



Published in final edited form as:

Metabolomics. 2007 September ; 3(3): 211–221. doi:10.1007/s11306-007-0082-2.

Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)

Lloyd W. Sumner,

The Samuel Roberts Noble Foundation, Ardmore, OK, USA

Alexander Amberg,

Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany

Dave Barrett,

Centre for Analytical Bioscience, School of Pharmacy, University of Nottingham, Nottingham, UK

Michael H. Beale,

National Centre for Plant and Microbial Metabolomics, Rothamsted Research, West Common, Harpenden, Herts, UK

Richard Beger,

National Center for Toxicological Research, Jefferson, AR, USA

Clare A. Daykin,

Division of Molecular and Cellular Science, School of Pharmacy, University of Nottingham, Nottingham, UK

Teresa W.-M. Fan,

Department of Chemistry, University of Louisville, Louisville, KY, USA

Oliver Fiehn,

UC Davis Genome Center, University of California, Davis, CA, USA

Royston Goodacre,

School of Chemistry and Manchester Interdisciplinary Biocentre, The University of Manchester, Manchester, UK

Julian L. Griffin,

The Department of Biochemistry, University of Cambridge, Cambridge, UK

Thomas Hankemeier,

Division Analytical Biosciences, Leiden University, Leiden, The Netherlands

Nigel Hardy,

Department of Computer Science, University of Wales, Aberystwyth, Aberystwyth, UK

James Harnly,

© Springer Science+Business Media, LLC 2007

Correspondence to: Lloyd W. Sumner, lwsunner@noble.org.

The contents of this paper do not necessarily reflect any position of the Government or the opinion of the Food and Drug Administration

Sponsor: Metabolomics Society

<http://www.metabolomicssociety.org/>

Reference: <http://msi-workgroups.sourceforge.net/bio-metadata/reporting/pcb/>

<http://msi-workgroups.sourceforge.net/chemical-analysis/>

Food Composition and Methods Laboratory, Beltsville Human Nutrition Research Center, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, MD, USA

Richard Higashi,

Department of Chemistry, University of Louisville, Louisville, KY, USA

Joachim Kopka,

Max Planck Institute of Molecular Plant Physiology, Golm, Germany

Andrew N. Lane,

James Graham Brown Cancer Center, University of Louisville, Louisville, KY, USA

John C. Lindon,

Department of Biomolecular Medicine, Imperial College London, London, UK

Philip Marriott,

School of Applied Sciences, RMIT University, Melbourne, Australia

Andrew W. Nicholls,

Investigative Preclinical Toxicology, GlaxoSmithKline, Ware, UK

Michael D. Reily,

Discovery Biomarkers, Pfizer Global R&D, Ann Arbor, MI, USA

John J. Thaden, and

College of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Mark R. Viant

School of Biosciences, The University of Birmingham, Birmingham, UK

Lloyd W. Sumner: lwsunmer@noble.org; Alexander Amberg: Alexander.Amberg@sanofi-aventis.com; Dave Barrett: David.Barrett@nottingham.ac.uk; Michael H. Beale: mike.beale@bbsrc.ac.uk; Richard Beger: richard.beger@fda.hhs.gov; Clare A. Daykin: Clare.Daykin@nottingham.ac.uk; Teresa W.-M. Fan: teresa.fan@louisville.edu; Oliver Fiehn: ofiehn@ucdavis.edu; Royston Goodacre: Roy.Goodacre@manchester.ac.uk; Julian L. Griffin: jlg40@mole.bio.cam.ac.uk; Thomas Hankemeier: hankemeier@chem.leidenuniv.nl; Nigel Hardy: nwh@aber.ac.uk; James Harnly: james.harnly@ars.usda.gov; Richard Higashi: rick.higashi@louisville.edu; Joachim Kopka: Kopka@mpimp-golm.mpg.de; Andrew N. Lane: anlane01@gwise.louisville.edu; John C. Lindon: j.lindon@imperial.ac.uk; Philip Marriott: philip.marriott@rmit.edu.au; Andrew W. Nicholls: andrew.w.nicholls@gsk.com; Michael D. Reily: Michael.Reily@pfizer.com; Mark R. Viant: m.viant@bham.ac.uk

Abstract

There is a general consensus that supports the need for standardized reporting of metadata or information describing large-scale metabolomics and other functional genomics data sets. Reporting of standard metadata provides a biological and empirical context for the data, facilitates experimental replication, and enables the re-interrogation and comparison of data by others. Accordingly, the Metabolomics Standards Initiative is building a general consensus concerning the minimum reporting standards for metabolomics experiments of which the Chemical Analysis Working Group (CAWG) is a member of this community effort. This article proposes the minimum reporting standards related to the chemical analysis aspects of metabolomics experiments including: sample preparation, experimental analysis, quality control, metabolite identification, and data pre-processing. These minimum standards currently focus mostly upon mass spectrometry and nuclear magnetic resonance spectroscopy due to the popularity of these techniques in metabolomics. However, additional input concerning other techniques is welcomed and can be provided via the CAWG on-line discussion forum at <http://msi-workgroups.sourceforge.net/> or [http://Msi-workgroups-feedback@lists.sourceforge.net](mailto:Msi-workgroups-feedback@lists.sourceforge.net). Further, community input related to this document can also be provided via this electronic forum.

Keywords

Metabolomics; Metabolite profiling; Metabolite identification; Minimum reporting standards; Chemical analysis; Mass spectrometry; Nuclear magnetic resonance; Flux; Isotopomer analysis; GC-MS; LC-MS; CE-MS; NMR; Quality control; Method validation

1 Introduction

The aim of the Chemical Analysis Working Group (CAWG) as part of the Metabolomics Standards Initiative (MSI) is to identify, develop and disseminate a consensus description for the best chemical analysis practices related to all aspects of metabolomics. Ideally, the proposed standards will consist of good analytical chemistry practices while providing specific provisions for metabolomic data (the main distinction being large numbers of data-sets each containing large numbers of measurements, and the need to compare them electronically and across different instrumental platforms). These practices will be aligned with those typically mandated by top quality analytical journals. The goal is not to *prescribe* how metabolomics experiments should be performed, but to formulate a minimum set of reporting standards that *describe* the experimental methods (i.e. the metadata or information describing the nature of the experiments and how they were actually executed) to maximize the utility of the data to other researchers. Consequently, there will be no attempt to restrict or dictate specific practices, but to develop consistent and appropriate descriptors to support the dissemination and re-use of metabolomic data. Such reporting standards will specify the metadata identified as necessary for complete and comprehensive reporting in a range of contexts, such as submission to academic journals and public databases. Data exchange standards will be developed to provide a transparent technical vehicle which meets or exceeds the requirements of reporting standards.

The scope of the CAWG includes sample preparation, experimental analysis, instrumental performance, method validation, metabolite identification, and data preprocessing. There is slight overlap in the sample preparation with the Biological Context Working Group and slight overlap in data pre-processing with the Data Processing Working Group. However, the scope and focus of the CAWG is upon the experimental aspects of sample processing, instrumental analysis, and commonly used data pre-processing methods which convert raw instrumental files into organized, tabulated file formats. The organized data are then used for further statistical and chemometric analysis which are the focus of the Data Processing Working Group.

The operational plan of the CAWG is to cooperatively draft a consensus document that describes a minimum core set of necessary metadata related to the chemical analyses associated with metabolomics experiments. This will be based upon community input from generalists and specialists relating to the most common technologies utilized in metabolomics. The CAWG will evaluate previous and relevant work in other specialist areas including similar work in transcriptomics and proteomics studies, and recent metabolomics standardization efforts. The group will pay careful attention to the distinction of best practice (which will evolve as the science and technology of metabolomics advances), reporting standards (which should have longer validity) and data exchange standards (which support reporting). It will work with relevant journals and editorial staff to review and advise on the practicality, acceptability, and support of standards.

The proposed CAWG standards were originally described during the NIH Metabolomics Workshop convened in August, 2005 (<http://www.niddk.nih.gov/fund/other/metabolomics2005/>) and are based upon significant literature (Bino 2004; Jenkins et al.

2004; Quackenbush 2004; Jenkins et al. 2005; Lindon et al. 2005; Fiehn et al. 2006, Rubtsov et al. 2007). Significant input has been provided related to mass spectrometry (MS) and nuclear magnetic resonance (NMR) based metabolomics, but the ultimate schema is aimed at all analytical approaches used in metabolomics. Input to date has been provided by a diversity of academic and commercial entities through personal communications and through the on-line discussion forum (<http://msi-workgroups.sourceforge.net/>).

2 Proposed minimum information for reporting chemical analysis

The following sections describe the proposed minimum information for reporting chemical analyses metadata that have been discussed to date. The proposed minimum reporting standard information is presented below as bulleted text which is augmented with numerous examples. The examples should not be viewed as required and are not meant to include an exhaustive list of all possibilities. However, the examples should help the reader better visualize the requested context of the proposed minimum information.

2.1 Proposed minimum metadata for sample preparation

Sample preparation is a vast topic which can vary dramatically for different species, tissues, cell cultures, and biofluids. However, it is fundamentally essential that sufficient information is provided about sample preparation to enable experimental reproduction as well as to provide convincing evidence of sample integrity. The initial stages of sample preparation are often generic, whereas the final stages are almost always technique-specific. Therefore, proposed minimum standards for generic sample preparation are provided here, whereas instrument specific sample preparation details are provided within the respective instrumental sections. Further, the issue of sample collection and processing is being addressed by multiple MSI working groups and thus, there is some overlap on this theme (Fiehn et al. 2007; Griffin et al. 2007; van der Werf et al. 2007). However, greater emphasis is provided here concerning the experimental aspects of the sample processing.

- Sampling process and protocol
 - Replicate sampling and analyses: Substantial biological variance exists within all organisms; therefore replicate sampling and analyses are critical to provide a statistical basis for data evaluation and interpretation. A minimum of triplicate ($n = 3$) biological sampling is proposed with $n = 5$ preferred. Biological replicates (repetitive analyses of samples obtained from different individuals or pooled individuals from a population) are preferred over analytical replicates (repetitive analyses of the same sample obtained from the same individual or pooled individuals) as biological variance almost always exceeds analytical variance.
 - Tissue harvesting method: For example, sample freezing method (e.g. liquid N₂, dry ice and acetone bath, freeze clamping, etc.), sample wash method for removing unwanted external components, time and duration for tissue collection (e.g. time from tissue resection to liquid N₂ freezing), temperature, and sample storage prior to further preparation (e.g. -80°C for 2 weeks). All temperatures should be measured if possible; however temperature set-points are acceptable assuming quality monitoring was performed and no abnormalities recorded.
 - Biofluid harvesting or collection method: For example, syringe, collection onto refrigerated surface, vacuum system/vacutainers used for blood collection, storage vessel and anticoagulant (if relevant), temperature, velocity and duration of centrifugation, and sample freezing method.

- Tissue processing method: For example, lyophilization, fresh tissue processing, pulverization/ homogenization, tissue cell lysis (e.g. liquid N₂ grinding, manual or electric homogenization, bead-based homogenization, ultrasonic cell lysis, buffer based lysis, etc.).
- Storage conditions prior to extraction or further processing (e.g. –80 °C, duration, atmospheric pressure or vacuum, desiccation, preservatives added).
- Relocation and shipping of tissues from one laboratory to another (if relevant).

Generic extraction and subsequent sample handling that are typically employed for most samples (instrument specific sample processing methods are provided in the respective sections, below).

- Extraction method
 - Solvent(s), pH and ionic strength of buffer, solvent temperature and volume(s) per quantity of tissue, number of replicate extracts, sequential extraction, and extraction time.
 - Example: 1 ml ice-cold methanol (MeOH) per 6 mg lyophilized tissue, two extractions combined, CHCl₃/MeOH (2/1, v/v) followed by 10% trichloroacetic acid extraction.
 - It is noted that degassing of solvents is important to minimize redox reactions of sensitive compounds such as ascorbate, cysteine, etc.
- Extract concentration, dilution, and resolubilization processes
 - Dried under nitrogen, resolubilized in H₂O or pyridine.
- Extract Enrichment (if relevant)
 - SPE (solid phase extraction column volume/mass, elutant, sorbent, manufacturer)
 - Desalting, molecular weight cut-off, ion exchange, etc.
- Extract Cleanup and/or Additional Manipulation
 - Ultrafiltration, removal of paramagnetic ions, addition of metal chelators such as EDTA, citrate
- Extract Storage and/or Relocation
 - Storage conditions prior to and during analysis
 - Relocation and shipping of extracts from one laboratory to another (if relevant)

2.2 Proposed minimum metadata relative to chromatography

The majority of mass spectrometry based metabolomics methods include sample introduction via hyphenated chromatography. This is also a feature of some NMR experiments (i.e. LC/NMR) as well as other analytical devices, e.g. photodiode arrays, Coulombic arrays, etc. Thus, it is critical to define the chromatographic parameters and the following metadata are suggested.

- Chromatography instrument description

- Manufacturer, model number, software package and version number or date.
- Auto-injector
 - Injector model/type, software version, injection volume, wash cycles (volumes), solvent.
- Separation column and pre/guard column
 - Manufacturer, model number/name, stationary media composition (support and coating, e.g. silica, C18, etc.) and physical parameters (i.e. coating thickness for GC/MS, particle size and pore size for LC/MS), internal diameter, and length.
- Technique-specific sample preparation
 - Resuspension of sample (e.g. in mobile phase), amount injected.
 - Derivatization reaction conditions if relevant, (e.g. OMS/trimethylsilyl; chemical manufacturer, temperatures, and duration).
 - Sample spiking e.g. internal standards, retention-index standards.
- Separation parameters
 - Method name (a detailed method can be published elsewhere and referenced here by a unique protocol identifier), injector temperature, split or splitless mode and ratio, LC post-column split, mobile phase compositions, mobile phase flow rates, pressure, thermal/solvent/solute gradient profiles.

2.3 Proposed minimum metadata relative to mass spectrometry

Mass spectrometry is a popular but complex technique used in metabolomics. Thus, it is necessary sufficient details to enable experimental replication and the following minimum reporting standards are proposed for mass spectrometry.

- Instrument description
 - Manufacturer, model number, software package and version (name, number or date).
- Sample introduction and delivery
 - From GC, from LC, direct infusion without chromatography, direct infusion using dedicated autosampler flow rate.
- Ionization source
 - Ionization mode (EI, APCI, ESI etc.), polarity (positive or negative-ion analysis), vacuum pressure, skimmer/focusing lens voltages (e.g. capillary voltage etc.), gas flows (e.g. nebulization gas, cone gas etc., source temperature). Although these values will vary between instruments, they should provide a cumulative view of the ionization conditions sufficient to enable reproduction of the experiment.
- Mass analyzer description and acquisition mode
 - Type (quadrupole, ion-trap, time-of-flight, FT-ICR, including combinations of these for hybrid instruments), acquisition mode (full scan, MSⁿ, SIM, MRM, etc.).

- Technique-specific sample preparation (if relevant)
 - Re-suspension of sample (e.g. in MeOH:water 1:1 with 0.2% formic acid), derivatization, volume injected, and internal calibrant(s) added (if relevant).
- Data acquisition parameters
 - Date, operator, data acquisition rate, m/z scan range, compounds used for m/z calibration, mass resolution, mass accuracy, logic program used for data acquisition (often reported for ion-traps), spectral acquisition rate, vacuum pressure, and/or lock spray (concentration, lock mass, flow rate, and frequency).

2.4 Proposed minimum metadata relative to nuclear magnetic resonance

NMR is a popular, but complex technique used in metabolomics. Thus, it is necessary sufficient details to enable experimental replication and the following minimum reporting standards are proposed for mass spectrometry.

- Instrument description
 - Manufacturer, model name/number, magnetic field strength in Tesla (example 14.1 T Varian Inova; 18.8 T Bruker Avance) or proton resonance frequency e.g. 600 MHz, and console description.
- Instrument configuration
 - VT control, pulsed field gradients (z or x,y,z) and maximum gradient strength (if used), number of shims, number of channels.
 - Probe type (e.g. 10 mm ^{31}P , 5 mm HCN cold probe, 3 mm flow-probe etc.), solution or solid-state, automation or manual operation, autotune or manual tune, and probe gas. For LC-NMR: sample handler, injection volumes, wash cycles and solvent.
- Instrument-specific sample preparation
 - Volume, extract/powder/intact organisms, tissue or cells, type of NMR tube (e.g. conventional, Shigemi, microcell etc.), pH, solvent (D_2O , CD_3OD , CDCl_3 etc.), buffer, chemical shift or calibration standard.
- Data acquisition parameters
 - For 1-D ^1H or X-nucleus NMR: temperature, observed nucleus, pulse sequence name, pulse sequence implementation (e.g. gradient selection, sensitivity enhancement), spin rate or statement of no spin, solvent saturation or decoupling method, presence or absence of heteronuclear decoupling (e.g. isotope-enriched samples), decoupling mode and bandwidth; spin lock field strength (in Hz) and duration (in sec), mixing time (for NOESY, ROESY etc.), spin echo time (e.g. for relaxation analysis or broadline suppression), RF pulse widths, any selective pulse shapes and durations used, magnetic field gradient pulse times and shapes, spectral width, acquisition time, relaxation delay and additional delays (mixing time, etc.), interpulse delay (or recycle time), digitization parameters, spectral width and acquisition time, number of transients, and number of steady states transients (i.e. dummy scans). For solvent suppression: technique, excitation maximum and bandwidth.

- Additional parameters for 2-D and higher dimensional NMR: observed nucleus in F2 and F1, pulse sequence, excitation pulse widths for relevant nuclei, spectral width in F2 and F1, solvent saturation method, number of transients in t_2 and number of increments in t_1 , acquisition times for t_2 and t_1 , phase sensitive or magnitude detection, pulsed field gradient strengths and shapes (z or x,y,z) and maximum gradient strength (if relevant to the pulse sequence).
- Additional parameters for X-nucleus 1D and higher dimensional NMR: direct or indirect detection, proton decoupling mode (Waltz, Garp, Wurst, Stud etc.) and effective band width, evolution time for constant time experiments, editing mode (cf. INEPT-based experiments), heteronuclear spin lock strength and mixing time (e.g. HCCH-TOCSY).
- Additional parameters for pseudo 2D NMR experiments: physical parameter varied in the t_1 dimension (e.g. T_2 , T_1 , diffusion period, chromatographic separation time as in LC-NMR, etc.), pulse sequence, array of values used for physical constants.

2.5 Proposed minimum metadata relative to stable isotopes & flux analysis

Many researchers utilize stable isotopes and flux analysis in metabolomics research to better understand mass flow through pathways. Therefore, the following minimum reporting standards are proposed for stable isotopes and flux analysis.

- Isotope labeled precursors used
 - Element/isotope, position(s), percent labeled; e.g. [^{13}C -1]-D-glucose (98%), [$^{15}\text{N}_2$]-L-glutamine (99%).
 - Isotope source (i.e. manufacturer), chemical purity of the labeled compound(s), concentration of the compound, fraction of total present (requires detailed breakdown of media composition for cell and tissue studies, including analysis of any added FCS or other growth supplements; labeling scheme).
 - Total number of moles isotope added during the experiment.
- Duration of pulse label or continuous addition

2.6 Proposed minimum metadata relative to Fourier transform infrared (FT-IR) spectroscopy

FT-IR spectroscopy has been used for metabolic fingerprinting and footprinting (Ellis and Goodacre 2006). In this approach the classification of samples is based on provenance of either their biological relevance or origin and does not usually give specific metabolite information. The following minimum reporting standards are proposed for FT-IR spectroscopy.

- FT-IR spectrometer instrument description
 - Manufacturer, model number, software name and version number or date.
- Instrument configuration
 - Type of sampling compartment used, including where necessary type of microscope employed.
 - Type of detector used (DTGS (deuterated triglycine sulphate), MCT (mercury cadmium telluride), and/ or FPA (focal plane array)).

- Technique-specific sample preparation
 - Resuspension of n mg ml⁻¹ sample into solvent, volume analysed.
- Sample presentation
 - Transmission measurement: in KBr, or on ZnSe, Si windows.
 - Reflectance measurement: on Si, Au, Al, or other defined metal sample carrier.
 - Diffuse reflectance measurement: on defined metal sample carrier.
 - Sampling area, and for imaging pixel size.
- Data acquisition parameters
 - Wavenumber (cm⁻¹) range.
 - Rate of acquisition.
 - Spectral resolution (in cm⁻¹).
 - Number of spectra co-added.
 - Number of data points in the resultant spectrum, and how this is displayed (absorbance or transmission).

2.7 Proposed minimum metadata relative to instrumental performance and method validation

Instrumental performance validation/qualification and method validation help ensure reliable data production and to demonstrate that a particular method used for quantitative measurement of an analyte(s) in a given biological matrix, such as plants, blood, plasma, serum, or urine, is reliable and reproducible for the intended use (Thompson et al. 2002; FDA 2001). These quality control procedures are fundamental components of Good Laboratory Practices (GLP), Good Analytical Practices (GAP), and Good Manufacturing Practices (GMP). Although instrumental performance and method validation are not mandated, they are recommended and the following descriptions are suggested.

- Minimum Reporting of Instrumental Performance Parameters is Encouraged. The nature and method(s) used to ensure sensitive and selective instrumental performance should be reported and the following details and descriptors are deemed appropriate.
 - Mass spectrometry instrument performance validation parameters reported might include chemical description of the m/z calibration standard used, accuracy of m/z calibration, mass resolution, and ion source optimization parameters. For hyphenated MS methods, suggested reporting parameters could include chromatographic resolution, accuracy and precision of internal standard(s) or retention time markers, accuracy and precision for replicated analyses, accuracy and precision for validation sample(s), and cycles per column/injector/septum/ blank.
 - NMR instrument performance verification parameters might include calibration standard used (name, chemical shift and concentration; e.g. 0.5 mM DSS or 1 mM TMS at 0.0 ppm), statement of line width of the standard at 50% and 1% of its full height (e.g. DSS, TSP or TMS methyl peak) or residual water, pH marker used (if relevant) and shift correction. For X nuclei: external or internal reference and conditions, and correction made for susceptibility effects. Reporting of shift referencing method for

indirect dimension in 2D experiments (direct or indirect based on ratios) would also be beneficial.

- Quantitative Method Validation. Two methods of quantitative analysis are typically used in metabolomics and include relative and absolute quantification. Relative quantification (i.e. reporting of metabolite(s) instrument response relative to an internal standard or another metabolite(s) level such as the sum of all metabolite abundance) is typically used in non biased metabolomics. Whereas, absolute quantification (determination of the absolute concentration of a metabolite(s) through correlation of its instrument response to that of a known concentration series of the same metabolite) is commonly used in targeted metabolite(s) analysis.
 - Relative Quantification reporting should include
 - a description and quantifier of the added exogenous isotopically labeled or unlabeled metabolite(s).
 - A description of the method used for assessing instrument response (e.g. peak integration, binning/bucketing or deconvolution method, intensity normalized to reference,
 - For NMR, descriptions for correction for saturation effects - T_1 values measured), and provide relaxation agents if added (type, amount). For direct X-detection (especially ^{13}C or ^{31}P), correction for nuclear Overhauser enhancement as well as saturation. For non-deuterated aqueous samples, state any corrections made for nonlinear excitation profile and method.
 - Reporting on replicate analyses, standard error/ deviation of quantification.
 - Absolute Quantification method validation is of higher rigor and performed to demonstrate that a particular method used for quantitative measurement of an analyte(s) in a given biological matrix, such as plants, blood, plasma, serum, or urine, is reliable and reproducible for the intended use (Thompson et al. 2002; FDA 2001). Suggested minimum reporting standards include:
 - Calibration curves should be generated for each metabolite to be quantified in the same biological matrix and include a sufficient number of standard solutions to adequately define the instrument response to concentration relationship (i.e. suggested minimum of at least one standard solution per order of change in concentration). The range of standard solutions used and the range of linearity with correlation coefficient should be reported.
 - A quantifier of the method accuracy (i.e. standard deviation, relative standard deviation, coefficient of variance) should be reported and bias assessed if possible (bias; due to method, lab, ion suppression, etc.).
 - A quantifier of the method precision (i.e. standard deviation, relative standard deviation, coefficient of variance) should be reported.

The lower limit of quantification (LLOQ) and confidence level should be reported. The LLOQ is defined as the minimum concentration generating an instrumental signal-to-noise response ratio of 10. The LLOQ has alternatively been defined as 5 times the limit of detection (LOD). The LOD is defined as the concentration that yields a minimum instrumental signal-to-noise ratio of 3.

Additional quantitative descriptions of recovery and/or stability provide additional method validation.

2.8 Proposed minimum metadata relative to data preprocessing

The scope of the CAWG data pre-processing standards focuses upon the conversion of raw instrumental files into organized/tabulated file formats. The organized data are then used for further statistical and chemometric analyses which are the focus of the Data Processing Working Group (Goodacre et al. 2007). The following minimum reporting standards are proposed for data pre-processing.

- Post Acquisition Data Pre-processing
 - Data file format used and/or conversion methods should be reported. Examples include conversion of proprietary file formats to more universal formats such as net.cdf, XML, MZmine, etc.
 - Details of any data pre-processing methods which convert raw instrumental data into organized or tabular file formats should be reported.

Examples for MS might include: background subtraction, noise reduction, curve resolution for temporal chromatographic alignment, peak picking, peak thresholding, spectral deconvolution, and/or metabolite identifications. Some comparative methods do not resolve or identify individual metabolites prior to comparative analysis. The general experimental details describing these methods should still be reported and should be sufficient so that others can replicate the data processing.

Examples for NMR data pre-processing might include phase-correction method (e.g. automatic, manual), conversion from time to frequency domain (e.g. Fourier Transform), degree of zero filling, degree of linear prediction; apodization parameters and window functions in all dimensions (exponential, Gaussian, sine bell etc.), baseline corrections (dc offset, linear or non-linear corrections), first point multipliers, any shifting of the free induction decays.

For data analysis of isotope labeling of flux experiments, the method for determining positional and fractional labeling, standard error of the estimates; and estimated isotope recovery in observable fractions (and fraction of total isotope supplied) should be described.

Examples for FT-IR spectroscopy might include conversion from time to frequency domain (e.g. Fourier Transform), and degree of zero filling. Baseline corrections parameters might include offsets, level and type of derivatisation (including

algorithm, window size for smoothing), and whether or not CO₂ was removed from spectra (deleted or a linear trend fitted).

2.9 Proposed minimum metadata relative to metabolite identification

Metabolite identification is a fundamental function that converts raw data into biological context. Thus, metabolite identifications are critical to the large-scale analysis of metabolites, i.e. metabolomics, and metabolite identifications should be of significant rigor to validate the identification. While it is difficult to prescribe a minimum reporting requirement for identification, the rigor of the metabolite identifications should be aligned with acceptable practices for chemical journals (see <http://pubs.acs.org/journals/jacst/> [http://www.rsc.org/Publishing/ReSource/AuthorGuidelines/ArticleLayout/sect3.asp](http://pubs.acs.org/journals/jacst/http://www.rsc.org/Publishing/ReSource/AuthorGuidelines/ArticleLayout/sect3.asp) https://paragon.acs.org/paragon/ShowDocServlet?contentId=paragon/menu_content/authorchecklist/CCcmk1.xls).

However, the exact basis for what constitutes a valid metabolite identification is still currently debated in the community and a consensus is still evolving.

Currently, four levels of metabolite identifications can be found in the published metabolomics literature. They include:

1. Identified compounds (see below).
2. Putatively annotated compounds (e.g. without chemical reference standards, based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries).
3. Putatively characterized compound classes (e.g. based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class).
4. Unknown compounds—although unidentified or unclassified these metabolites can still be differentiated and quantified based upon spectral data.

Authors should clearly differentiate and report the level of identification rigor for all metabolites reported.

The majority of metabolite identifications reported are typically non-novel as they have been previously characterized, identified, and reported at a rigorous level in the literature. Thus, non-novel metabolites not being identified for the first time are often identified based upon the co-characterization with authentic samples. However, it is generally believed that a single chemical shift, *m/z* value, or other singular chemical parameter is insufficient for non-novel metabolite identification. Thus, the following minimum standards for level 1, non-novel metabolite identification are proposed.

- A minimum of two independent and orthogonal data relative to an authentic compound analyzed under identical experimental conditions are proposed as necessary to validate non-novel metabolite identifications (e.g. retention time/index and mass spectrum, retention time and NMR spectrum, accurate mass and tandem MS, accurate mass and isotope pattern, full ¹H and/or ¹³C NMR, 2-D NMR spectra). The use of literature values reported for authentic samples by other laboratories are generally believed insufficient to validate a confident and rigorous identification. The use of literature or external laboratory data result in level 2 identifications.

- If spectral (MS or NMR) matching is utilized in the identification process then the authentic spectra used for the spectral matching should be described appropriately or libraries made publicly available. It is preferred that the reference spectra are made available at no cost, but the CAWG recognizes that this may not always be possible for commercialized libraries (NIST, Wiley, etc.). However, the premise of this minimum is that authors document and provide the spectral evidence to validate the metabolite identifications. If the authors choose not to provide the experimental evidence to support the identifications, then the identifications should be reported as 'putative identifications'.
- Metabolite identifications based upon additional orthogonal data (i.e. more than two) are highly advantageous, provide additional confidence, and are often necessary to provide unambiguous identification of stereo configuration. Additional data consistent with best chemical practices might include: selective solvent extraction, retention time, m/z , photodiode array spectra, \max and \max , chemical derivatization, isotope labeling, 2D NMR, IR spectra, etc.

2.9.1 Nomenclature for non-novel metabolites—The standard for compound nomenclature is provided by the International Union of Pure and Applied Chemistry (IUPAC, <http://www.chem.qmul.ac.uk/iupac/>). However, these rules typically result in very complex and lengthy names. As a result, IUPAC names are traditionally replaced with shorter more common names, e.g. rutin as compared to 2-(3,4-dihydroxyphenyl)-5,7-dihydroxy-3-[(2S,3R,4S, 5S,6R)-3,4,5-trihydroxy-6-[[[(2R,3R,4R,5R,6S)-3,4,5-trihydroxy-6-methyl-oxan-2-yl]oxymethyl]oxan-2-yl]oxy-chro-men-4-one. Compounds can also be referenced by numerical identifiers such as:

Chemical Abstract Service (CAS; <http://www.cas.org/>)

Chemical Entities of Biological Interest (ChEBI; <http://www.ebi.ac.uk/chebi/>)

Molfile

PubChem compound identifier (CID; <http://pubchem.ncbi.nlm.nih.gov/>)

Simplified Molecular Input Line Entry Specification (SMILES; Anderson et al. 1987; Weininger 1988; <http://www.daylight.com/smiles/>)

IUPAC International Chemical Identifier (InChI; <http://inchi.info/>)

Generally, CAS numbers are less favored due to the proprietary nature of these numbers, whereas CID, SMILES, and INCHI codes are more preferred. It is the CAWG current opinion that INCHI codes offer a favorable format for data exchange and database communication. Thus, it is suggested that authors report a minimum of one chemical name (IUPAC or common) and one structural code for all identified metabolites for publication.

2.9.2 Novel metabolite identifications—Metabolites identified for the first time and which represent novel identifications should include sufficient evidence for full stereochemical structural identification and acceptable criteria are clearly defined by most journals (i.e. <http://pubs.acs.org/journals/jacst/>, and <http://www.rsc.org/Publishing/ReSource/AuthorGuidelines/ArticleLayout/sect3.asp>, https://paragon.acs.org/paragon/ShowDocServ-let?contentId=paragon/menu_content/authorchecklist/CCCmk1.xls). This traditionally involves extraction, isolation, and purification followed by elemental analysis, accurate mass measurement, ion mass fragmentation patterns, NMR (^1H , ^{13}C , 2D), and other spectral data such as IR, UV, or chemical derivatization. The CAWG fully supports these traditional criteria for novel metabolite identifications.

2.9.3 Nomenclature for novel metabolites—For novel metabolites identified for the first time and/or compounds that are not yet included in PubChem (<http://pubchem.ncbi.nlm.nih.gov/>), formal naming should be consistent with IUPAC nomenclature and common naming is left to the author's discretion. However, author(s) are encouraged to (a) submit novel structures to PubChem and/ or (b) release an electronic code for the structure, i.e. the INCHI code that is recommended by IUPAC and NIST. The INCHI code and software to generate this code for chemical drawings is freely available (http://inchi.info/software_en.html).

2.10 Proposed minimum metadata relative to reporting of unknown metabolites

Within most metabolomics datasets, there are typically many unknown analytes, i.e. level 3 and 4 compounds. Obviously, those deemed highly important to the study should be rigorously identified according to the metabolite identification discussions above. This is not possible in all cases due to time restrictions or the lack of authentic material for unambiguous assignment. However, these unknown metabolites can often still be differentiated based upon unique experimental data, i.e. spectral or chromatographic features, and it is valuable to systematically report such “unique unknowns” in a meaningful manner to other researchers. The following minimum reporting standards are suggested for systematically naming unidentified metabolites.

2.10.1 Nomenclature for unknown metabolites

- For NMR, the exact chemical shift and multiplicity of at least one nucleus in the metabolite should be part of the unknown nomenclature. For example, an unidentified triplet at 1.16 ppm could be reported as: ‘unknown (1.16 ppm, triplet)’. When such a signal can be correlated with other atoms in the same molecule using multidimensional or multi-pulse techniques, the chemical shifts and the connectivity of such correlated nuclei in the unknown should be reported in the work. In such cases the molecular fragment may be identified, such as ‘isopropyl group’.
- For MS, the retention time, retention index, and/or prominent ions in the mass spectrum should be reported along with MS-MS data if available (also see Bino et al. 2004).
- Xenobiotics (e.g. administered drugs, related drug metabolites) or other exogenous compounds such as herbicides, pesticides, etc. should be rigorously distinguished from endogenous metabolites for all unknowns and if possible.

3 Discussions and conclusions

The Chemical Analysis Working Group will continue to work cooperatively on a consensus document that describes a minimum core set of necessary data related to the chemical analyses associated with metabolomics experiments. Further, the CAWG will work cooperatively with other MSI groups to build an integrated consensus document. The primary motivation is to establish acceptable practices that will maximize the utility, validity, and understanding of metabolomics data. It is envisioned that the proposed MSI minimum reporting standards will eventually lead to the generation of a schematic representation and model of the reporting standards to assist potential users and developers to better understand, evaluate, and utilize the proposed metadata. However, it is the general consensus of the MSI working groups that it is still a little early for this effort and additional input is needed prior to this next step. During the interim, the MSI Exchange format working group has initiated efforts to define data exchange formats and to produce a schema for such

operations that cover all aspects of the metadata, the analytical data (both spectroscopic and chromatographic) and the data analysis.

The above proposed standards do not cover all aspects of chemical analysis. Significant input is still needed within the specific areas of capillary electrophoresis, electrochemical detection, and numerous other techniques. There are also specialist areas of the mass spectrometry and NMR spectroscopy sections which may need revision or expansion to cover future consideration (e.g. in vivo NMR spectroscopy). However, we believe that the above texts provide general guidelines for improving the quality and utility of published metabolomics datasets. To achieve this objective, the CAWG invites feedback and input from the greater scientific community on the technologies and standards, and an internet discussion site has been established at <http://msi-workgroups.sourceforge.net/> or [http://Msi-workgroups-feedback@lists.sourceforge.net](mailto:Msi-workgroups-feedback@lists.sourceforge.net) to facilitate such feedback. Only through active community involvement will a functional solution be achieved.

References

- Anderson, E.; Veith, GD.; Weininger, D. SMILES: A line notation and computerized interpreter for chemical structures Report No EPA/600/M-87/021. U.S. EPA, Environmental Research Laboratory-Duluth; Duluth, MN 55804: 1987.
- Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW. Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*. 2004; 9:418–425. [PubMed: 15337491]
- Ellis DI, Goodacre R. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst*. 2006; 131:875–885. [PubMed: 17028718]
- FDA. Guidance for industry Bioanalytical method validation. 2001. <http://www.fda.gov/cder/guidance/4252f1.pdf>
- Fiehn O, Kristal B, van Ommen B, Sumner LW, Assunta-Sansone S, Taylor C, Hardy N, Kaddurah-Daouk R. Establishing reporting standards for metabolomic and metabonomic studies: A call for Participation. *Omic*s. 2006; 10:158–163. [PubMed: 16901221]
- Fiehn O, Sumner LW, Ward J, Dickerson J, Lange MB, Lane G, Roessner U, Last R, Rhee SY, Nikolau B. Minimum reporting standards for plant biology context in metabolomics studies. *Metabolomics*. 2007; 3 this issue.
- Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Capuani G, Craign A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M, Trygg J, Wlfert F. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*. 2007; 3 this issue.
- Griffin JL, Nicholls AW, Daykin C, Heald S, Keun H, Schuppe-Koistinen I, Griffiths JR, Cheng L, Rocca-Serra P, Rubtsov DV, Robertson D. Standard reporting requirements for biological samples in metabolomics experiments: mammalian/in vivo experiments. *Metabolomics*. 2007; 3 this issue.
- Jenkins H, Hardy N, Beckmann M, Draper J, Smith A, Taylor J, Fiehn O, Goodacre R, Bino R, Hall R, Kopka J, Lane G, Lange B, Liu J, Mendes P, Nikolau B, Oliver S, Paton N, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner L, Wang T, Walsh S, Wurtele E, Kell D. A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology*. 2004; 22:1601–1606.
- Jenkins H, Johnson H, Kular B, Wang T, Hardy N. Toward supportive data collection tools for plant metabolomics. *Plant Physiology*. 2005; 138:67–77. [PubMed: 15888680]
- Lindon J, Nicholson J, Holmes E, Keun H, Craig A, Pearce J, Bruce S, Hardy N, Sansone S, Antti H, Jonsson P, Daykin C, Navarange M, Beger R, Verheij E, Amberg A, Baunsgaard D, Cantor G, Lehman-McKeeman L, Earll M, Wold S, Johansson E, Haselden J, Kramer K, Thomas C, Lindberg J, Schuppe-Koistinen I, Wilson I, Reily M, Robertson D, Senn H, Krotzky A, Kochhar S, Powell J, van der Ouderaa F, Plumb R, Schaefer H, Spraul M. Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology*. 2005; 23:833–838.
- Quackenbush J. Data standards for 'omic' science. *Nature Biotechnology*. 2004; 22:613–614.

- Rubtsov DV, Jenkins H, Ludwig C, Easton J, Viant MP, Guenther U, Griffin JL, Hardy N. Requirements for the description of NMR-based metabolomics experiments. *Metabolomics*. 2007; 3 this issue.
- Thompson, et al. Harmonized guidelines for single laboratory validation of methods of analysis (IUPAC Technical Report). *Pure and Applied Chemistry*. 2002; 74:835–855.
- van der Werf MJ, Takors R, Smedsgaard J, Nielsen J, Ferenci T, Portais JC, Wittmann C, Hooks M, Tomassini A, Oldiges M, Fostel J, Sauer U. Standard reporting requirements for biological samples in metabolomics experiments: Microbial and in vitro biology experiments. *Metabolomics*. 2007; 3 this issue.
- Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*. 1988; 28:31–36.