repositories**
  There are also substantial barriers to re-using metabolomics data that have been publicly deposited. For example, the raw or interpreted data, original experimental parameters, processing protocols, and/or the relevant software details may be missing, not defined, or unavailable. Also missing could be the type of statistical analysis methods, criteria for identifying statistical significance, experimental design, and whether relative or absolute concentration changes were used. Without the necessary metadata (for example the experimental design information), it is impossible to carry out secondary processing and analysis of the data. In addition, proper interpretation of the existing resultant analyses in the repository is limited.
    - **Experiences in reusing data in Metabolomics Workbench**
      The Metabolomics Workbench is an extensive data repository that includes metadata and experimental metabolomics data for 2,000+ studies. Tremendous progress has been made in building this repository since its inception in 2013 and it has significantly contributed to making metabolomics data easily

accessible to the scientific community. However, some technical issues remain that limit the ease of reusing the data in this or other repositories. These issues include, but are not limited to:

- Unavailability of the raw data for a substantial portion of the studies.
- Lack of mapping between the raw data files to the samples in the experimental design. Even though it is still possible to reprocess the raw data files, the missing mapping prevents any statistical analysis of the data to make biological interpretations.
- Unable to convert raw LC-MS, GC-MS or NMR data into open data formats (mzML, mzXML, CDF, nmrML *etc.*).
- Lack of unambiguous metabolite identification or validation of metabolite identification.
- Lack of in-house physical reference standards or publicly available compound library information used for the reported, identified, or annotated metabolites.
- Insufficient data curation
- Limited query capabilities.
- Lacking capabilities to visualize processed or raw MS and NMR data.
- Large portion of the data embargoed for long periods after deposition.
- Deposition or data retrieval failures due to a range of technical problems

Please see publication [4] here: https://pubs.acs.org/doi/10.1021/acs.analchem.1c00355 for an example workflow.

2. Data and metadata needed for use and reuse of metabolomics data. Significant effort has been invested in building consensus reporting standards for metabolomics data [5-8]. Here we revisit the issues and provide a current perspective from the MANA NMR Interest group.

- **Required Data**
  - raw data from study samples, QC samples, and blank samples in open data formats (mzML, mzXML, CDF, *etc.*) and vendor file formats (Bruker, Agilient, JEOL, Thermo Fisher, Sciex, LECO, Waters, *etc.*) for using and reusing the raw data
  - Result tables following data preprocessing, processing, and the analysis produced by the original data generators for using and reusing such results
- **Recommended Meta-Data**
  - Accurate mapping of sample identifiers to raw and/or processed data file names
  - Defined ontologies that help automation/reuse and standardize the field.
  - Detailed information on experimental design and factors impacting results or the nature of the study
  - Number and type of experimental groups, number of biological/analytical replicates per group, number/type of controls.
  - Instrumentation details: manufacturer, software version, spectrometer frequency, nuclei, NMR probe, mass analyzer, automation used (SampleJet, automatic tune match, *etc.*)
  - Analytical method used (1D/2D NMR, LC-MS, GC-MS), NMR pulse sequence used with parameters, use of isotope labeling, *etc.*
  - LC/GC details: column type, column dimensions, solvents, gradient-elution parameters.
  - Detailed description of sample type, sample collection, handling, and sample preparation:
    - pH, buffer, solvent, temperature, chemical shift/mass internal reference.
    - Cell/tissue lysis or homogenization method (sonication, bead-beating, *etc.*).
    - Precipitation or filtering to remove large biomolecules
    - Sample storage temperature and duration (*if available*).
  - Description of the quality control strategy and QC results
    - pooled samples (pooling strategy and use), buffer blanks, extraction blanks, standards, internal/external references, sample preparation run order and/or instrument analysis order.
  - Description of processing/preprocessing parameters
    - Baseline correction, phasing, normalization and scaling method, window function, zero-filling, removal of spectral regions, alignment, and reference methods
    - Binning/bucketing, peak-picking/feature selection criteria (CV, %missing, fold-change), missing data imputation method
    - Description of the system environment (OS, RAM, CPU, Python version, R version), or Docker if available

- Description of statistical methods and validation
  - Univariate and/or multivariate methods, artificial intelligence, or deep learning methods
  - Minimal fold change, reported *p*-value for significance, false-discovery rate or multiple hypothesis correction, reported R^2/Q^2 values, proper validation methods reported for supervised statistical models
- Public availability of software tools to reproduce the original data used for processing and analyzing the data and for statistical and machine learning analysis.
  - software versions, data processing and analysis parameters or scripts
- Persistent link to source code of in-house software/tools used for analyzing the data

- **Other valuable resources**
  - Link to published paper(s) describing the deposited data set and/or relevant to associated study.
  - Link to published papers or documents describing the experimental procedures or protocols applied to the deposited data.
  - The processing of NMR raw data can be facilitated by a data repository integrating with NMRBox that maintains a depository of relevant software with version history.

3. A uniform data processing pipeline is essential for use and reuse of metabolomics data

The metadata outlined in section ***2: Data and Metadata needed for use and reuse of metabolomics data*** above is essential for the use and reuse of metabolomics data because the field currently lacks defined and globally employed best practices. Instead, a diversity of experimental protocols, statistical methods, and software are used by the metabolomics community - each investigator or research group follows a different if not unique data processing pipeline. This is further exacerbated by the lack of publicly available benchmarking datasets, which makes development, assessment, optimization, and comparison of data analysis steps/pipelines/tools difficult if not impossible. While organizations like MANA and Metabolomics Quality Assurance & Quality Control Consortium (mQACC) strive to establish and inform best practices, it is a daunting task that lacks financial support. Ideally, we would have standard protocol(s) (*likely sample type dependent)* that would convert the raw NMR or MS spectral data set into a list of statistically significant metabolites, concentration changes, and associated pathways. These uniform data processing pipelines could then be employed by the data repository to simplify and automate the use and reuse of the metabolomics data and solve the problems outlined in ***section 1: Experience in generating, using, and reusing metabolomics data sets*** above. Importantly, including these data processing toolsets and pipelines within the data repository would be a valuable step towards establishing standardization. Until this laudable goal is achieved, the metabolomics data in a repository will be an invaluable resource to inform how to establish a uniform data processing pipeline. These metabolomics datasets will establish best practices by identifying what we are doing as a community, what protocols are working well, and what approaches are problematic.

4. Experiences in and recommendations for incorporating metabolomics data into multi-omics studies

- **Experiences:** A multi-omics approach generally includes any combination of genomics, transcriptomics, proteomics, metabolomics, and lipidomics data sets that originate from multiple analytical sources like NMR, LC-MS, GC-MS, and FTIR. Research projects employing multi-omics approaches are becoming more common and are slowly growing in popularity. Multi-omics studies may become a major source of metabolomics data sets in the near future. Because of the complexity of the data set structure, multi-omics data or links to multi-omics data sets are rarely available in metabolomics data repositories.

- **Recommendations:** Data repositories need to accommodate multi-omics data deposits, simplify the deposition of multi-omics data sets, and enable the cross-linking of multi-omics data sets across multiple repositories that accept only a single data type. The UK biobank may provide a valuable example for a repository of multi-model data. The community should require data repositories to share multi-omics datasets or provide links to various components of the datasets if the data is stored in different data repositories. A common quality control material or reference material may be an important tool as a data anchor in merging multi-omics data sets.

5. Experiences in and recommendations for selecting and using software and informatics tools for metabolomics

- **Experiences:** Both commercial and freely available software tools have been developed for processing MS- and NMR-based metabolomics data. Commercial software tools for NMR- and MS-based metabolomics such as

MNova, Chenomx, Progenesis QI (Nonlinear Dynamics), and Compound Discoverer (Thermo Fisher) are quite costly compared to freely available software tools. In addition, the inner workings of these commercial software tools are not transparent making it difficult for users to evaluate the results generated by the software tools. At least one commercial vendor (Bruker) now offers an AI-based identification and quantification model built entirely with proprietary databases and algorithms that has a fee structure and is completely opaque to subscribers. On the other hand, commercial software tools developed by professional software developers may be more stable and user friendly compared to open-source software tools that are typically developed in academic laboratories. Academic software tools developed are not always well-maintained, are outdated, no longer functional or available (*e.g.,* GitHub) due to lack of resources or funding to maintain such software tools. This can result in significant loss of knowledge and tremendous loss of returns on investments to funding agencies.

- **Recommendations:** The development and sharing of well documented, open-source software tools for metabolomics should be strongly encouraged by funding agencies. This will address the need for establishing best practices for the field and a uniform data processing pipeline. In the meantime, more funding is needed to attract and retain talented professional level software developers in academia, considering the higher salary for software developers generally offered in industry compared to salaries offered at universities. The resulting software tools should also be made publicly available and hosted on sites like GitHub and NMRBox.

6. How the interoperability of different file formats has promoted or impeded the reuse of metabolomics data

- **Experiences:** Raw data in mass spectrometry vendor-specific formats makes reusing the metabolomics data difficult, if not impossible. The software *msConvert* addresses this issue to a large extent by converting the vendor-specific formats to open data formats [9]. However, *msConvert* sometimes is not capable of converting the file formats for some raw data files, and certain instrument-specific meta information contained in the vendor-specific data formats is lost in the conversion process. This issue is not as challenging for NMR metabolomics as there exists many tools to export spectra from vendor-specific platforms. But there is still a need to convert the vendor file format to the format used by the processing software, and, of course, each processing software creates a unique processed file format. Again, creating difficulties for data visualization, analysis and reuse/mining repositories.

- **Recommendations:** Encourage mass spectrometry vendors to work with the metabolomics community to use and improve open data formats (mzML, mzXML, cdf, nmrML [10]) to encode vendors' data. Similarly, establishing a uniform file format for the processed spectral data will facilitate the reuse of metabolomics data for analysis.

7. Suggested best practices to enhance harmonization across datasets

- **What can the data repositories do?**
  - **Experiences**
    - Our collective experience suggests that non-uniform organization structures of data archives for different studies make it very challenging to automate data harmonization across a large number of datasets. On the data uploading side, it has also been our experience that uploading data to publicly accessible data repositories can be excessively time consuming, burdensome, and leads to unnecessary delays in final acceptance of manuscripts in journals.
    - Database errors are routinely caused by the human depositor, where the complexity of the deposition process is a contributing factor.
    - Problems with proliferations of redundant databases and depositories
    - For studies that combine metabolomics, lipidomics, proteomics, and/or genomics, different types of omics data are stored in different data repositories, or some components of the data are not publicly available.

  - **Recommendations**
    - Enable and encourage a uniform or standard structure of data archives uploaded into the data repositories to allow automated processing of data from different studies.
    - Define and encourage the use of common ontologies specific for metabolomics. This is foundational to the development of tools to improve the link between experiment and data deposition, ultimately aiding compliance and facilitating data reuse.

    - Provide software tools with a graphical user interface to guide and assist data uploaders to prepare uniformly formatted data archives with minimum efforts. Include example template files or pulldown menus for users to format their data according to the templates, or items in menus.
    - Harmonize data format between MetaboLights and Metabolomics Workbench so that data can be easily deposited in both or switched from one database to the other, if necessary.
    - For existing data in data repositories that are missing essential information and may be deemed not reusable, encourage data generators to re-upload their data to comply with the new uniform organization structure and provide all necessary information.
    - Need a uniform approach for metabolite annotations. InChi and SMILES still have accuracy limitations and are problematic for non-experts. An agreed-upon unique database identifier (HMDB, BioCyc, KEGG, *etc.*) may be a better choice.
    - Use existing data repositories, like wwPDB, as a model system to guide the metabolomics data repository and perform a feasibility study.
    - Providing metadata can be the most time-consuming and frustrating aspect of the data deposition process. Making this process simpler and easier would be highly beneficial. For example, enabling the copying or linking of common group-specific protocols across multiple depositions or developing text-parsing capabilities to populate forms from methodology documents or papers would be helpful.

- **What can funding agencies and journal publishers do?**

  Experiences: While the literature is currently populated with numerous replicate metabolomics studies (*e.g.,* identifying metabolomics biomarkers to diagnose a variety of human diseases), few if any of these studies that deal with the same sample sources, species, and disease have been deposited in a database and are reusable. For cases where there are sufficient replicate studies with similar sources, the studies themselves are difficult to co-analyze due to the lack of a common link or ready access to raw or processed data, or the necessary metadata. There is a significant lack of sample specific reference materials, which limits what can be truly achieved with data harmonization. We will continue to struggle with the use and reuse of metabolomics data without tangible incentives.

  Recommendations:
  - Establish policies regarding usage of deposited datasets, including considering uploaded datasets as "copyright" materials of the original creator. The original creator should be acknowledged or invited as co-authors if the reused data are in publication forms. The re-users will need to agree with these policies before accessing the datasets. Similarly, the depositor(s) will need to agree to policies that enable fair usage of their data and analysis results.
  - Ask all journals to require deposition of metabolomics data in public repositories prior to manuscript publications.
  - Provide incentives for data generators, who received federal funding for their study, to deposit their data in publicly accessible repositories, even if these data are no longer used by the group that generated the data.
    - A lot of unpublished metabolomics data that could be very informative remains inaccessible to the general public long after the studies were completed. Providing general access to these datasets will increase both the diversity of studies and the number of similar studies in publicly accessible data repositories for data harmonization and biomarker discovery.
  - Ask all journals to require acknowledgements of the original data when re-analysis of the original data is reported in publications.
  - Establish guidance for standardized data sharing when receiving federal or other funding, using similar guidance used by the NIH for genomic data sharing (https://sharing.nih.gov/genomic-data-sharing-policy/developing-genomic-data-sharing-plans).
  - Encourage the development and use of reference materials. Human plasma and urine have become the first sample types for standardized reference materials; however, cell lines, tissues, microbiomes, and other biofluids are critical for inter-study, across sample types, and the historical analysis of publicly available data, as well as allowing for the characterization of matrix specific metabolome.

8. Other recommendations

- A primary goal should be accuracy, transparency, flexibility, and then finally user-friendly.
- There should be both browsing and search options and ranked display of results.
- We encourage developers of data repositories to seek input from web designers as well as the scientific community.

References

1. https://www.metabolomicsworkbench.org/
2. https://www.ebi.ac.uk/metabolights/
3. https://massive.ucsd.edu/
4. Smirnov A, Liao Y, Fahy E, Subramaniam S, Du X, ADAP-KDB: A Spectral Knowledgebase for Tracking and Prioritizing Unknown GC-MS Spectra in the NIH's Metabolomics Data Repository**.** *Anal Chem* **2021**, 93(36):12213-12220.
5. Fiehn, O., Robertson, D., Griffin, J., van der Werf , M., Nikolau, B., Morrison, N., Sumner, L.W., Goodacre, R., Hardy, N.W., Taylor, C., Fostel, J., Kristal, B., Kaddurah-Daouk, R., Mendes, P., van Ommen, B., Lindon, J.C., Sansone, S.-A., The Metabolomics Standards Initiative (MSI), *Metabolomics*, **2007**, 3(3) 175-178.
6. Sumner, L.W., Amberg, A., Barret, D., Beale, M., Beger, R., Daykin, C., Fan, T., Fiehn, O., Goodacre, R., Griffin, J., Hardy, N., Higashi, R.M., Lane, A., Lindon, J., Marriott, P., Nicholls, Reily, M., Viant, M., Proposed minimum reporting standards for chemical analysis, *Metabolomics*, **2007**, 3(3) 211-221.
7. Sansone, S.A., Nikolau, B., van Ommen, B., Kristal, B.S., Taylor, C., Robertson, D., Lindon, J., Griffin, J.L., Sumner, L.W., van der Werf, M., Hardy, N.W., Morrison, N., Mendes, P., Kaddurak-Daouk, R., Goodacre, R., Fan, T., Fiehn, O. The Metabolomics Standards Initiative, *Nature Biotechnology*, **2007**, 25(7) 846-848.
8. Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., Ewald, J., Fraser, P., Giavalisco, P., Hall, R., Heinemann, M., Link, H., Luo, J., Neumann, S., Nielsen, J., Perez de Souza, L., Saito, K., Sauer, U., Schroeder, F., Schuster, S., Siuzdak, G., Skirycz, A., Sumner, L., Snyder, M., Tang, H., Tohge, T., Wang, Y., Wen, W., Wu, S., Xu, G., Zamboni, N., Fernie, A., Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nature Methods*, **2021**, 18, 747-756.
9. https://proteowizard.sourceforge.io/
10. https://nmrml.org/

9. MANA NMR Interest Group Members

- Leo L. Cheng, Massachusetts General Hospital, Harvard Medical School
- Chaevien S. Clendinen, Environmental Molecular sciences Laboratory (EMSL), Pacific Northwest National Laboratory (PNNL)
- Valérie Copié, Department of Chemistry and Biochemistry, Montana State University-Bozeman
- John R. Cort, Biological Sciences Directorate, Pacific Northwest National Laboratory, and Institute of Biological Chemistry, Washington State University
- Xiuxia Du, Department of Bioinformatics and Genomics, University of North Carolina at Charlotte
- Art Edison, Department of Biochemistry, University of Georgia
- Hamid R. Eghbalnia, UConn Health
- Candace Fleischer, Department of Radiology and Imaging Sciences, Emory University School of Medicine
- Goncalo J. Gouveia, Institute for Bioscience and Biotechnology, National Institute of Standards and Technology and University of Maryland
- Nathaniel Mercaldo, Department of Radiology, Massachusetts General Hospital and Harvard Medical School
- Wimal Pathmasiri, Nutrition Research Institute, Department of Nutrition, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
- Robert Powers, Department of Chemistry, University of Nebraska-Lincoln
- Daniel Raftery, Department of Anesthesia and Pain Medicine, University of Washington.
- Tracey Schock, Chemical Sciences Division, National Institute of Standards and Technology
- Jane Shearer. Department of Biochemistry and Molecular Biology. University of Calgary
- Lloyd W. Sumner, Department of Biochemistry, Director, Missouri University Metabolomics Center, University of Missouri
- Panteleimon G. Takis, National Phenome Center, Imperial College London