Data Sharing, Standards and workflows in Metabolomics: Towards reproducible Metabolomics

Reza Salek PhD
Metabolism and Molecular Informatics
The European Bioinformatics Institute (EMBL-EBI)
Email: Reza.salek@ebi.ac.uk

NETTAB Rome 2016
To make efficient use of data...

- What do I mean by reproducibility
- Data Sharing;
  - data publication is insufficient
  - data “sharing” is needed for reuse, mashup, and integration of the data

- Standards;
  - standardization of terminology – MI and SOP
    - via Society or international organizations
  - standardization of data exchange format
  - clarification of rules regarding data exchange (copyright, personal information, etc...)
  - Privacy and ethic were applicable

- Creation and management of a portal website from which users access existing DB
  - Data analysis reproducibility
Reproducibility is a **cornerstone of the scientific** process: only if my colleagues can reproduce my work should they trust its veracity.

Christian Collberg & Todd Proebsting

http://reproducibility.cs.arizona.edu/
Metabolomics Standard Initiative (WG)

- 5 Workgroups
  - Biological context metadata WG
  - Chemical analysis WG
  - Data processing WG
  - Ontology WG
  - Exchange format WG

Roy Goodacre Metabolomics (2014) 10:5-7
Capturing Metadata: ISA-Tab format

Developed a user friendly way to capture standards-compliant metadata

https://github.com/ISA-tools/ISACreator
https://github.com/ISA-tools/ISACreator/wiki/API
https://github.com/ISA-tools/ISATab-Viewer

Susanna-Assunta Sansone and ISA team

P Rocca-Serra et al., ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, Bioinformatics, 2010.
ISAcreator – Using Ontologies

4 terms from 2 ontologies

- NCBITaxon - NCBI organismal classification
  - Caenorhabditis elegans (obo:NCBITaxon_6239)
  - Caenorhabditis elegans/Monacrosporium haptotylum mixed EST library (obo)
- NEWT - NEWT UniProt Taxonomy Database

Selected term. (You can also enter freetext here): NCBITaxon:Caenorhabditis elegans
MetaboLights – Current submission pipeline

Study upload
Aspera, High-Speed File

Share private prepublication studies with reviewers and other trusted parties.

MetaboLights

Users browse investigations, query and view experimental metadata, and access associated data files via the web application.
MetaboLights – Study Validation Status

A metabolomic study of urine changes in type 2 diabetes in human compared to the control group

Validations Status: 

Release date: 14-Feb-2012

Organism: Homo sapiens

Study Factors:
- Gender
- Metabolic syndrome

Study identifier: MTBL51
Total Study size: 229.97MB
Submitted by: Reza Salek

doi:10.1093/nar/gks1004
## MetaboLights – Study Validation details

Validation marked with (*) are specially approved by the MetaboLights Curators.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Status</th>
<th>Description</th>
<th>Requirement</th>
<th>Group</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Study Title</td>
<td>MANDATORY</td>
<td>STUDY</td>
<td>OK</td>
</tr>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Study Description</td>
<td>MANDATORY</td>
<td>STUDY</td>
<td>OK</td>
</tr>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Study text successfully parsed</td>
<td>OPTIONAL</td>
<td>STUDY</td>
<td>OK</td>
</tr>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Study Contact(s) have listed email</td>
<td>MANDATORY</td>
<td>CONTACT</td>
<td>OK</td>
</tr>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Sample(s)</td>
<td>MANDATORY</td>
<td>SAMPLES</td>
<td>OK</td>
</tr>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Sample Name consistency check</td>
<td>MANDATORY</td>
<td>ASSAYS</td>
<td>OK</td>
</tr>
<tr>
<td>❌</td>
<td>FAILS</td>
<td>Publication(s) associated with this Study</td>
<td>MANDATORY</td>
<td>PUBLICATION</td>
<td>Study Publication is missing</td>
</tr>
<tr>
<td>✓</td>
<td>PASSES</td>
<td>Minimal Experimental protocol</td>
<td>MANDATORY</td>
<td>PROTOCOLS</td>
<td>OK</td>
</tr>
<tr>
<td>❌</td>
<td>INCOMPLETE</td>
<td>Comprehensive Experimental protocol</td>
<td>OPTIONAL</td>
<td>PROTOCOLS</td>
<td>Protocol description is not sufficiently detailed or not all required fields are provided. Missing field(s): Chromatography, Metabolite identification, Sample collection</td>
</tr>
<tr>
<td>❌</td>
<td>FAILS</td>
<td>Sample Collection protocol</td>
<td>MANDATORY</td>
<td>PROTOCOLS</td>
<td>Sample data is provided but no 'Sample collection' protocol is described</td>
</tr>
</tbody>
</table>
Data sharing repositories

- [Metabolomics Workbench](http://www.metabolomicsworkbench.org/)
- [MetaboLights](http://ebi.ac.uk/metabolights/)
Use case for metadata: OmicsDI – Collection of omics

Omics Discovery Index

Examples: cancer, Homo sapiens, Orbitrap, Q9HAU5, Phospho, Hela

expression

OmicsDI

Omics Discovery Index

46 Q Results for search term: *:* AND instrument_platform:"Orbitrap" AND omics_type:"Metabolomics"

Show results for
- Proteomics (0)
- Metabolomics (2)
- Transcriptomics (0)
- Genomics (0)
- Multi-Omics (0)

Repository

Find your repositories
- GNPS (0)
- Metabolights (0)
- MetabolomicsWorkbench (0)

Define by
- OMICS (2)
- BioAssay (0)
- Database (0)
- Experimenter (0)
- Journal (0)
- Platform (4)
- Repository (2)
- Study (1)
- Technology (0)
- Topic (0)

Sort by:
- Accession
- Relevance
- Publication data

Page size:
- 15
- 20

Showing 1 - 15 of 46

- Metabolomics-based elucidation of active metabolic pathways in erythrocytes and HSC-derived reticulocytes
  - Human stem cell derived reticulocytes were compared with mature erythrocytes by metabolomics analysis.
  - ORGANISM(S): Homo sapiens
  - Repository: MetabolomicsWorkbench

- TC and B6 untreated plasma in lupus-prone mice lipidomics (part-I)
  - compare plasma samples from 3 month old lupus-prone (TC) and control (B6) mice
  - ORGANISM(S): Mus musculus
  - Repository: MetabolomicsWorkbench

- Untargeted metabolomic analysis of the small intestinal content of malnourished mice
  - A total of 8 samples from 6 week old, female C57BL/6 mice, treated for 3 weeks with a malnourished diet or a control-fed isocaloric diet.
  - Samples were taken from the small intestinal fecal content at the terminus of the ileum.

Leading to data discovery

Omics Discovery Index

Dataset Information

Metabolomics-based elucidation of active metabolic pathways in erythrocytes and HSC-derived reticulocytes

**ABSTRACT**: Human stem cell derived reticulocytes were compared with mature erythrocytes by metabolomics analysis.

**DATA PROTOCOL**: HILIC positive ion mode

**INSTRUMENT(S)**: Orbitrap

**ORGANISM(S)**: Homo sapiens

**TISSUE(S)**: Blood

ST000403 | MetabolomicsWorkbench

Similar Datasets

- Hs_GSCs
  2014-12-31 | PeptideAtlas

- Test Metabolomics set
  2014-05-021 | E-TABM-289 | ArrayExpress

- Metabolomic profiling of twenty metabolites from human tissues in six studies
  2014-05-02 | E-TABM-290 | ArrayExpress

- Quantitative Proteomics Reveals Metabolic Differences in Homing and Non-Homing Glioma Stem Cell Xenografts and Stromal Cells
  2016-02-17 | PXD001775 | PRIDE
DATA FORMATS
Why different formats?

COordination of Standards in MetabOlomicS - FP7
**Data Standards, Exchange: What is XML?**

- XML stands for **E Xtensible Markup Language**
- XML is a markup language much like HTML
- XML was designed to carry data, not to display data
- XML is designed to be self-descriptive

**NMR analysis**

All spectra were recorded on a

```xml
<Varian NMR Instrument> Varian VNMRS 600 NMR Spectrometer </Varian NMR Instrument>
```

operating at a proton NMR frequency of

```xml
<Irradiation frequency>599.83 <Megahertz>MHz</Megahertz> </Irradiation frequency>
```

using a `<cryoprobe>5 mm inverse detection cryoprobe</cryoprobe>`.  

```xml
<acquisition nucleus>1H</acquisition nucleus> NMR spectra were recorded [...].
```
Data exchange standards in MS

- Metadata: ISAtab
- Raw data: mzML
- Metabolite Identification: mzTab
- Metabolite Quantification: mzTab

Slide from: Steffen Neumann (IPB-Halle), Proteomics and HUPO-PSI community
Generating ISA-Tab metadata files from metabolomics XML data

https://github.com/ISA-tools/mzml2isa
Tools the way forward!

ISA API
The open source ISA metadata tracking tools help to manage an increasingly diverse set of life science, environmental and biomedical experiments that employing one or a combination of technologies.

Built around the "Investigation" (the project context), "Study" (a unit of research) and "Assay" (analytical measurement) general-purpose Tabular format, the ISA tools helps you to provide rich description of the experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships) so that the resulting data and discoveries are reproducible and reusable.

To find out more about ISA, see www.isa-tools.org
To find out who’s using ISA and about the ISA development and user community, see www.isacommons.org

The ISA API aims to provide you, the developer, with a set of tools to help you easily and quickly build your own ISA objects, validate, and convert between serializations of ISA-formatted datasets and other formats/schemas (e.g. SRA schemas). The ISA API is published on PyPI as the isatools package.

*Note that the current version is a very early stage release (v0.1)*

Authors: The ISA team.
License: This code is licensed under the CPAL License.
Repository: https://github.com/ISA-tools/isa-api
ISA team email: https://groups.google.com/forum/#!topic/isatools
ISA discussion group: https://groups.google.com/forum/#!topic/isatools
Github issue tracker: https://github.com/ISA-tools/isa-api/issues

MZmine

3
Current way and ideal

- Planning of Study, experiment design, samples selection and treatments
- Experimental realisation
- Perform assays
- Metabolomics raw data
- Data Processing
- Raw Data conversion
- Transfer Meta Data from lab book
- ISA-Tab: i_Investigation, s_Samples, a_Assays, m_Metabolites

Wetlab

- Planning of Study, experiment design, samples selection and treatments
- ISA-Tab: i_Investigation, s_Samples, a_Assays, m_Metabolites
- Experimental realisation
- Perform assays
- <mzML> or <nmrML>

Drylab

- Metabolomics raw data
- ISA-Tab: i_Investigation, s_Samples, a_Assays, m_Metabolites
- Raw Data conversion
- <mzML> or <nmrML>
- Data analysis
- Statistics
- Network analysis

Data re-use

- MetaboLights
- Data Processing
- Data analysis
- Statistics
- Network analysis

```
Data standards can boost metabolomics research, and if there is a will, there is a way
```
DATA ANALYSIS

THE MISSING LINK
What software we have – which one should I choose?

<table>
<thead>
<tr>
<th>Type of Tool</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>LC/MS</td>
<td>93</td>
</tr>
<tr>
<td>GC/MS</td>
<td>32</td>
</tr>
<tr>
<td>NMR</td>
<td>24</td>
</tr>
<tr>
<td>Metabolite Identification</td>
<td>29</td>
</tr>
<tr>
<td>Lipid Identification</td>
<td>6</td>
</tr>
<tr>
<td>Lipidomics - Other</td>
<td>2</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>13</td>
</tr>
<tr>
<td>Pathway Analysis</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>208</td>
</tr>
</tbody>
</table>

Rachel Spicer
PhenoMeNal - Goal

Data Producer

Data container

Tool maker

Packaged tool

Infrastructure provider

Compute Infrastructure

PhenoMeNal

VRE Portal

PhenoMeNal; http://phenomenal-h2020.eu/
Key objectives

- Understand the computational needs of the Metabolomics Community.
- Integrate and scale **existing Open Source** tools into a well-tested e-infrastructure.
Major revolution
Same in software

- Developer’s
- PI’s
- Cluster
- Cloud
- Collaborator’s

PhenoMeNa1; EMBL-EBI
VRE Portal

- Three usability rounds
- 80% functionality running.
- Public instance access.
- App Library, hooked to EGI AppDB.
- Documentation.

http://portal.phenomenal-h2020.eu/
Why workflow?

View, share, edit, rerun workflow
Birmingham galaxy–m metabolomics workflow

- Many tools
- Many languages
- Complex to learn
- Many parameters
- Complex to report

Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data

Robert L. Davidson†, Ralf J. M. Weber†, Haoyu Liu, Archana Sharma-Oates and Mark R. Viant ‡

† Contributed equally

Received: 5 March 2015 | Accepted: 6 February 2016 | Published: 23 February 2016
MassCascade, Tomato samples
LC-MS positive ion mode
mzML:
- 3 pooled samples
- 3 WT QC samples

https://bitbucket.org/sbeisken/masscascadeknime/wiki/ExampleWorkflows
MetaboLights – The team

Kenneth Haug
Kalai Jayaseelan
Venkata Chandrasekhar
Jose Ramon Macias Gonzalez
Christoph Steinbeck

Reza Salek
Mark Williams
Keeva Cochrane
Xuefei Li (MRC)
Jules Griffin (UC/MRC)

Previous: Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, Pablo Conesa
EBI PhenoMeNal – The team

Kenneth Haug

Reza Salek

Pablo Moreno

Sijin He

Christoph Steinbeck

Namrata Kale
COSMOS consortium
PhenoMeNal consortium
Funding and Collaborators

- BBSRC: bioscience for the future
- EMBL-EBI
- isa software suite
- UNIVERSITY OF CAMBRIDGE
- MRC: Human Nutrition Research
- SEVENTH FRAMEWORK PROGRAMME
- European Union flag