Ontology-based metabolomics data integration with quality control

Patricia Buendia¹, Ray M Bradley¹, Thomas J Taylor¹, Emma L Schymanski^{2,3}, Gary J Patti⁴ & Mansur R Kabuka^{*,1}

¹INFOTECH Soft, Inc., 1201 Brickell Ave. Suite 220, Miami, FL 33131, USA

²Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Campus Belval, 6 Avenue du Swing, Belvaux L-4367, Luxembourg

³Eawag – Swiss Federal Institute of Aquatic Science & Technology, Überland Strasse 133, Dübendorf 8600, Switzerland

⁴Departments of Chemistry, Genetics, & Medicine. Washington University, Saint Louis, MO 63110, USA

*Author for correspondence: kabuka@infotechsoft.com

Aim: The complications that arise when performing meta-analysis of datasets from multiple metabolomics studies are addressed with computational methods that ensure data quality, completeness of metadata and accurate interpretation across studies. **Results & methodology:** This paper presents an integrated system of quality control (QC) methods to assess metabolomics results by evaluating the data acquisition strategies and metabolite identification process when integrating datasets for meta-analysis. An ontology knowledge base and a rule-based system representing the experiment and chemical background information direct the processes involved in data integration and QC verification. A diabetes meta-analysis study using these QC methods finds putative biomarkers that differ between cohorts. **Conclusion:** The methods presented here ensure the validity of meta-analysis when integrating data from different metabolic profiling studies.

First draft submitted: 16 November 2018; Accepted for publication: 1 May 2019; Published online: 10 June 2019

Keywords: data integration • diabetes use case • meta-analysis • metabolomics • ontology-based expert system • quality control

Metabolomics investigations have been applied in various research areas including environmental and biological stress studies, functional genomics, biomarker discovery and integrative systems biology [1]. The next step in the metabolomics field is the development of algorithms and methods to integrate metabolite data from different studies as is currently done for genetic variation by genome-wide association studies [2]. In metabolomics, as in any biomedical field, understanding what mediates the connection between a risk factor and an outcome can be improved with larger sample sizes [3,4]. There are several analytical technologies and separation methods used in metabolomics, with the most sophisticated techniques being gas chromatography (GC), liquid chromatography (LC) normally hyphenated with mass spectrometry (MS) as (GC–MS) and (LC–MS) respectively, and nuclear magnetic resonance (NMR) [5]. These methods differ in specificity and sensitivity. Moreover, studies may differ in sample collection and preparation protocols used. Therefore, there can be distinctly different metabolites measured across different technologies and studies making it necessary to understand the protocols that were used to generate the data for proper interpretation and re-analysis of results [6,7]. In this paper, we present methods that may be used in assessing metabolomics data quality in large-scale and multisite repositories or in medical research datasets from individual laboratories with sights on data reuse as presented in a use case study. We also describe how to apply these methods to build MODIS, a Metabolomics Data Integration System that annotates and combines metabolomics data from different sources.

Several metabolomics meta-analysis studies have been performed in recent years studying various diseases from diabetes to cancer [8–10]. It has been argued that the variation of measurements in metabolomics experiments needs to be accounted for when doing a re-analysis or meta-analysis of the results. The Consortium of Metabolomics Studies (COMETS) promotes collaborations among various prospective cohort studies and is seeking to identify the best strategies for performing metabolite profiling meta-analyses across multiple studies [11]. Three comprehensive

Figure 1. A representative metabolomics workflow. From sample preparation to identification and analysis of metabolites, quality control considers each procedure applied to the data that affects the results.
MS: Mass spectrometry; NMR: Nuclear magnetic resonance; QC: Quality control.

Figure 2. Data integration process and components in metabolomics data integration system. Dark blue boxes represent data handling processes, light blue boxes represent components of the knowledge base. See also Supplementary Table 15. MIME: Minimum information about metabolomics experiments; MODIS: Metabolomics data integration system; TCG: Technology compatibility graph; QC: Quality control.

quality control (QC) studies that compare results from different laboratories using the same samples were performed in recent years [12–14], revealing in one study discrepancies in metabolite detection across laboratories, and in the other studies, concordance of results when the same software and protocols were used. Given these recent initiatives in assessing the reliability of metabolomics results, a first step in meta-analysis or re-analysis of metabolomics results is to validate them for correct biological interpretation [15–17]. The Chemical Analysis Working Group (CAWG) of the Metabolomics Standards Initiative (MSI) recommended minimum reporting standards for chemical analysis in metabolomics [18] that have been adopted in MODIS. The compliance of 399 public datasets, from four major metabolomics data repositories, was evaluated to the biological context MSI reporting standards, with the finding that none of the reporting standards were complied with in every publicly available study [19]. Other groups have issued new recommendations that were also adopted in MODIS and focus on assigning a confidence level (CL) to the metabolites identified in a study [17,18,20,21]. A 2014 Guideline Document by the EU Initiative ‘Coordination of standards in metabolomics (COSMOS) [22] suggests the use of OWL and SWRL, the Semantic Web Rule Language [23], for validation checks of MetaboLights data [24], a public database storing metabolomics studies and data, but an implementation has not been reported.

The following sections describe MODIS, a prototype implementation of a minimum reporting standard for collecting information about a metabolomics experiment (MIME) and an OWL/SWRL-based implementation of a metabolomics quality assurance method, both applied to a meta-analysis example using datasets from multiple diabetes studies. The description of the tool prototypes shown in Figure 2 will allow bioinformaticians and

Table 1. Minimum information about metabolomics experiment with mass spectrometry list of required information.

Column title	Description	Example	Col Type	Score	Value domain under→ MODIS→ MODIS experimental procedures→ MS
Analysis type	Analysis type	Targeted tandem MS	CODE	1	→ MS analysis type
MS ionization type	MS ionization type	Negative ESI	CODE	1	→ MS ionization type
MS instrument name	MS instrument name	Agilent 6530 QTOF	CODE	1	→ Instrument name
MS instrument type	MS instrument type	QTOF	CODE	1	→ Instrument type
Source of QC or reference sample	Source of QC or reference sample	QC from pooled samples	CODE	1	→ Standard or QC reference materials
Spectral MS library	Spectral libraries of MS standard references used for metabolite matching	NIST11 library	CODE	1	→ Spectral libraries of standard references

ESI: Electrospray ionization; MODIS: Metabolomics data integration system; MS: Mass spectrometry; QC: Quality control; QTOF: Quadrupole time-of-flight.

researchers to build and customize a MODIS system and use it to evaluate the chosen metabolomics workflow procedures used in a study of interest, as shown in Figure 1, before using its results in meta-analysis studies.

Methods

Minimum information about metabolomics experiments

For public dissemination, it is essential that metabolomics datasets include sufficient metadata to allow other researchers to understand the biological and technical origins of the data in enough detail to be able to reproduce essentially the same results. We designed MIME to capture the minimum required information as well as other relevant but optional information that needs to be available at the time of data integration in order to perform QC verification of the metabolomics experiment and of the methods that produced the results. Figure 1 illustrates the general tasks involved in collecting, storing and preparing samples, generating and processing metabolomics data, with MIME collecting details from each of these steps in order to assess confidence in the data and guide in the reproducibility of experiments.

MIME builds upon the ISA-TAB framework [25] which is built around an Investigation-Study-Assay data model and has been developed with several international collaborators, including MAGE-Tab. MIME is organized as a Project-Study-Sample-Experiment-Metabolite hierarchy with three sections: experiment QC covering project, study, sample and experiment; data processing QC covering bioinformatics data processing; and metabolite identification. MIME conforms to the metabolomics ‘best practices’ proposed by the MSI [26] for reporting standards and guidelines for metabolomics metadata and data processing [27], including MSI ontology concepts [28], and reporting biological context for mammalian clinical trials and human studies, and chemical analysis [18].

Experiment QC

Experimental QC includes considerations that match the ISA-Tab ‘Investigation’ (the project context), ‘Study’ (a unit of research) and the Sample and Experiment information adjusted to the reporting standards for metabolomics studies [1]. The required information to describe the sampling and extraction process was defined based on [18] and the Metabolomics Workbench Template [29]. Experimental information includes details about the technology used to generate the metabolomics datasets. Tables 1 and 2 show required and optional information linked to the MS technology. The ‘Value’ domain column links to a standard terminology or a set of enumerated values, which in MODIS is encoded in an ontology, an approach adopted from the i2b2 project [30], with the column showing the path in the ontology. The score of the information is a value that defines the significance of the information and is multiplied by an evidence score from the ontology; if the information is required, the score is 1. Optional details, such as seen in Table 2 (e.g., capillary temperature celsius and capillary voltage for MS information) are used to compute a quality protocol score (QPS) which is described in the MODIS Ontology section.

Data processing QC

Data processing QC covers commonly used methods which convert raw instrument files into organized, tabulated file formats ready for statistical analysis. Batch correction is especially necessary for MS, where batch-to-batch variation is commonly seen. Batch variation can be dealt with in different ways, for example, by using internal

Table 2. Excerpt of minimum information about metabolomics experiment with mass spectrometry list of optional information.

Column title	Description	Example	Column type	Score	Value domain under→ MODIS→ MODIS experimental procedures→ MS→ Analysis
Capillary temperature celsius	Capillary temperature celsius	300	Decimal	0.5	→ Capillary temperature celsius
Capillary voltage	Capillary voltage	3500	Decimal	0.5	→ Capillary voltage
Collision energy	Collision energy V	10	Decimal	0.5	→ Collision energy V
Collision gas	Collision gas	Argon	CODE	0.5	→ Collision gas
Dry gas flow	Dry gas flow l/h	900	Decimal	0.5	→ Dry gas flow l/h
Dry gas temperature celsius	Dry gas temperature celsius	325	Decimal	0.5	→ Dry gas temperature celsius
Fragment voltage	Fragment voltage	120	Decimal	0.5	→ Fragment voltage
Fragmentation method	Fragmentation method	CID	CODE	0.5	→ Fragmentation method

CID: Collision-induced dissociation; CODE: Code linking to ontology term; MODIS: Metabolomics data integration system; MS: Mass spectrometry.

standards as controls, or by injecting reference or QC samples at regular intervals [16,31–33] making it necessary to collect that information when trying to reproduce the data processing and analysis steps. QC samples are used in targeted analytical methods to determine the accuracy and precision of the analytical method for each metabolite [16]. The batch correction method used in a study is captured in MIME. The most often used batch correction methods are regression models, among them, locally estimated scatterplot smoothing (LOESS) or the quality control-robust LOESS signal correction (QC-RLSC). is a widely accepted method for QC-based shift signal correction for integration of metabolomic data from multibatch experiments [31].

Metabolite identification QC & CLs

Metabolite identification QC fields define clear levels of confidence for identified and putatively identified metabolites so that a CL can be calculated. Guidelines were established by the MODIS Advisory Board based on recommended standards [17,18,20,21]. Different technologies require that different information be supplied to assess the quality of the experiment and the confidence in the metabolite identification. As shown in Tables 3 & 4, five levels, from 1 to 5, are assigned to metabolites, from the highest level (1) for identified compounds to the lowest level (5) for unknown or unclassified metabolites with a mass or spectral data. The assignment of a CL is performed automatically by the rule-based system in MODIS as described in the ‘Technical Implementation’ section.

Guidelines to assign CLs

Tables 3 and 4 present the guidelines to assigning a CL for metabolite identification based on the evidence that needs to be present. A status value of ‘published’ indicates whether the evidence was previously proposed [18,20,21], while ‘draft’ indicates whether it was proposed by the MODIS Advisory Panel as an addition to the previous requirements. A distinction is made between physicochemical properties that are collected in a results matrix per identified metabolite, and sample-based results, such as an MS spectra, that are collected in study data files. The Code or Path column describes ontology paths or ontology term codes.

Quantitative scoring system

A quantitative scoring system for summarizing a metabolite identification confidence based on spectral match was adopted from [34] with the addition of a score for GC–MS, a technology which produces metabolite identifications with higher level of confidence than LC–MS. The scores (see Supplementary Table 9) can be modified or augmented by a MODIS QC administrator. The identification score is computed from the sum of all the data types that support the identification, then multiplied by either 2 when matching to an authentic reference compound because it provides the highest confidence, or multiplied by 1.5 when matching to literature values or other externally generated spectral libraries due to a lower confidence.

Table 3. Metabolite identification quality control confidence level 1.

Information	Status	CODE or path
All experiments		
Experiment spectral data	Published	See section 'Experiment Spectral Data' (In Supplementary Material)
A reputable database ID was provided for an identified compound that is not a class	Published	→ MODIS→ Metabolite identification→ Metabolite identifier (RM)
GC-MS		
Authentic reference spectra used for spectral matching	Published	METID:AuthRefStandSpectra (STDF), OR METID:MASS+INTENSITY.AUTHREF.STANDARD (RM)
Retention index	Published	→ MODIS→ Metabolite identification→ Metabolite evidence→ Experiment evidence\retention index (RM)
Confident RI match value	Draft	METID:RI.MATCH.VALUES (RM)
Confident electron impact spectral match value	Draft	METID:EI.MATCH.VALUES (RM)
GC-MS/MS		
Authentic reference spectra used for spectral matching	Published	METID:AuthRefStandSpectra (STDF), OR METID:Majorionfragint.AuthRefStandard (RM), OR METID:FRAGMENTS.INTENSITIES.AUTHREF.NEGATIVE (RM)
Retention index	Published	→ MODIS→ Metabolite identification→ Metabolite evidence→ Retention index
Accurate mass of parent ion (m/z)	Published	METID:ACCMASSPARENTION (RM), OR METID:ACCMASS.PARENT.ION.NEGATIVE (RM)
Confident RI match to reference standard	Draft	METID:RI.MATCH.VALUES (RM)
Spectral match value	Draft	METID:SPECTRAL.MATCH.VALUE (RM)
LC-MS/MS		
Retention time	Published	METID:RT (RM)
Retention time of authentic standard	Draft	METID:RT.AUTH.STANDARD (RM)
Spectral match value	Draft	METID:SPECTRAL.MATCH.VALUE (RM)
2D NMR		
Authentic reference spectra used for spectral matching	Published	METID:AuthRefStandSpectra (STDF)
NMR chemical shift file	Published	METID:NMR.ChemicalShiftsFile (STDF)
Pulse sequence	Published	Always required in MODIS. Stored in CustomPulseSequence column of NMR Protocol Dimension
(1) Level 1: IDENTIFIED CompoundA Level 1 confidence is given to an IDENTIFIED compound when these orthogonal data relative to an authentic standard match are provided and a minimum quality protocol score threshold is reached as computed from the supplied protocol details. 2D-NMR: 2-dimensional NMR; GC-MS: Gas chromatography–mass spectrometry; GC-MS/MS: Gas chromatography–tandem mass spectrometry; LC-MS/MS: Liquid chromatography–tandem mass spectrometry; MODIS: Metabolomics data integration system; NMR: Nuclear magnetic resonance; RM: Results matrix; RI: Retention index; STDF: Study data files.		

Results & discussion

Technical implementation

This section describes how the components of the MODIS are incorporated into a Web-based platform to address the challenges of annotating and combining metabolomics data from different sources and preparing it for analysis by running QC checks on the data. This is achieved through several components that work together as shown in Figure 2: Data from a study are ingested into the knowledge base using the MIME specification to collect information on the quality and completeness of the data, while the MODIS ontology is used to annotate the fields in the data, and the expert rules to validate the quality of the data. Additional details about metabolomics data integration and the MODIS prototype is available at www.infotechsoft.com/metabolomics.

MIME templates for a dimensional database model

MIME-based annotation combines the use of defined data structures and controlled vocabularies to facilitate computer-based processing of annotated data based on published guidelines [17,18,20,21]. The MIME protocol and data sections are translated into a computer-readable JavaScript Object Notation (JSON) Warehouse Template (WHT) to ensure that all data within the MODIS warehouse are annotated in a standardized and systematic manner and in conformance to the specification. The WHT organizes protocols and metabolite information in

Table 4. Metabolite identification quality control confidence levels 2 through 5.

Information	Status	CODE or path
Level 2: PROBABLE IDENTIFIED compoundA		
All experiments:		
– Name of external library used for matching	Published	METID:EXTERNAL_LIBRARY (RM)
– Experiment spectral data	Published	See section ‘Experiment Spectral Data’ (In Supplementary Material)
– A reputable database ID was provided for an identified compound that is not a class	Published	→ MODIS→ Metabolite identification→ Metabolite identifier (RM)
– Matched spectrum identifier from external library	Draft	METID:MATCHED_ID.EXTERNAL_LIB (RM)
For GC–MS, GC–MS/MS, LC–MS/MS, 1D-NMR, 2D-NMR:		
– The same information as in level 1 but measurements from the authentic spectra are replaced with measurements from the external laboratory data or literature as specified in the Supplementary Material.		
Level 3: PUTATIVELY CHARACTERIZED structure		
Experiment spectral data	Published	See section ‘Experiment Spectral Data’ (In Supplementary Material)
A reputable database ID or an identifiable metabolite name was provided	Published	→ MODIS→ Metabolite identification→ Metabolite identifier (RM)
For GC–MS experiments:		
– Retention index	Published	→ MODIS→ Metabolite identification→ Metabolite evidence\retention index (RM)
For GC–MS/MS, LC–MS, LC–MS/MS:		
– Retention time or retention index	Published	METID:RT (RM), OR → MODIS→ Metabolite identification→ Metabolite evidence\retention index (RM)
– Accurate mass of parent ion (m/z)	Published	METID:ACCMASSPARENTION (RM)
Level 4: UNKNOWN OR UNCLASSIFIED metabolite with a molecular formula or known identifier		
Unequivocal molecular formula Published METID: molecular formula (RM)		
A reputable database ID or an identifiable metabolite name was provided	Published	→ MODIS→ Metabolite identification→ Metabolite identifier (RM)
Level 5: UNKNOWN OR UNCLASSIFIED metabolite with an exact mass (m/z) or spectral data		
Accurate mass of parent ion (m/z)	Published	METID:ACCMASSPARENTION (RM)
For GC–MS or NMR:		
– Experiment spectral data	Published	See section ‘Experiment Spectral Data’ (In Supplementary Material)
(2) Level 2 confidence is given to a PROBABLE IDENTIFIED compound when: spectral matching is used and all the information as in level 1 is provided, but an authentic reference standard for unequivocal proof is not provided.		
(3) Level 3 confidence are given to either: a compound class (a parent term in PubChem); a metabolite, not a class, was uniquely identified, but some information required in level 1 or 2 are missing; for MS: Level 3 even for a uniquely identified metabolite, not a class, when identification is not definitive as with LC–MS, but this information is required.		
(4) Level 4 confidence is given to UNKNOWN, UNIDENTIFIED OR UNCLASSIFIED metabolites with a molecular formula that can still be differentiated and quantified based upon spectral data.		
(5) Level 5 confidence is given to UNKNOWN, UNIDENTIFIED OR UNCLASSIFIED metabolites with an exact mass (m/z).		
1D-NMR: 1D-Nuclear magnetic resonance, 2D-NMR: 2D-Nuclear magnetic resonance; CODE: code linking to ontology term; MODIS: Metabolomics data integration system; NMR: Nuclear magnetic resonance; RM: Results matrix;		

dimension tables and measurement data in fact tables to build a dimensional database model. The WHT in turn is used in combination with a wizard template (WZT) to automatically build a customized user interface. The WZT defines: names of metadata fields; data type and range of allowed values that a field may be assigned; links to controlled vocabularies; written instructions for appropriate completion of each field; details about the equipment, software, and protocol of the experiment; additional fields that must be completed based on the responses; and set of expected data artifacts and file formats for the given experiment. The resulting user interface guides the user in the submission of metabolomics data and experiment protocols, providing two main sections, ‘protocols’ and ‘data’, and multiple steps in each to submit information about a study to the dimensional database. (See [Supplementary Table 12](#) for a generic WZT design.)

MODIS ontology

The MODIS ontology represents the standard vocabulary and a knowledge repository used to annotate the data in the MODIS warehouse. The MODIS Ontology is a virtual ontology that is customizable by the scientists involved in metabolomics research. See [Supplementary Table 6](#) for a list of external ontologies added to MODIS. The ontology is closely linked to MIME as it describes the set of allowed terms with their evidence scores used to

annotate a metabolomics experiment and its results. Evidence scores are used to calculate the QPS, which is the sum of the evidence scores for information collected from all protocols and is used to describe the level of confidence for a complete experiment.

With the introduction of the Semantic Web [35], the term ‘ontology’ gained usage in Computer Science as a technique used to represent and share knowledge about a domain using standard, machine-processable languages (e.g., RDF [36] and OWL [37]). Figure 3A shows the Metabolomics category with MODIS Technology expanded. The MODIS ontology contains knowledge from chemical experiments and technologies, specifically metabolomics-related knowledge, and includes ontologies from peer-reviewed, large-scale ontology repositories such as the Open Biomedical Ontologies (OBO) Foundry established by the OBO consortium [38,39], among them, the Chromatography ontology seen in Figure 3A, which was extracted from the chemical methods ontology (see Supplementary Table 6). The MODIS ontology supports knowledge-based decisions, terminology management and systems interoperability and integration with an evidence score or rule that is assigned to ontology terms by an ontology curator. The evidence score is linked to MIME and describes the significance of annotating a study with that term. The more important the information, the higher the score. The ontology term ‘evidence reference’ field contains references, as internet hyperlinks, to the source that justifies the evidence score. As an example, a small anticoagulant ontology is shown in Figure 3B, in which citrate has an evidence score of 0.5. The evidence reference contains the URL for the Metabolon Sample Preparation Guidelines document [40], which states the following with regards to preference for anticoagulants for plasma samples: best results: EDTA (K2, K3, Na; avoid Li); acceptable results: citrate; and less desirable: heparin. Based on those guidelines, K2, K3 and Na were given a reference score of 1, Citrate a 0.5, Heparin a 0.2 and Li a -1.0 (negative 1). When a MODIS user selects an anticoagulant from the ontology, the system multiplies this score, which describes the content or value of required information, by the MIME evidence score (column score in Tables 1 & 2), which describes how significant the information is, and adds it to the cumulative QPS score. When all information is collected using the MIME wizard described earlier, the QPS score is computed and compared with the minimum value allowed, which is defined by the system administrator.

Identifier matching

Metabolomic datasets use identifiers from publicly available databases such as KEGG, HMDB, PubChem and others. In some cases, metabolites are identified by compound names without referencing a database. Compound names are very poor descriptors as names often cannot be unambiguously mapped to authentic chemical structures. A complication when integrating datasets is that the same metabolite in different studies is associated with different database IDs. An entity mapping method was therefore developed to combine metabolite result matrices obtained from different sources that use different database IDs. The solution involves decoupling of the source entities from the data warehouse entities and addresses issues with using multiple identifiers for the same metabolite. In MODIS, a ‘map’ table is used to link the record in the source system with the corresponding entity in the dimensional warehouse to encapsulate these identifier relationships (See Supplementary Figure 3).

The metabolite identification that is selected in the results of a study affects the assignment of the CL, as compounds may be characterized by the same name or identifier, or by different names that map to the same compound, or by a formula that may be linked to a unique compound or multiple isomers.

Table 6 in the Results section offers examples of issues that occur during identifier mapping. The BridgeDb framework [41] and the Chemical Translation Service [42] are used within MODIS for standardized access to metabolite identifier mapping services.

MODIS expert system

A QC rule-based expert system is applied to new data submitted to the MODIS warehouse to estimate CLs of identified metabolites. The CLs can be computed when data are stored in the database or recomputed, if rules change, from the data already stored in the database. The system builds on standard ontologies from the OBO Foundry [38]. The primary base ontology is the ontology for biomedical investigations [43], which provide concepts for describing all aspects of how investigations in the biological and medical domains are conducted.

Metabolite profiling, as shown in Figure 4, is modeled as multiple processes that include the assay; preprocessing of spectra produced by the assay, such as the peak matching performed by XCMS; and finally, metabolite matching where the processed spectra are matched to spectrum libraries, or the spectrum of an authentic compound.

Figure 3. Metabolomics data integration system ontology in the metabolomics data integration system user interface. (A) High-level concepts within metabolomics ontology. **(B)** Detailed view of the anticoagulant ontology with selected term 'citrate' showing evidence scores and references.

Table 5. Semantic web rule language rules for a confidence level 1 for a metabolite detected by GC–MS Metabolite profiling.

Natural language	SWRL
IF information about the experiment was acquired AS PART OF a MPE AND mass spectrum datasets were acquired THEN spectral data were uploaded	$\text{MetaboliteProfiling}(?a) \wedge \text{InformationAcquisition}(?b) \wedge \text{MassSpectrumDataSet}(?c) \wedge \text{hasPart}(?a, ?b) \wedge \text{hasSpecifiedOutput}(?b, ?c) \rightarrow \text{hasUploadedSpectraData}(?a, \text{true})$
IF the GC-MS technology was used AS PART OF a MPE THEN GCMP was performed	$\text{MetaboliteProfiling}(?a) \wedge \text{GcMassSpectrometry}(?b) \wedge \text{hasPart}(?a, ?b) \rightarrow \text{GCMetaboliteProfiling}(?a)$
IF ACM was performed AS PART OF a GCMP AND the experiment mass spectrum was uploaded AND retention index IS INPUT OF the ACM AND the authentic mass spectrum IS INPUT OF the ACM AND the experiment mass spectrum IS INPUT OF the ACM AND an IMR IS OUTPUT OF the ACM AND a metabolite identifier IS PART OF the IMR THEN it satisfies confidence level 1	$\text{GcMetaboliteProfiling}(?a) \wedge \text{hasUploadedSpectralData}(?a, \text{true}) \wedge \text{RetentionIndex}(?c) \wedge \text{MassSpectrum}(?d) \wedge \text{MetaboliteId}(?e) \wedge \text{MassSpectrum}(?g) \wedge \text{AuthenticCompoundMatching}(?h) \wedge \text{IdentifiedMetaboliteRecord}(?i) \wedge \text{hasPart}(?a, ?h) \wedge \text{hasSpecifiedInput}(?h, ?c) \wedge \text{hasAuthenticMSSpectrumInput}(?h, ?d) \wedge \text{hasSpecifiedInput}(?h, ?e) \wedge \text{hasExperimentMSSpectrumInput}(?h, ?g) \wedge \text{hasSpecifiedOutput}(?h, ?i) \wedge \text{hasPart}(?i, ?e) \rightarrow \text{satisfiesConfidenceLevel}(?e, 1)$

ACM: Authentic compound matching; GCMP: GC–MS metabolite profiling; IMR: Identified metabolite record; MPE: Metabolite profiling experiment; SWRL: Semantic web rule language.

Table 6. Confidence levels and quantitative score for metabolites identified in two different studies.

Metabolite name	Mapped PubChem ID	Project; study	CL	QS	Source of evidence
Citrate	311	PR000330; ST000383	1	7	GC–MS: HR retention index, GC–MS spectra, nominal mass of parent ion, mass spec list of mass + intensity, PubChem ID, Kegg ID
Citric acid	311	PR000330; ST000443	4	0	GC–MS
Asparagine	6267	PR000330; ST000383	1	7	GC–MS: HR retention index, GC–MS spectra, nominal mass of parent ion, mass spec list of mass + intensity, PubChem ID, Kegg ID
Asparagine	236	PR000330; ST000435	4	0	LC–MS/MS
1-Monostearin	24699	PR000330; ST000383	1	7	GC–MS: HR retention index, GC–MS spectra, nominal mass of parent ion, mass spec list of mass + intensity, PubChem ID, Kegg ID
1-Octadecanoyl-rac-glycerol	24699	PR000330; ST000421	4	0	UPLC–ToF MS+
2-Deoxyerythritol NIST	18302	PR000330; ST000383	1	7	GC–MS: HR retention index, GC–MS spectra, nominal mass of parent ion, mass spec list of mass + intensity, PubChem ID, Kegg ID
C4 H10 O3	18302, 640997, 6994279	PR000330; ST000421	4	2	UPLC–ToF MS+: Molecular formula
281363 (BinBase ID)		PR000330; ST000383	3	7	GC–MS: HR retention index, GC–MS spectra, nominal mass of parent ion, mass spec list of mass + intensity, BinBase ID
194.115@0.8462586		PR000330; ST000421	5	4	UPLC–ToF MS+: Accurate mass of parent ion, retention time

CL: Confidence level; HR: High resolution; NIST: National Institute of Standards and Technology; QS: Quantitative score; UPLC–ToF MS: Ultra performance liquid chromatography time-of-flight mass spectrometry.

Figure 4. Subprocesses of metabolite profiling. The squares are used to represent metabolite profiling processes, while the ovals are used to represent information content entities from the information artifact ontology.

Information passed between the processes is captured using terms from the information artifact ontology [44], an ontology for capturing the essence of information entities and how their components relate to each other.

Rules for the expert system were defined using SWRL [23], which extends the set of OWL axioms to include Horn-like rules allowing them to be combined with an OWL knowledge base. As can be seen in Table 5, which shows an example SWRL rule for assigning a CL of 1, the rules take the form of an implication between an antecedent (body) and consequent (head). Tables 3 and 4 list the rules that were defined to determine the possible CL of metabolites identified within a metabolite profiling process.

Application of the rules to determine the CL of metabolite identification was performed by first transforming experiment data submitted to the system into a fact base conforming to the previously described knowledge base (see Supplementary Figure 1). The knowledge base, fact base and rules are then loaded into the Drools rule engine using the SWRLAPI Drools [45] bindings. Once loaded into Drools, the rules are executed and the results extracted using a SPARQL query to retrieve the maximum satisfied CL.

Technology compatibility graph

For meta-analysis, it is important to understand the technical compatibility of study results. A problem in untargeted metabolomic studies is that one only sees a subset of the metabolites depending on the technology. Biological information also causes variability, such as peaks in fecal samples appearing larger than other samples. Some of this variability will be captured by the technology compatibility graph (TCG). The TCG is used to warn about merging data together that is not compatible; for example, a metabolite that is over-represented in one study but not present in another study that examines the same condition could be erroneously discarded as a biomarker, but with the TCG, the researcher will know not to discard it if it was never detected in that type of sample or by that particular experiment, software or instrument.

The TCG is a collection of bipartite graphs with weighted edges. One set of vertices contains named metabolites or compound classes, the other set contains nonmetabolite terms related to a metabolomics sample, experiment or analysis from the ontology. The TCG is built periodically from the data stored in the MODIS warehouse by running ontology-based queries with different study conditions as selected search terms. The weight of the edges is the number of times the link between metabolite and ontology term was discovered in the warehouse. Beyond evidence scores and CLs, as another layer of QC validation, the TCG reveals how often or seldom a metabolite is detected using the methods described in a study. This information aids in determining a compatibility score when data are being merged or studied together. An example of an unweighted TCG graph is shown in Figure 5, in which the ontology term is 'experiment type' and used to select metabolites common across technologies (Supplementary Figure 4 shows the TCG for all metabolites across experiment types).

Data analysis framework

A data analysis framework (DAF) was developed for meta-analysis of integrated metabolomics data as described in the use case below using conceptual integration and concatenation-based integration. The construction of meta-analysis workflows is based on XML input/output definitions maintained by a tool catalog and coordinated by the DAF client. The DAF client features a user interface, the Workflow Builder, to enable the design and execution of data analysis workflows. It presents a selection of metabolomics data processing tools from the tool catalog that are available for use in the design of a workflow and enables augmenting the tool catalog with new tools. A prototype of the DAF execution environment was installed on multiple Linux and Windows computing nodes. HTCondor was installed on each computing node to provide high throughput grid computing and the Pegasus Workflow Management System was installed to manage the DAF scientific workflow execution on the HTCondor grid. The DAF prototype implementation features one cluster node supporting tools for metabolomics analysis. To expand it to statistical analysis, the cluster was expanded with an R cluster node. The MetaboAnalystR package was installed in the R node.

Diabetes studies use case

To demonstrate the metabolomics integration techniques presented in this paper, a search for case-control studies with metabolomics data was performed in the public metabolomics databases Metabolomic Workbench and MetaboLights. The search criteria required availability of phenotype information, the table of identified metabolites with measurements, the same medical condition, same specimen type and same sample collection protocol. Intervention study designs were filtered out, as interventions are rarely the same between studies and bias the

Figure 5. Graph of common metabolites labeled as detected by all technologies (GC-MS, LC-MS and NMR).
Metabolites: pink, organic acids: orange and volatiles: red.

measurements. Four diabetes studies matched the filter criteria, but only two had complete information to pass the QPS threshold. Two datasets were downloaded from the Metabolomics Workbench, a Type 1 diabetes study with project ID PR000330 with five linked substudies using LC-MS, LC-MS/MS and GC-MS and a Type 2 diabetes study with project ID PR000300 with a single linked study, ST000383, that used the GC-MS technology. In both cases, plasma was collected after an overnight fast.

The study protocol information was entered into the MODIS system for both studies and obtained a QPS score above the current required threshold of 22, as both studies were well annotated. PR000300 listed identified metabolites with multiple equivalent database IDs, such as Binbase, KEGG and PubChem. PR000330, however, in some cases only listed metabolite names, while in other cases it listed molecular formulas or mass and retention time. The MODIS ID mapper was used to standardize and map identifiers for the linked substudies.

Table 6 shows the assigned CL and quantitative score (QS) for a selection of metabolites and the evidence used to compute the values. A lower CL number indicates greater confidence in metabolite identification. The source of evidence column describes how the CL and QS numbers were computed. For example, the QS for study ST000383 is computed as follows: metabolites had a 'high resolution retention index' (score: 2), a 'nominal mass of parent ion' (score: 0.5), and a 'GC-MS mass spec list of mass + intensity' (score: 1), which adds up to $2 + 0.5 + 1 = 3.5$. As authentic reference spectra were used in the study, the score is multiplied by 2, resulting in a final QS of 7. PR000330 used different technologies in the substudies. For the untargeted metabolomics studies, the metabolite identification varied from a name to a formula or the mass and retention time, which explains the varying QS and CL scores.

The same metabolite could be detected in both projects even when it was described by completely different names, as is the case with 1-octadecanoyl-rac-glycerol (in PR000330) and 1-monostearin (in PR000300). In some cases, the mapped ID conflicted with the ID recorded in the data, for example, asparagine in Table 6 had different PubChem IDs, but the metabolites were matched based on other database IDs that matched, for asparagine it was through HMDB ID HMDB00168.

Once the metabolites of all studies are assigned QC scores, the researchers can select a threshold score to filter out the metabolites that do not meet the threshold criteria. In our meta-analysis use case, only the metabolites that were detected in both projects and had a confidence level 1 through 4 were kept, as mass values alone are not comparable between studies. Between the two projects, 38 common metabolites were found conforming to the chosen CLs.

Statistical analysis

Conceptual data integration

Conceptual integration refers to the situation where multiple datasets are analyzed separately, and then, the resulting conclusions are matched. It was applied to our use case study and produced valuable insights into putative biomarkers that differentiate poor glycemic control (gc-) from good glycemic control (gc+), but no clear biomarkers differentiated them from nondiabetics (ND). The datasets from PR000300 and PR000330 and their linked substudies were filtered, normalized and then submitted to statistical analysis. T-tests, principal component analysis, partial least squares determinant analysis (PLS-DA), significance analysis of metabolites (SAM), volcano plot, fold change analysis and random forest analysis were performed on different combination of datasets. For PR000330, four substudies were planned, pairwise comparison of (gc+ and ND), (gc- and ND) and (gc+ and gc-), and comparison of all three groups. Two analyses were carried out, one for all targeted metabolites and one only for amino acids.

In the all-metabolite analysis, random forest analysis of PR000330 for the four substudies had high out-of-bag (OOB) errors of 0.264 and larger, indicative of poor predictive accuracy. OOB error, also called OOB estimate, is a method of measuring the prediction error of random forests. For PR000300, the OOB error was 0.143. Very little overlap between any of the important factors were found via t-tests, random forests, or PLS-DA.

In the amino acid study, 27 amino acids of study ST000435 and related compounds were matched to the ST000383 study. ST000435 (ND_gc-) identified five features among its top ten in common with ST000383: aspartic acid, serine, glycine, histidine and phenylalanine. The gc+/- datasets returned a random forest OOB error of 0 indicative of a high predictive accuracy for the classifier (see Figure 6). The same set of amino acids in the ST000383 study has an OOB error of 0.196. A t-test on all 27 amino acids also selects the first four metabolites with an FDR corrected p-value of <0.0000156 as significant features that separate good from bad glycemic control. The high predictive accuracy of the random forest classifier for the gc+/- amino acids in the ST000435 dataset indicates that aspartic acid, phosphoethanolamine, proline, cystine and histidine may be putative biomarkers that differentiate good from bad glycemic control.

Concatenation-based integration

A dataset concatenation-based integration method (see Supplementary Table 13) was used to analyze the combined ST000435 and ST000383 amino acid datasets. The data tables produced by each study were first normalized to median, mean-centered and divided by the standard deviation of each variable. Once normalized, the datasets were concatenated into a single data table with two groups: ND and diabetic. Principal component analysis and PLS-DA did not obtain good separation of the groups. The random forest classifier had an OOB error of 0.295. However, serine was found to be a significant feature distinguishing ND from diabetic cases by the analysis methods SAM ($q = 0.028$), volcano plot ($p = 0.0046$) and fold-change analysis (2.0851) and the top feature identified by PLS-DA, sPLS-DA, oPLS-DA and random forest (see Figure 6). Serine was observed to have a twofold increase in diabetics versus NDs. This result is corroborated by a 2016 Type 2 diabetes study by the NIH Diabetes Prevention Program, which identified the metabolites betaine and serine as associated with incident diabetes even after adjustment for clinical measures [46].

Discussion

Meta-analysis of metabolomics data has been attempted only a few times [47–49]. The studies have highlighted problems with inadequate metabolite identification and reporting that hamper the potential of meta-analyses for biomedical discovery, calling for improved standardization of metabolomics studies. This paper demonstrates how to incorporate method information into data annotation with the purpose of assessing data quality for meta-analysis. The diabetes use case presented here shows the value of maintaining a well annotated in-house repository of omics data collected from existing studies and public databases related to the disease area being studied so that the data can be re-analyzed. In the paper that describes the ST000435 study [50], the authors compared glycemic control against the ND group and did not discover the pattern discovered through MODIS. The results suggest that among the ND population there are individuals with poor and good glycemic control making it harder to find a distinctive classifier that distinguishes them from the diabetic group but that the good glycemic control group differs markedly from the poor glycemic control in the metabolism of aspartic acid, serine, glycine and histidine. Finally, the concatenation-based integration study suggests to study serine as biomarker for diabetes, and both results taken together suggest that serine could be indicative of a prediabetic state.

Figure 6. Integrated statistical analysis. (A) Conceptual data integration showing good separation between poor and good glycemic controls using sparse partial least squares discriminant analysis with a volcano plot showing aspartic acid and phosphoethanol as significant markers, and (B) concatenation-based integration analysis showing bad separation between diabetics and nondiabetics for most features, but finding serine to be a significant feature by multiple methods, including by the volcano plot shown here. ND: Nondiabetic; T1D: Type 1 diabetic.

Conclusion

Documenting the origin of biomolecular datasets and the protocols used in the experiments or analysis that generated them is important for reuse and re-analysis of data. This is especially necessary in metabolomics research where variation of measurements occurs frequently, even for experiments run on the same instrument at different times. Journals, grant-issuing institutes and the research community ask for data sharing but publicly available data are of minimal value for re-analysis if not well annotated and described. When reusing datasets from collaborators or from published studies, it is impossible to align data with confidence if the information that describes the experiment and the results is missing.

Figure 6. Integrated statistical analysis (cont.). (A) Conceptual data integration showing good separation between poor and good glycemic controls using sparse partial least squares discriminant analysis with a volcano plot showing aspartic acid and phosphoethanol as significant markers, and (B) concatenation-based integration analysis showing bad separation between diabetics and nondiabetics for most features, but finding serine to be a significant feature by multiple methods, including by the volcano plot shown here. ND: Nondiabetic; T1D: Type 1 diabetic.

The MODIS knowledge base and components constitute an innovative contribution to the field of data integration for metabolic profiling. It promotes pooling and leveraging of existing datasets for designing of new screening methods or discovery of disease biomarkers. The system provides computational methods that include a dimensional warehouse, a knowledge base and an expert system for the storage, processing, QC and integration of metabolite data across various laboratory platforms and analytical technologies. We highlight the use of the system through a meta-analysis study of publicly available diabetes metabolomics data. We show that new insights can be gained from re-analyzing data from existing studies, which in turn demands that the data be well annotated, have undergone QC procedures, be assigned a CL and be easily extracted in a format ready for analysis.

Future perspective

As researchers collect new data and publish studies, online public repositories will host more studies and support the type of analysis presented in this paper. Currently, there are not enough studies in MetaboLights and the Metabolomics Workbench that can be used to run the analysis we presented here. A recent study came to a similar conclusion when it looked at MSI reporting standards compliance of 399 public datasets from four major metabolomics data repositories [19]. We expect to see in the near future public repositories adopt standards that facilitate meta-analysis of public datasets. Moreover, the development of a system as described here that is scalable to other omics data are inevitable. The delay in its implementation is due to a current lack of standards and a lack of knowledge of how results are affected by the study design. The bioanalysis computer software of the future will perform many of the QC data processing steps on data from existing studies and redirect the processed data into statistical analysis and machine learning pipelines within a cluster-computing framework.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/bio-2018-0303

Acknowledgments

The authors would like to thank the members of the MODIS Advisory Panel for their participation in the project (I Adroulakis, A Cheema, C Clish, O Fiehn, M Loda, G Michailidis, G Patti, H Ressom, E Schymanski, L Sumner and W Dunn).

Financial & competing interests disclosure

This work was supported by a contract from the National Cancer Institute (SBIR contract 261201600065C) and this article will be deposited in NIHMS system/PMC. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Support for writing and revising the manuscript was provided to GJ Patti by NIH grant ES028365, and EL Schymanski by EC FP7 grant 603437.

Ethical conduct of research

Data relating to human or animal experimental investigations was downloaded from publicly accessible databases (Metabolomics Workbench). The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Executive summary

Aim

- Reuse, re-analysis and meta-analysis of data from published metabolomics studies requires quality control (QC) measures and verification of data compatibility.

Methods

- An ontology-linked minimum information about metabolomics experiments specification is the basis for the QC method.
- Minimum information about metabolomics experiments is then translated into a system of expert rules and a dimensional warehouse model with validation rules.

Results & discussion

- The method was used to develop components for a metabolomics QC software system.
- The method was applied to published diabetes studies and the meta-analysis revealed putative biomarkers for diabetes or poor glycemic control.

Conclusion

- As researchers collect new data and publish studies, online public repositories will host more studies and require the type of QC-oriented data processing presented in this paper.

References

1. Zhou B, Xiao JF, Tuli L, Ressom HW. LC-MS-based metabolomics. *Mol. Biosyst.* 8(2), 470–481 (2012).
2. National Cancer Institute NIH/NCI 341: development of metabolomics data integration methods and software. NCI: SBIR & STTR (2015). <https://sbir.cancer.gov/funding/contracts/nihnci341>
3. McDermott JE, Wang J, Mitchell H *et al.* Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin. Med. Diagn.* 7(1), 37–51 (2013).
4. Lin D, Zhang J, Li J, He H, Deng HW, Wang YP. Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. *Front. Cell Dev. Biol.* 2, 62 (2014).
5. Aretz I, Meierhofer D. Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *Int. J. Mol. Sci.* 17(5), 632 (2016).
6. Yin P, Lehmann R, Xu G. Effects of pre-analytical processes on blood samples used in metabolomics studies. *Anal. Bioanal. Chem.* 407(17), 4879–4892 (2015).
7. Zanetti KA, Mette E, Maruvada P *et al.* The future of metabolomic profiling in population-based research: opportunities and challenges. *J. Anal. Bioanal. Tech.* 5(203), 2 (2014).
8. Guasch-Ferré M, Hruby A, Toledo E *et al.* Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. *Diabetes Care* 39(5), 833–846 (2016).
9. Goveia J, Pircher A, Conradi LC *et al.* Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. *EMBO Mol. Med.* 8(10), 1134–1142 (2016).

10. Reznik E, Luna A, Aksoy BA, Liu EM, Sander C. The landscape of metabolic changes across tumor types. *Cell Syst.* 6(3), 301–313 (2017).
11. National Cancer Institute. Think tank on metabolomics and prospective cohorts: how to leverage resources (2014). <https://epi.grants.cancer.gov/comets/think-tank.html>
12. Schymanski EL, Singer HP, Slobodnik J *et al.* Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.* 407(21), 6237–6255 (2015).
13. Cheema AK, Asara JM, Wang Y, Neubert TA, Tolstikov V, Turck CW. The ABRF metabolomics research group 2013 study: investigation of spiked compound differences in a human plasma matrix. *J. Biomol. Tech.* 26(3), 83–89 (2015).
14. Martin JC, Maillot M, Mazerolles G *et al.* Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. *Metabolomics* 11(4), 807–821 (2015).
15. Salek RM, Steinbeck C, Viant MR, Goodacre R, Dunn WB. The role of reporting standards for metabolite annotation and identification in metabolomic studies. *Gigascience* 2(1), 13 (2013).
16. Dunn WB, Wilson ID, Nicholls AW, Broadhurst D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* 4(18), 2249–2264 (2012).
17. Creek DJ, Dunn WB, Fiehn O *et al.* Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* 10(3), 350–353 (2014).
18. Sumner LW, Amberg A, Barrett D *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 3(3), 211–221 (2007).
19. Spicer RA, Salek R, Steinbeck C. Compliance with minimum information guidelines in public metabolomics repositories. *Sci. Data* 4, 170137 (2017).
20. Sumner LW, Lei Z, Nikolau BJ, Saito K, Roessner U, Trengove R. Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* 10(6), 1047–1049 (2014).
21. Schymanski EL, Jeon J, Gulde R *et al.* Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* 48(4), 2097–2098 (2014).
22. Salek RM, Neumann S, Schober D *et al.* COordination of Standards in MetabOlogicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics* 11(6), 1587–1597 (2015).
23. World Wide Web Consortium (W3C). SWRL: a semantic web rule language combining OWL and RuleML (2004). www.w3.org/Submission/SWRL/
24. Haug K, Salek RM, Conesa P *et al.* MetaboLights – an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41(D1), D781–D786 (2012).
25. Rocca-Serra P, Brandizi M, Maguire E *et al.* ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics* 26(18), 2354–2356 (2010).
26. Metabolomics Society, Inc. Metabolomics society: standardization in metabolomics experiments (2014). <http://metabolomicssociety.org/resources/metabolomics-standards>
27. Fiehn O, Robertson D, Griffin J *et al.* The metabolomics standards initiative (MSI). *Metabolomics* 3(3), 175–178 (2007).
28. Sansone SA, Schober D, Atherton HJ *et al.* Metabolomics standards initiative: ontology working group work in progress. *Metabolomics* 3(3), 249–256 (2007).
29. Metabolomics Workbench. Metabolomics Workbench – Home (2019). <http://metabolomicsworkbench.org/>
30. i2b2 tranSMART Foundation. i2b2: Informatics for integrating biology & the bedside i2b2 (2019). www.i2b2.org/resrcs/index.html(2019)
31. Dunn WB, Broadhurst D, Begley P *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* 6(7), 1060–1083 (2011).
32. Kirwan JA, Weber RJM, Broadhurst DI, Viant MR. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci. Data* 1, 140012 (2014).
33. Hendriks MMWB, van Eeuwijk FA, Jellema RH *et al.* Data-processing strategies for metabolomics studies. *TrAC Trends Anal. Chem.* 30(10), 1685–1698 (2011).
34. Sumner LW, Lei Z, Nikolau BJ, Saito K, Roessner U, Trengove R. Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* 10(6), 1047–1049 (2014).
35. Berners-Lee T, Hendler J, Lassila O. The semantic web. *Sci. Am.* 284(5), 29–37 (2001).
36. World Wide Web Consortium (W3C). Resource description framework (RDF) model and syntax specification (1999). www.w3.org/TR/1999/REC-rdf-syntax-19990222/
37. W3C Semantic Web Standards OWL – web ontology language (2012). www.w3.org/OWL/

38. Smith B, Ashburner M, Rosse C *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25(11), 1251 (2007).
39. Whetzel PL, Noy NF, Shah NH *et al.* BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 39(Suppl. 2), W541–W545 (2011).
40. Metabolon. Sample preparation and shipping guidelines. (2017). www.metabolon.com/application/files/7415/0293/7140/Sample-Preparation-and-Shipping-Guidelines_Feb-2017.pdf
41. van Iersel MP, Pico AR, Kelder T *et al.* The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics* 11(1), 5 (2010).
42. Chemical Translation Service. Fiehn Lab at UC Davis – simple conversion (2019). <https://cts.fiehnlab.ucdavis.edu>
43. Bandrowski A, Brinkman R, Brochhausen M *et al.* The ontology for biomedical investigations. *PLoS ONE* 11(4), e0154556 (2016).
44. BioPortal. Information artifact ontology. (2017). <https://biportal.bioontology.org/ontologies/IAO>
45. Drools. Drools – business rules management system. (2019). www.drools.org/
46. Walford GA, Ma Y, Clish C, Florez JC, Wang TJ, Gerszten RE. Metabolite profiles of diabetes incidence and intervention response in the diabetes prevention program. *Diabetes* 65(5), 1424–1433 (2016).
47. Guasch-Ferré M, Hruby A, Toledo E *et al.* Metabolomics in prediabetes and diabetes: a systematic review and meta-analysis. *Diabetes Care* 39(5), 833–846 (2016).
48. Goveia J, Pircher A, Conradi LC *et al.* Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. *EMBO Mol. Med.* 8(10), 1134–1142 (2016).
49. Patti GJ, Tautenhahn R, Siuzdak G. Meta-analysis of untargeted metabolomic data: combining results from multiple profiling experiments. *Nat. Protoc.* 7(3), 508–516 (2012).
50. Dutta T, Kudva YC, Persson X-MT *et al.* Impact of long-term poor and good glycemic control on metabolomics alterations in Type 1 diabetic people. *J. Clin. Endocrinol. Metab.* 101(3), 1023–1033 (2016).