Supplementary Information

mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics

Nils Hoffmann[1], Joel Rein[2], Timo Sachsenberg[3], Jürgen Hartler[4,5], Kenneth Haug[6], Gerhard Mayer[7], Oliver Alka[3], Saravanan Dayalan[8], Jake TM Pearce[9], Philippe Rocca-Serra[10], Da Qi[11,12], Martin Eisenacher[7], Yasset Perez-Riverol[6], Juan Antonio Vizcaíno[6], Reza M Salek[13,*], Steffen Neumann[14,15,*], Andrew R Jones[12,*]

Abstract

Additional details are provided here on the relationship between mzTab-M and mzTab 1.0, on how lipid identifications can be reported in mzTab-M, and on the representation of complex experimental designs in mzTab-M.

[1] Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Otto-Hahn-Straße 6b, 44227 Dortmund, Germany
[2] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA Cambridge, United Kingdom
[3] Applied Bioinformatics Group, Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany
[4] Institute of Computational Biotechnology, Graz University of Technology, Petersgasse 14, 8010 Graz, Austria
[5] Center for Explorative Lipidomics, BioTechMed-Graz, Graz, Austria
[6] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.
[7] Medizinisches Proteom Center (MPC), Ruhr-Universität Bochum, Universitätsstraße 150, D-44801 Bochum, Germany
[8] Metabolomics Australia, The University of Melbourne, Parkville, Australia
[9] MRC-NIHR National Phenome Centre, Department of Surgery & Cancer, Imperial College London, London, SW7 2AZ, United Kingdom
[10] University of Oxford, e-Research Centre, 7 Keble Road, Oxford OX1 3QG, United Kingdom
[11] BGI-Shenzhen, Shenzhen, 518083, China
[12] Institute of Integrative Biology, University of Liverpool, United Kingdom
[13] International Agency for Research on Cancer, 150 cours Albert Thomas, 69008 Lyon, France
[14] Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Halle, Germany
[15] German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig Deutscher, Platz 5e, 04103 Leipzig, Germany

* = Joint corresponding authors

Relationship with mzTab 1.0

The mzTab 1.0 format is able to capture summary details about quantification of small molecules and lipids by MS in a single SML table, but cannot capture richer evidence about identification, and does not fully adequately handle ambiguity in identification. As such, mzTab 1.0 is now deprecated for use in metabolomics and lipidomics and all developers are strongly encouraged to implement version 2.0 of the standard. At present, mzTab 1.0 remains in use for proteomics, although the PSI is starting to work to update mzTab to a new version, following some of the changes made to the core of the format (e.g. around experimental design) suffixed to become mzTab-P. The working groups developing mzTab-M and mzTab-P share several members in common to ensure that the formats remain sufficiently well aligned so that laboratories performing both MS proteomics and metabolomics should be able to adapt in house software straightforwardly to cover both cases.

Reporting of identifications of lipids and other compound classes

Since lipids can rarely be fully structurally annotated by MS fragmentation spectra, a standard nomenclature was introduced in 2013 to provide information of how much structural information is obtainable by the performed MS experiments[1]. This nomenclature differentiates between identifications where only the head-group, total lengths and number of double bonds of the attached fatty acyl and/or alkyl/1-alkenyl (FA) chains are known (lipid species), or where more structural information is available (lipid molecular species). These structural ambiguities hamper immediate fragment-based identification and are not limited to lipidomics, but are also inherent in related fields like natural product and environmental chemistry, and glycomics.

The mzTab-M standard allows the reporting of identification and quantification information for highly regularly structured compound classes like lipids, whose MS^1 or MS^2 fragments can be identified on different levels of structural resolution. For lipids, a domain-specific standard nomenclature[1] can capture structural knowledge at the lipid species (lipid class, subclass and bond type) or molecular lipid species levels (fatty acyl composition, position, and fatty acid/sphingoid base structure, full structural elucidation) and, separately, for their fragment ions[2]. This information can be reported in the *chemical_name* columns in the summary and evidence sections of mzTab-M. Other, more specific information specific to the identification or the quantification workflow[3] should be reported in the metadata section or as optional columns in the SMF and SME tables, respectively. The ChEBI resource provides a curated database of compound classes beyond lipids. Additionally, software such as the ClassyFire system[4] uses a rule-based approach to calculate hierarchical compound classes for any given molecular structure, which can also be reported as optional columns in the evidence tables.

Current limitations in MS technology however hamper the full structural elucidation of such molecules with a single analytical approach, which is why specific lipid classes are characterized with combinations of different MS platforms, chromatographic separation, and different ionization modes. In practice, some lipids or actually their fragments are better characterizable in positive ionization mode, while others are better detected in negative mode. It is therefore common to report both positive and negative mode evidence for lipids together, which is also supported in mzTab-M. We provide an example derived from the Lipid Data Analyzer 2 software[5] in the GitHub repository, which uses a rule-based identification approach.

In the near future, it is expected that data standards task group from the recently inaugurated Lipidomics Standards Initiative (https://lipidomics-standards-initiative.org/) will publish extended guidelines and reporting recommendations, together with an extended semantic validation for

lipidomics, which may also serve as a blueprint for customizations of the format in other domains[6]. Additional guidelines and nomenclature will be encoded via adding CV terms and rules governing their usage rather than changing the core of the format, allowing lipidomics development teams to start implementations with mzTab-M immediately.

Experimental Design in mzTab-M

The main text of the manuscript contains a simple experimental design example in Figure 2. A more complex example is provided in Supplementary Figure 1, including a time course design, biological and technical replicates.

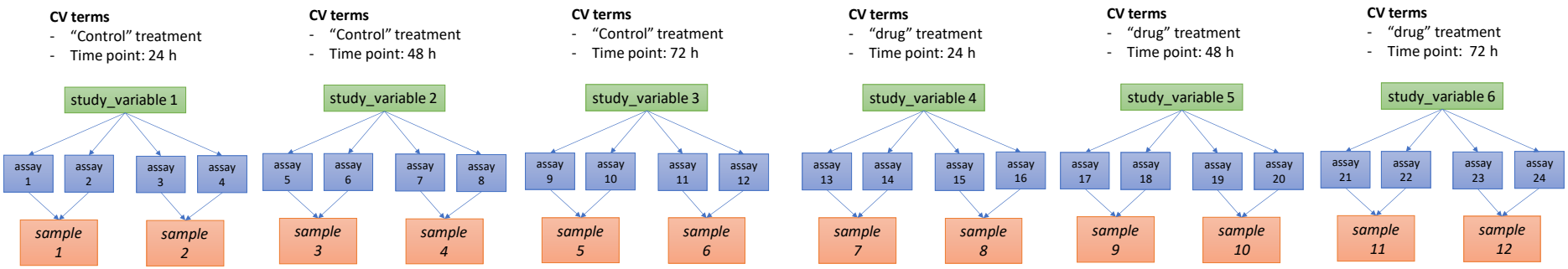

Figure S1. An experimental design represented in mzTab-M in which there are two treatments: "control" vs "drug" measured on a time course at 24, 48 and 72 hours. There are two technical replicates and two biological replicates: 2 X 3 X 2 X 2 = 24 assays. It is assumed that no-prefractionation of samples took places there would also be 24 ms_run elements (not shown, giving the locations of the raw MS data files). The SML and SMF tables would have quantitative values reported for 24 assays and 6 study_variables.

References

(1) Liebisch, G.; Vizcaino, J. A.; Kofeler, H.; Trotzmuller, M.; Griffiths, W. J.; Schmitz, G.; Spener, F.; Wakelam, M. J. *Journal of lipid research* **2013**, *54*, 1523-1530.
(2) Pauling, J. K.; Hermansson, M.; Hartler, J.; Christiansen, K.; Gallego, S. F.; Peng, B.; Ahrends, R.; Ejsing, C. S. *PLoS One* **2017**, *12*, e0188394.
(3) Holcapek, M.; Liebisch, G.; Ekroos, K. *Anal Chem* **2018**, *90*, 4249-4257.
(4) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. *Journal of Cheminformatics* **2016**, *8*, 61.
(5) Hartler, J.; Triebl, A.; Ziegl, A.; Trotzmuller, M.; Rechberger, G. N.; Zeleznik, O. A.; Zierler, K. A.; Torta, F.; Cazenave-Gassiot, A.; Wenk, M. R.; Fauland, A.; Wheelock, C. E.; Armando, A. M.; Quehenberger, O.; Zhang, Q.; Wakelam, M. J. O.; Haemmerle, G.; Spener, F.; Kofeler, H. C.; Thallinger, G. G. *Nature methods* **2017**, *14*, 1171-1174.
(6) Liebisch, G.; Ekroos, K.; Hermansson, M.; Ejsing, C. S. *Biochim Biophys Acta* **2017**, *1862*, 747-751.