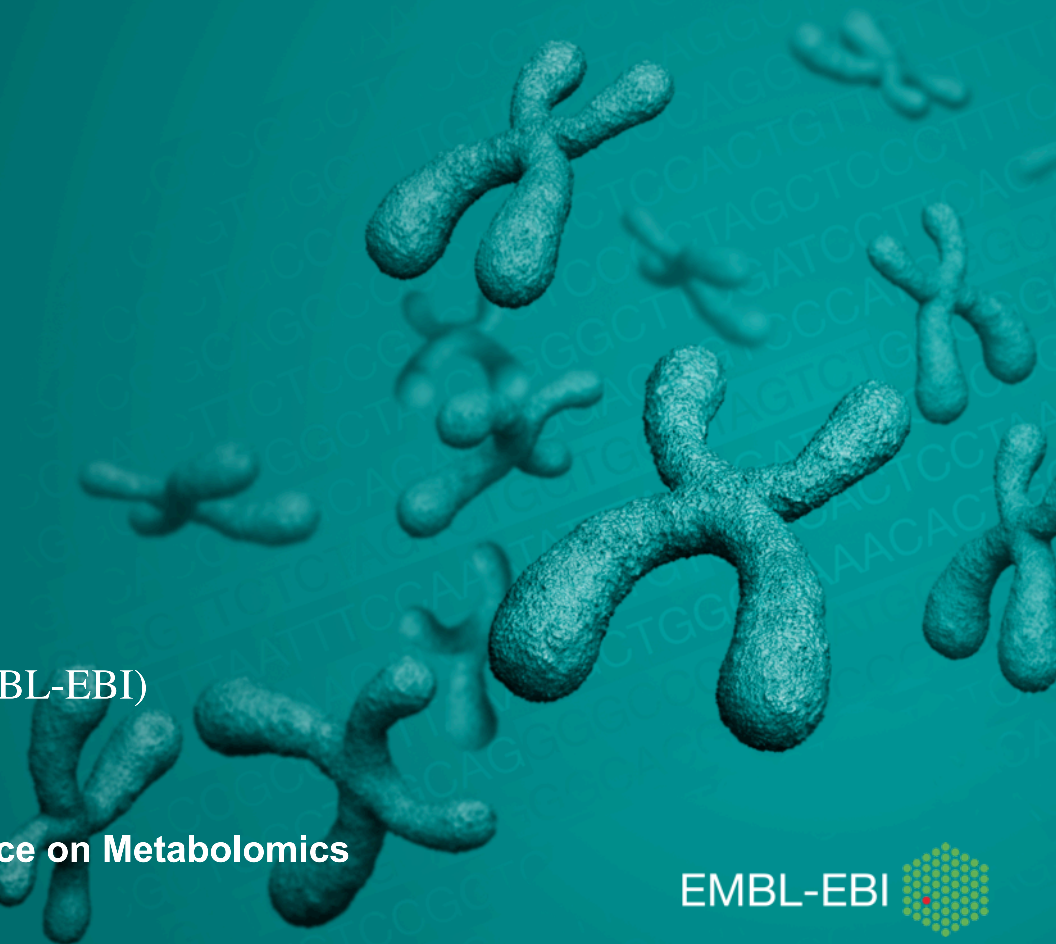
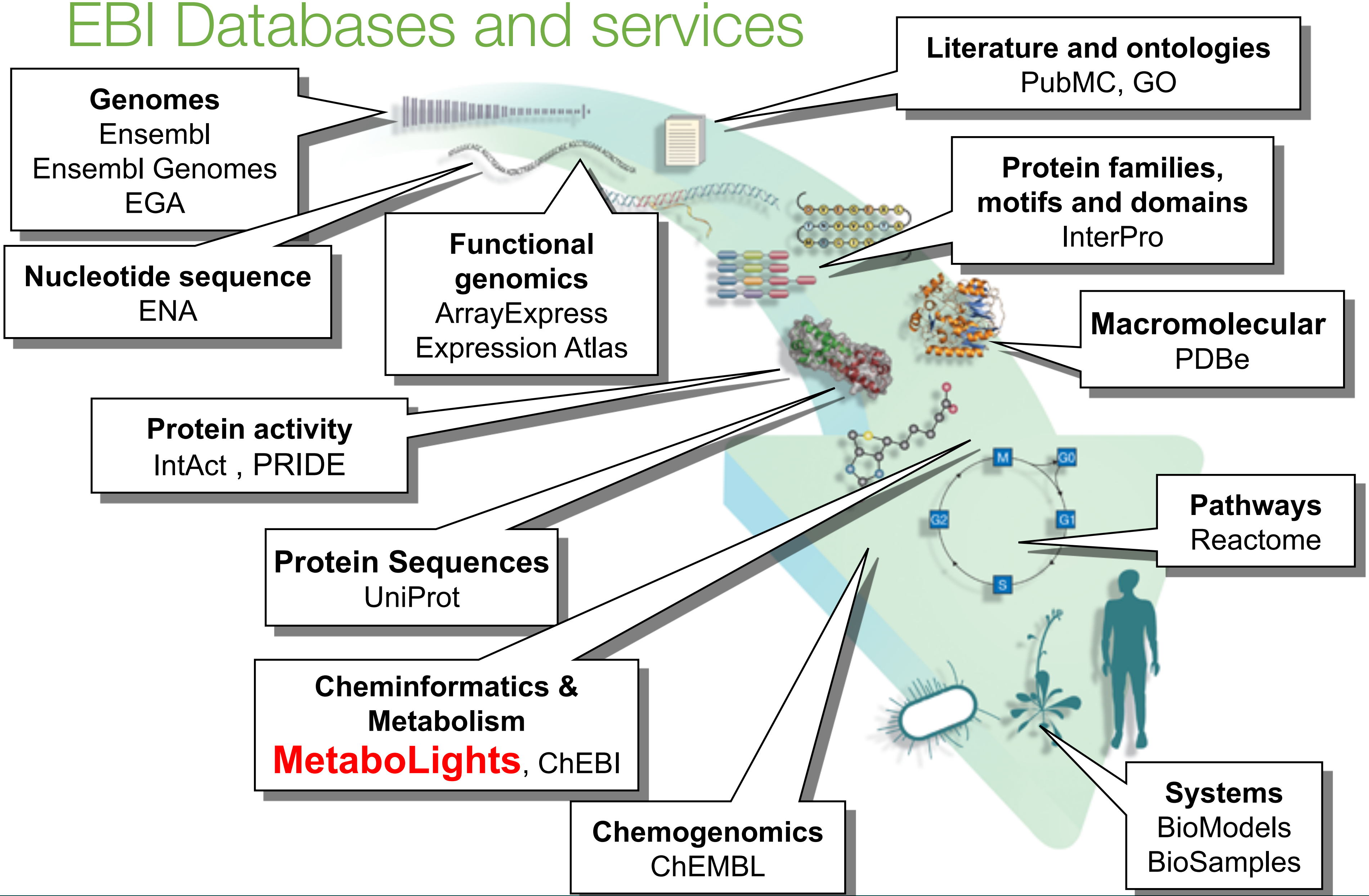
Towards standard, accessible and reproducible Metabolomics



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

The 1st International Electronic Conference on Metabolomics

EBI Databases and services



Literature and ontologies
PubMC, GO

Genomes
Ensembl
Ensembl Genomes
EGA

Protein families, motifs and domains
InterPro

Nucleotide sequence
ENA

Functional genomics
ArrayExpress
Expression Atlas

Macromolecular
PDBe

Protein activity
IntAct , PRIDE

Pathways
Reactome

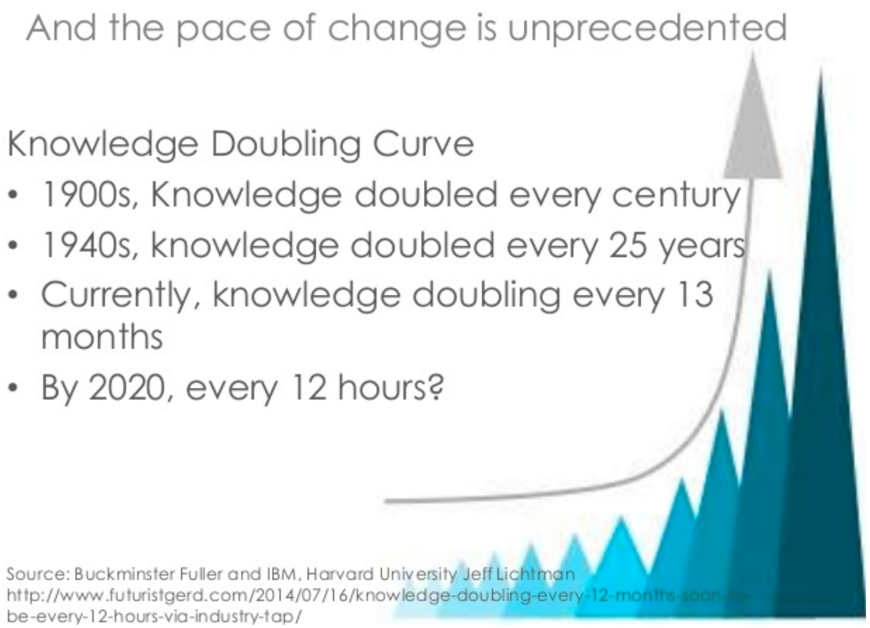
Protein Sequences
UniProt

Cheminformatics & Metabolism
MetaboLights, ChEBI

Chemogenomics
ChEMBL

Systems
BioModels
BioSamples

Is data growth, FAIR?

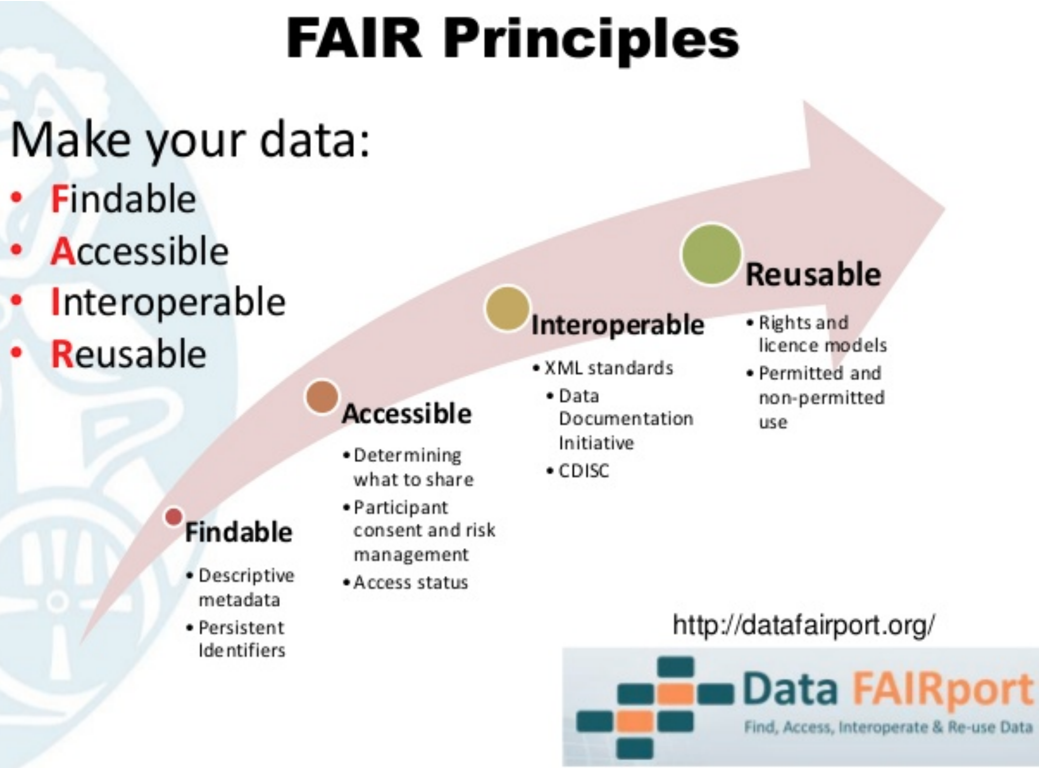


And the pace of change is unprecedented

Knowledge Doubling Curve

- 1900s, Knowledge doubled every century
- 1940s, knowledge doubled every 25 years
- Currently, knowledge doubling every 13 months
- By 2020, every 12 hours?

Source: Buckminster Fuller and IBM, Harvard University Jeff Lichtman
http://www.futuristgerd.com/2014/07/16/knowledge-doubling-every-12-months-soon-to-be-every-12-hours-via-industry-tap/



FAIR Principles

Make your data:

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Findable
- Descriptive metadata
- Persistent Identifiers

Accessible
- Determining what to share
- Participant consent and risk management
- Access status

Interoperable
- XML standards
  - Data Documentation Initiative
  - CDISC

Reusable
- Rights and licence models
- Permitted and non-permitted use

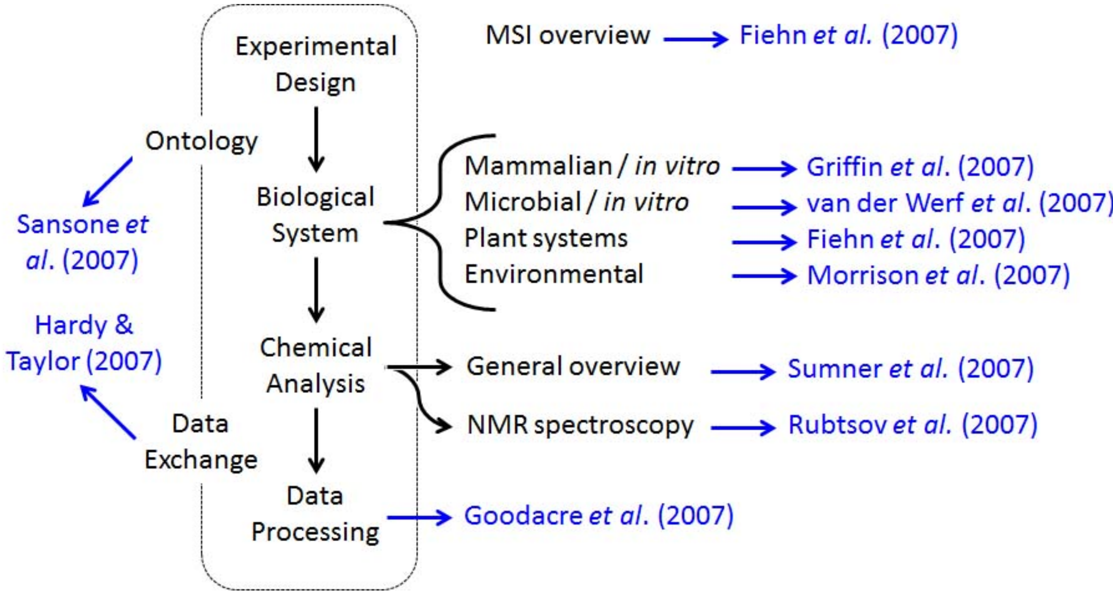
http://datafairport.org/

Data FAIRport

Find, Access, Interoperate & Re-use Data

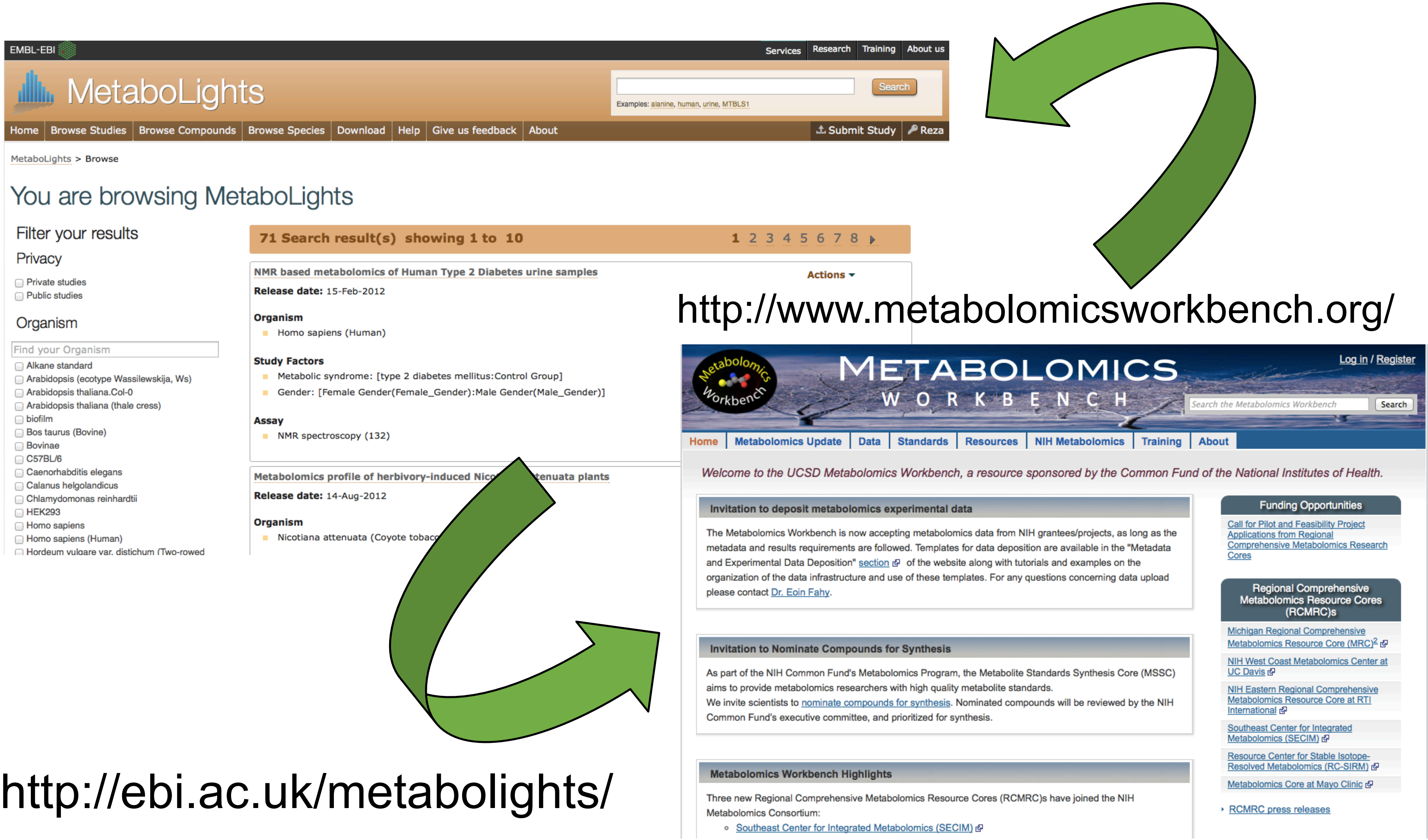
Metabolomics Standard Initiative (WG)

- Lives at http://msi-workgroups.sourceforge.net
- 5 Workgroups
  - Biological context metadata WG
  - Chemical analysis WG
  - Data processing WG
  - Ontology WG
  - Exchange format WG



Roy Goodacre Metabolomics (2014) 10:5-7

Data sharing repositories



Examples: alanine, human, urine, MTBLS1

MetaboLights > Browse

You are browsing MetaboLights

Filter your results

Privacy

☐ Private studies
☐ Public studies

Organism

Find your Organism

☐ Alkane standard
☐ Arabidopsis (ecotype Wassilewskija, Ws)
☐ Arabidopsis thaliana.Col-0
☐ Arabidopsis thaliana (thale cress)
☐ biofilm
☐ Bos taurus (Bovine)
☐ Bovinae
☐ C57BL/6
☐ Caenorhabditis elegans
☐ Calanus helgolandicus
☐ Chlamydomonas reinhardtii
☐ HEK293
☐ Homo sapiens
☐ Homo sapiens (Human)
☐ Hordeum vulgare var. distichum (Two-rowed

71 Search result(s) showing 1 to 10 1 2 3 4 5 6 7 8

NMR based metabolomics of Human Type 2 Diabetes urine samples Actions

Release date: 15-Feb-2012

Organism

- Homo sapiens (Human)

Study Factors

- Metabolic syndrome: [type 2 diabetes mellitus:Control Group]
- Gender: [Female Gender(Female_Gender):Male Gender(Male_Gender)]

Assay

- NMR spectroscopy (132)

Metabolomics profile of herbivory-induced Nico **tenuata plants**

Release date: 14-Aug-2012

Organism

- Nicotiana attenuata (Coyote tobac

http://www.metabolomicsworkbench.org/

Welcome to the UCSD Metabolomics Workbench, a resource sponsored by the Common Fund of the National Institutes of Health.

Invitation to deposit metabolomics experimental data

The Metabolomics Workbench is now accepting metabolomics data from NIH grantees/projects, as long as the metadata and results requirements are followed. Templates for data deposition are available in the "Metadata and Experimental Data Deposition" section of the website along with tutorials and examples on the organization of the data infrastructure and use of these templates. For any questions concerning data upload please contact Dr. Eoin Fahy.

Invitation to Nominate Compounds for Synthesis

As part of the NIH Common Fund's Metabolomics Program, the Metabolite Standards Synthesis Core (MSSC) aims to provide metabolomics researchers with high quality metabolite standards.
We invite scientists to nominate compounds for synthesis. Nominated compounds will be reviewed by the NIH Common Fund's executive committee, and prioritized for synthesis.

Metabolomics Workbench Highlights

Three new Regional Comprehensive Metabolomics Resource Cores (RCMRC)s have joined the NIH Metabolomics Consortium:

- Southeast Center for Integrated Metabolomics (SECIM)

Funding Opportunities

- Call for Pilot and Feasibility Project Applications from Regional Comprehensive Metabolomics Research Cores

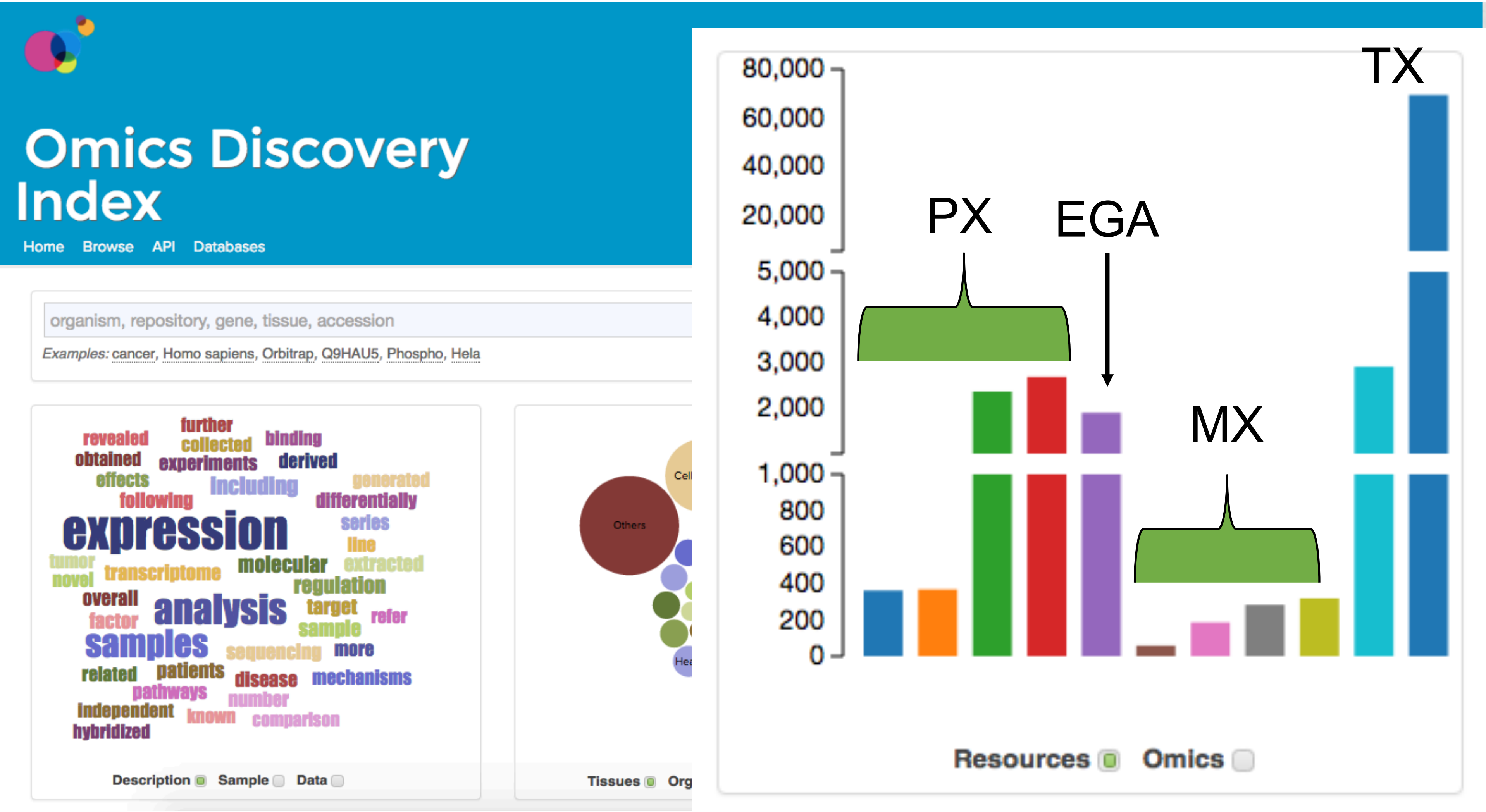
Regional Comprehensive Metabolomics Resource Cores (RCMRC)s

- Michigan Regional Comprehensive Metabolomics Resource Core (MRC)[2]
- NIH West Coast Metabolomics Center at UC Davis
- NIH Eastern Regional Comprehensive Metabolomics Resource Core at RTI International
- Southeast Center for Integrated Metabolomics (SECIM)
- Resource Center for Stable Isotope-Resolved Metabolomics (RC-SIRM)
- Metabolomics Core at Mayo Clinic

- RCMRC press releases

http://ebi.ac.uk/metabolights/

OmicsDI – Collection of omics



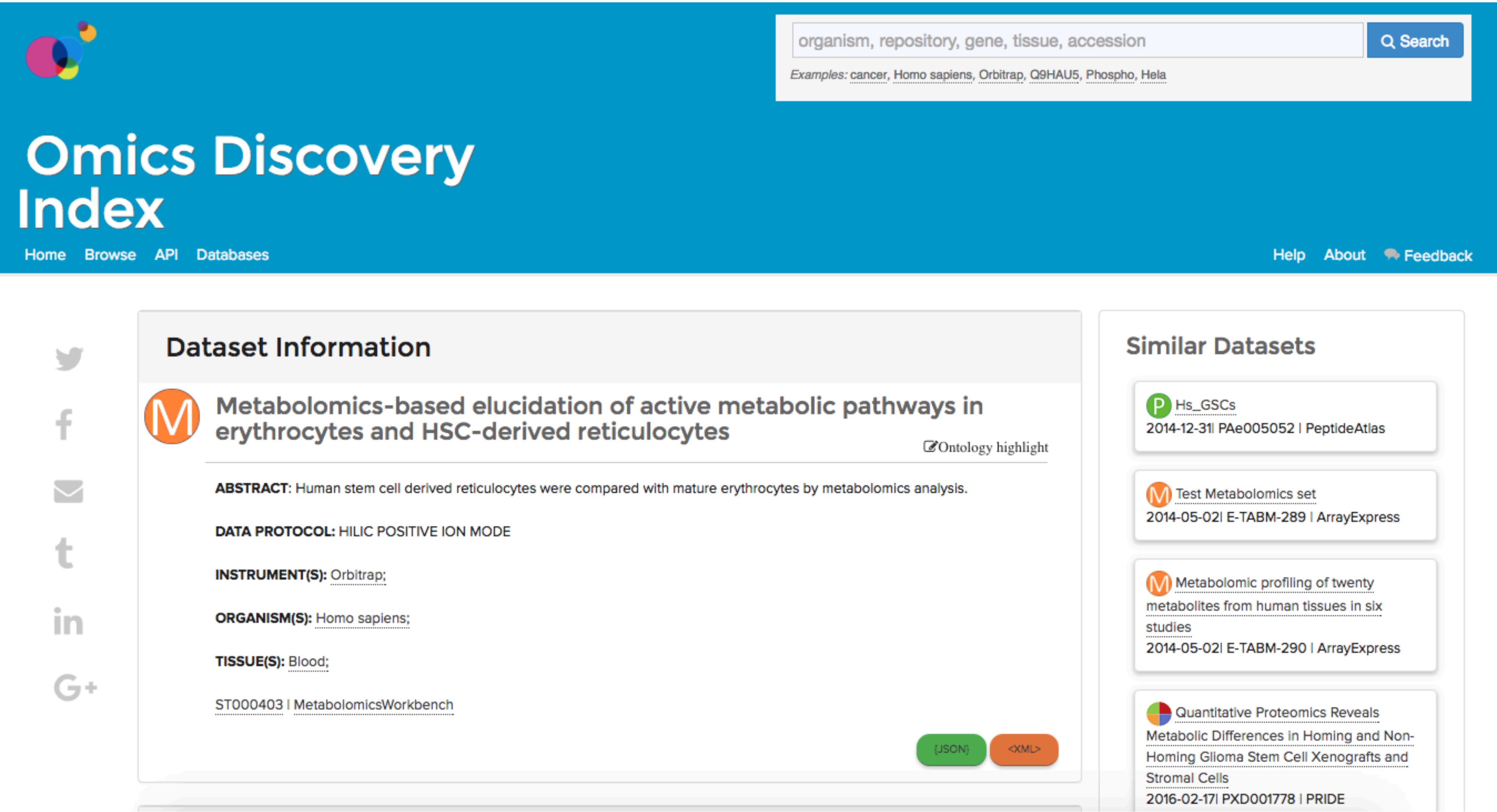
Omics Discovery Index

Home Browse API Databases

organism, repository, gene, tissue, accession

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

Leading to data discovery



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

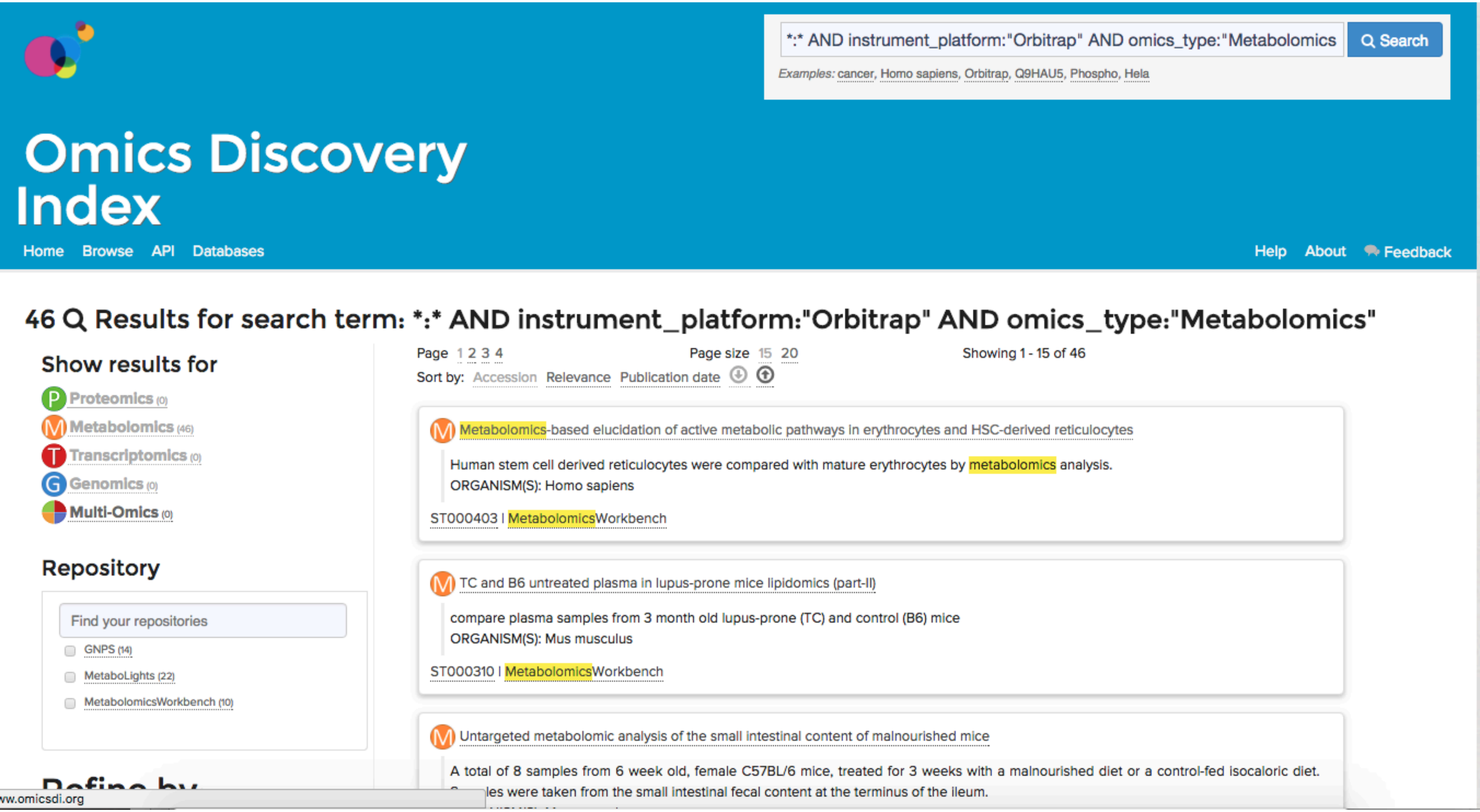
Hs_GSCs
2014-12-31| PAe005052 | PeptideAtlas

Test Metabolomics set
2014-05-02| 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| PXD001778 | PRIDE

OmicsDI



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

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

ST000403 | MetabolomicsWorkbench

TC and B6 untreated plasma in lupus-prone mice lipidomics (part-II)

compare plasma samples from 3 month old lupus-prone (TC) and control (B6) mice
ORGANISM(S): Mus musculus

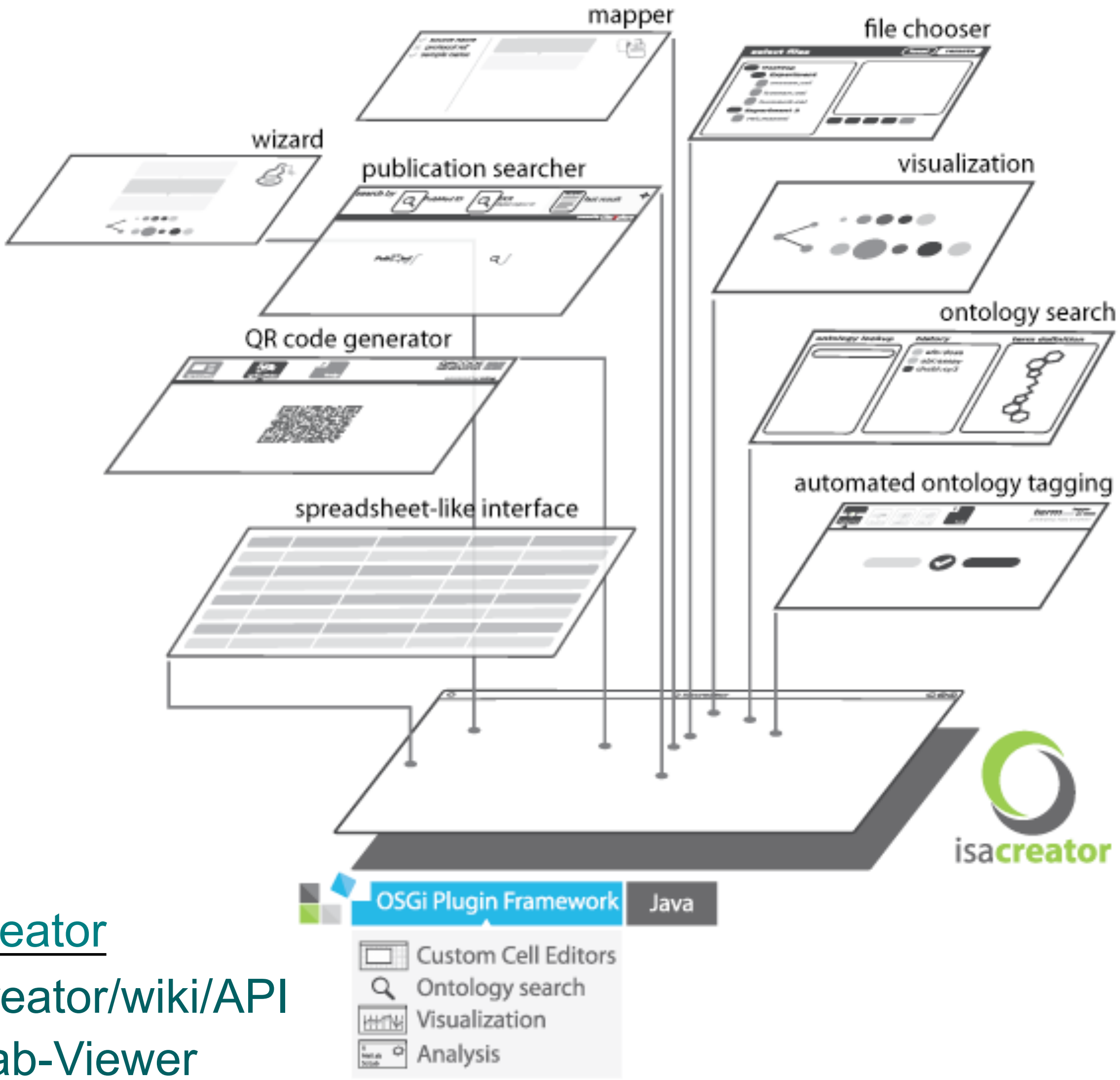
ST000310 | MetabolomicsWorkbench

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. les were taken from the small intestinal fecal content at the terminus of the ileum.

Capturing Metadata: ISA-Tab format

Developed a user friendly way to capture standards-compliant metadata

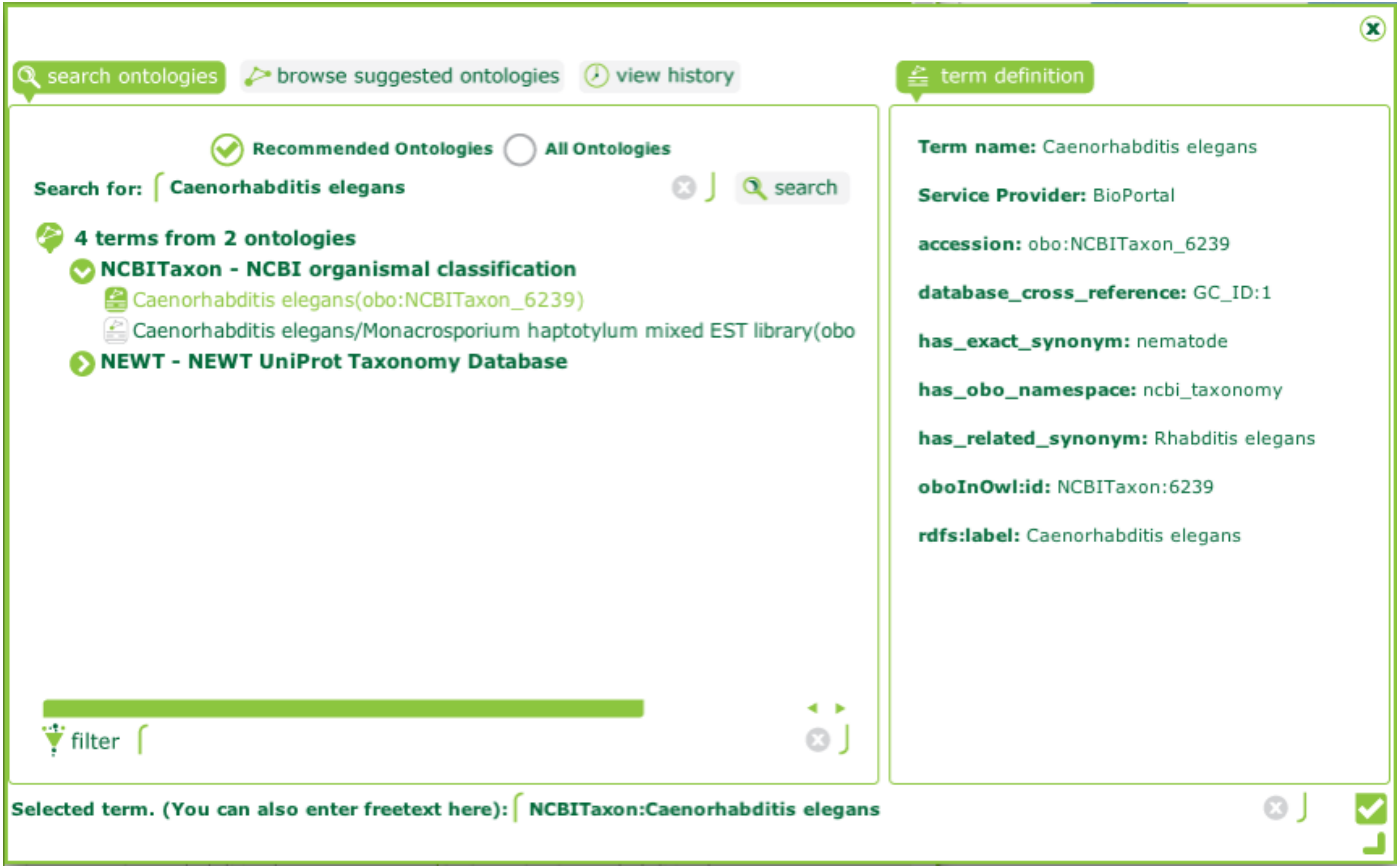


https://github.com/ISA-tools/ISAcreator

https://github.com/ISA-tools/ISAcreator/wiki/API

https://github.com/ISA-tools/ISATab-Viewer

ISAcreator – Using Ontologies



✓ **Recommended Ontologies** ○ **All Ontologies**

Search for: Caenorhabditis elegans

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

Term name: Caenorhabditis elegans

Service Provider: BioPortal

accession: obo:NCBITaxon_6239

database_cross_reference: GC_ID:1

has_exact_synonym: nematode

has_obo_namespace: ncbi_taxonomy

has_related_synonym: Rhabditis elegans

oboInOwl:id: NCBITaxon:6239

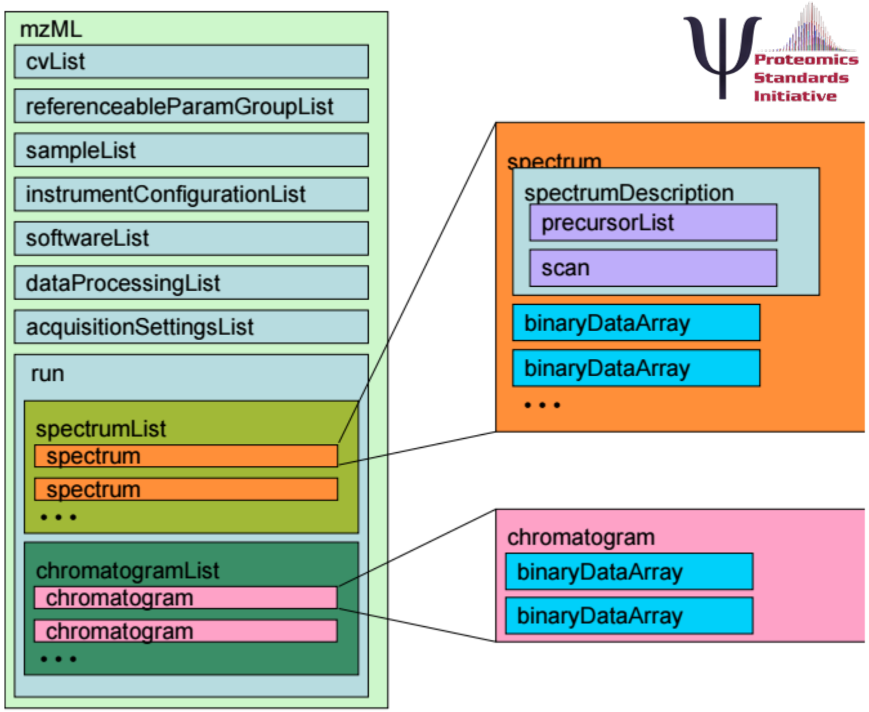
rdfs:label: Caenorhabditis elegans

Selected term. (You can also enter freetext here): NCBITaxon:Caenorhabditis elegans

Data Standards ; What is XML?

- XML stands for EXtensible 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 <Varian NMR Instrument> Varian VNMRS 600 NMR Spectrometer </Varian NMR Instrument> operating at a proton NMR frequency of <Irradiation frequency>599.83 <Megahertz>MHz</Megahertz> </Irradiation frequency> using a <cryoprobe>5 mm inverse detection cryoprobe</cryoprobe>. <acquisition nucleus>1H</acquisition nucleus> NMR spectra were recorded […].



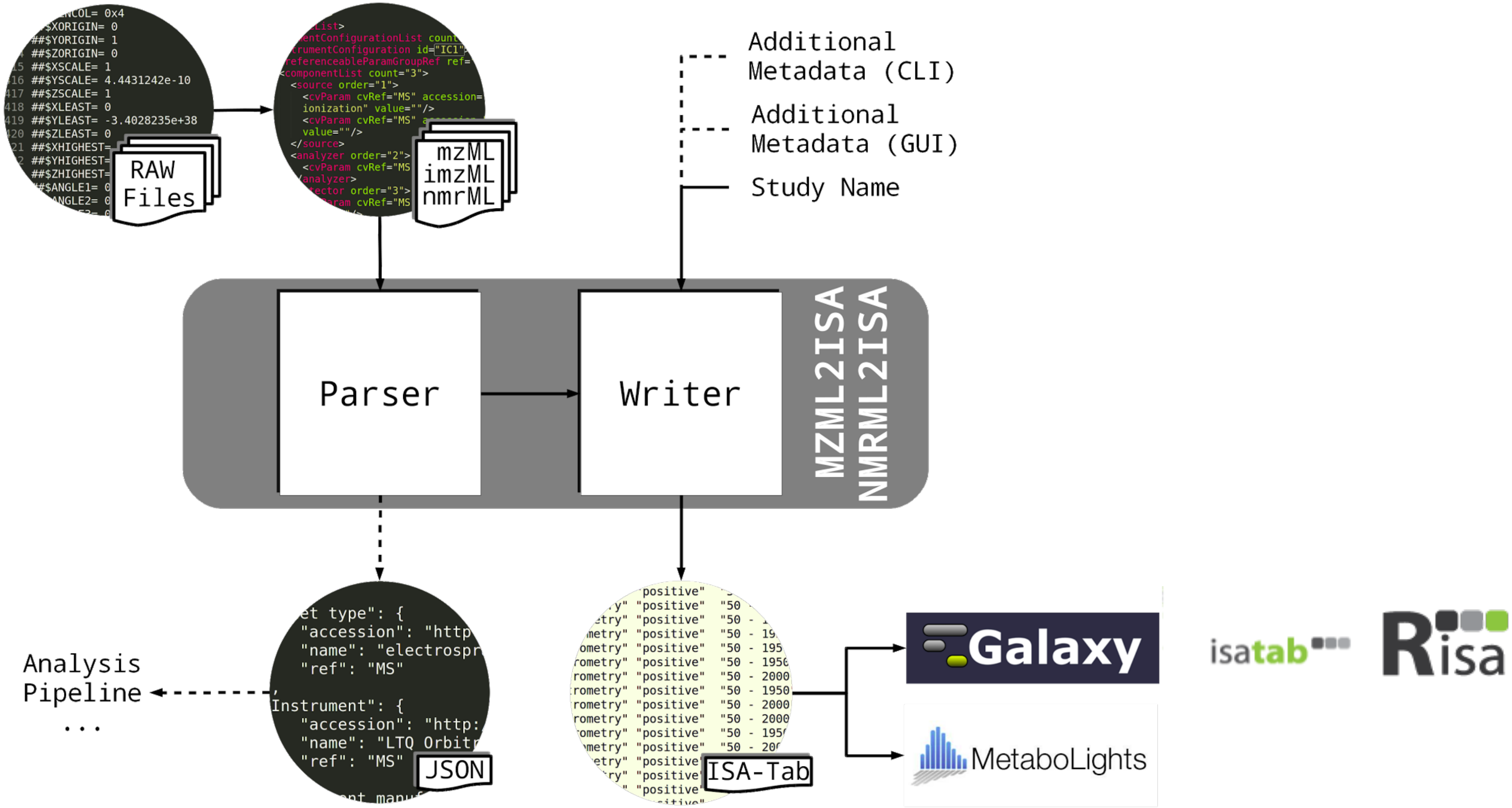
- mzML
  - cvList
  - referenceableParamGroupList
  - sampleList
  - instrumentConfigurationList
  - softwareList
  - dataProcessingList
  - acquisitionSettingsList
  - run
    - spectrumList
      - spectrum
      - spectrum
      - • • •
    - chromatogramList
      - chromatogram
      - chromatogram
      - • • •

- spectrum
  - spectrumDescription
    - precursorList
    - scan
  - binaryDataArray
  - binaryDataArray
  - • • •

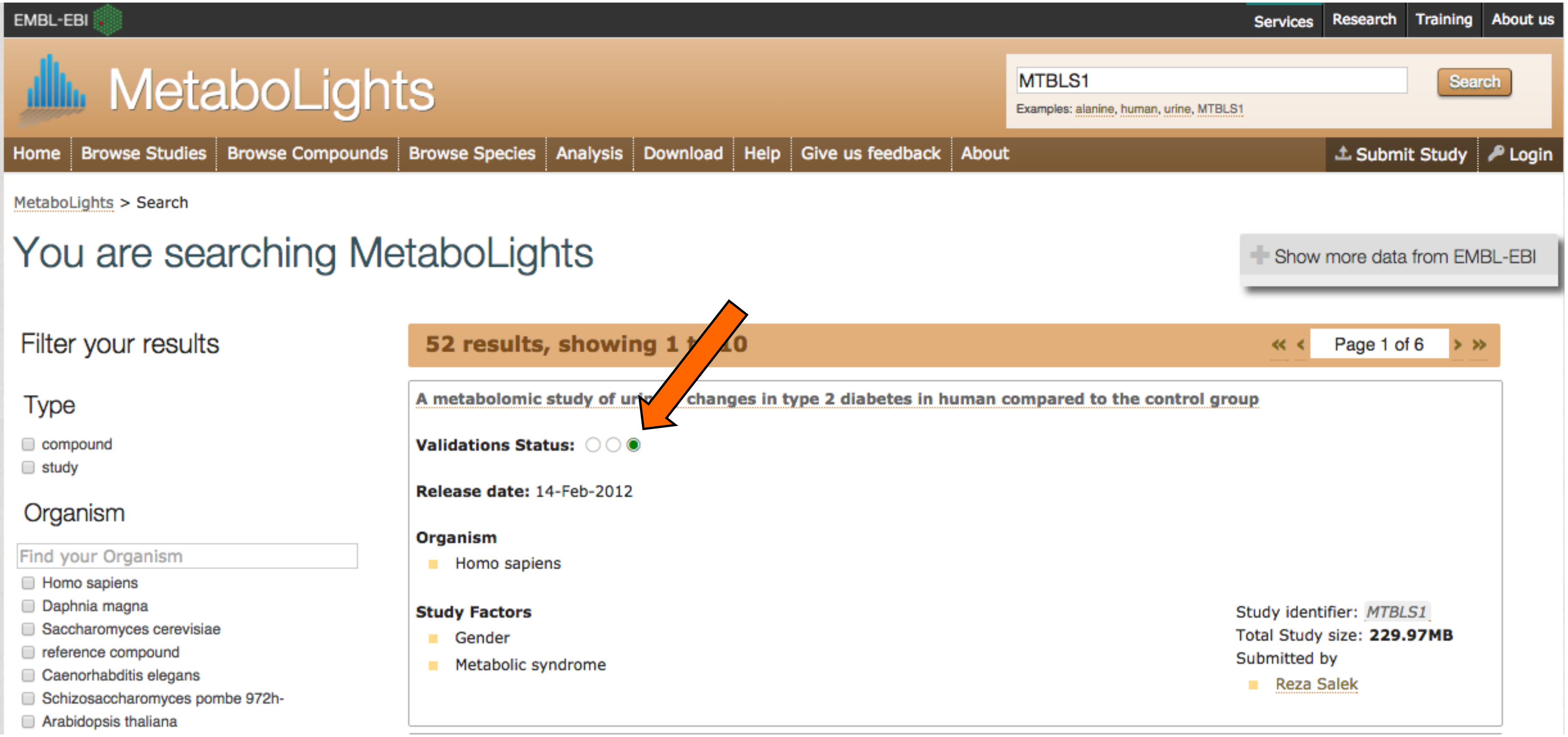
- chromatogram
  - binaryDataArray
  - binaryDataArray

Proteomics Standards Initiative

Generating ISA-Tab metadata files from metabolomics XML data



MetaboLights – Study Validation Status



MTBLS1

Examples: alanine, human, urine, MTBLS1

You are searching MetaboLights

Filter your results

Type

☐ compound
☐ study

Organism

Find your Organism

☐ Homo sapiens
☐ Daphnia magna
☐ Saccharomyces cerevisiae
☐ reference compound
☐ Caenorhabditis elegans
☐ Schizosaccharomyces pombe 972h-
☐ Arabidopsis thaliana

52 results, showing 1 t 10

A metabolomic study of uri 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

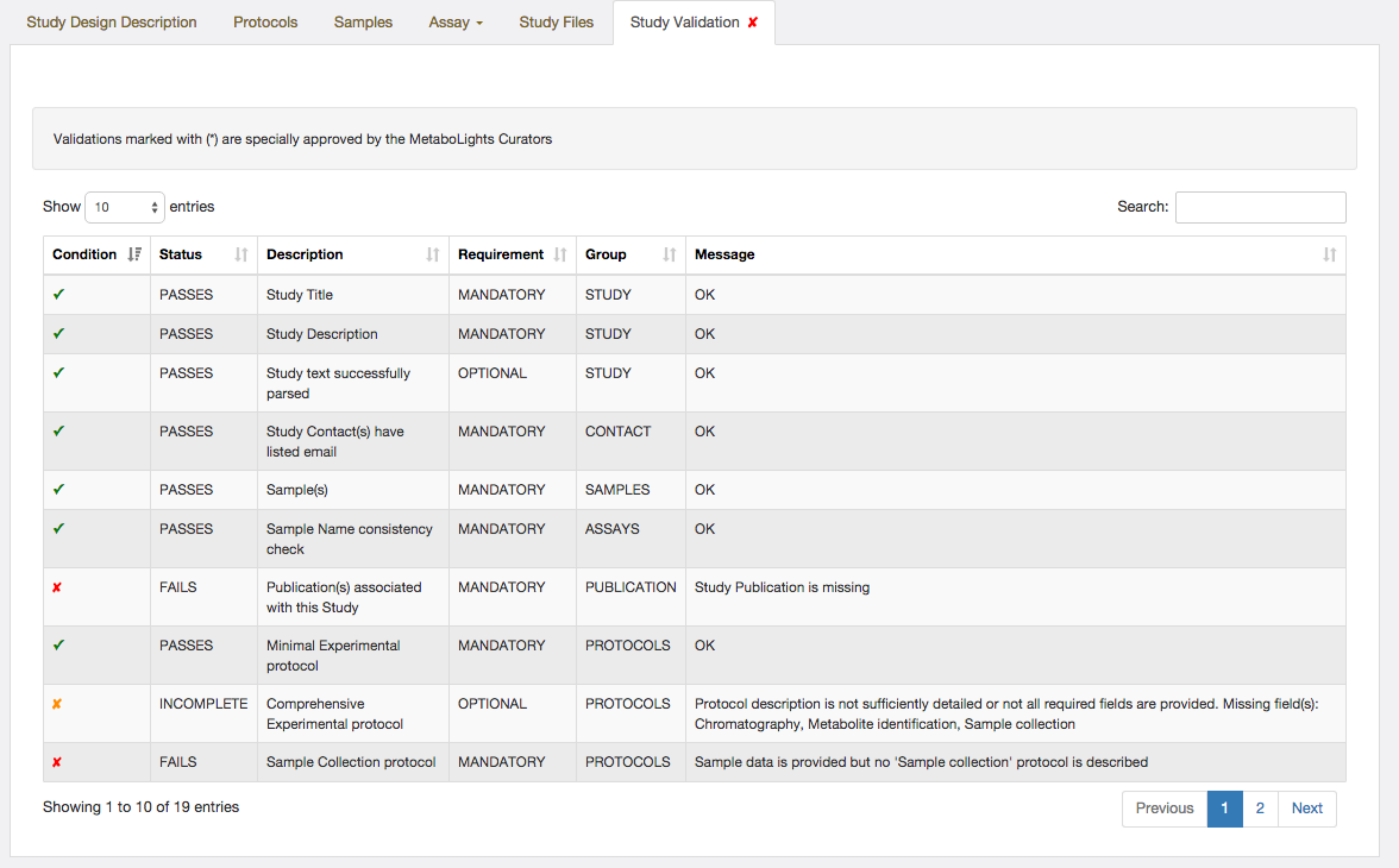
Study identifier: *MTBLS1*
Total Study size: **229.97MB**
Submitted by

- Reza Salek

MetaboLights - an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucl. Acids Res. (2012)* *[doi:10.1093/nar/gks1004*

EMBL-EBI

MetaboLights – Study Validation details

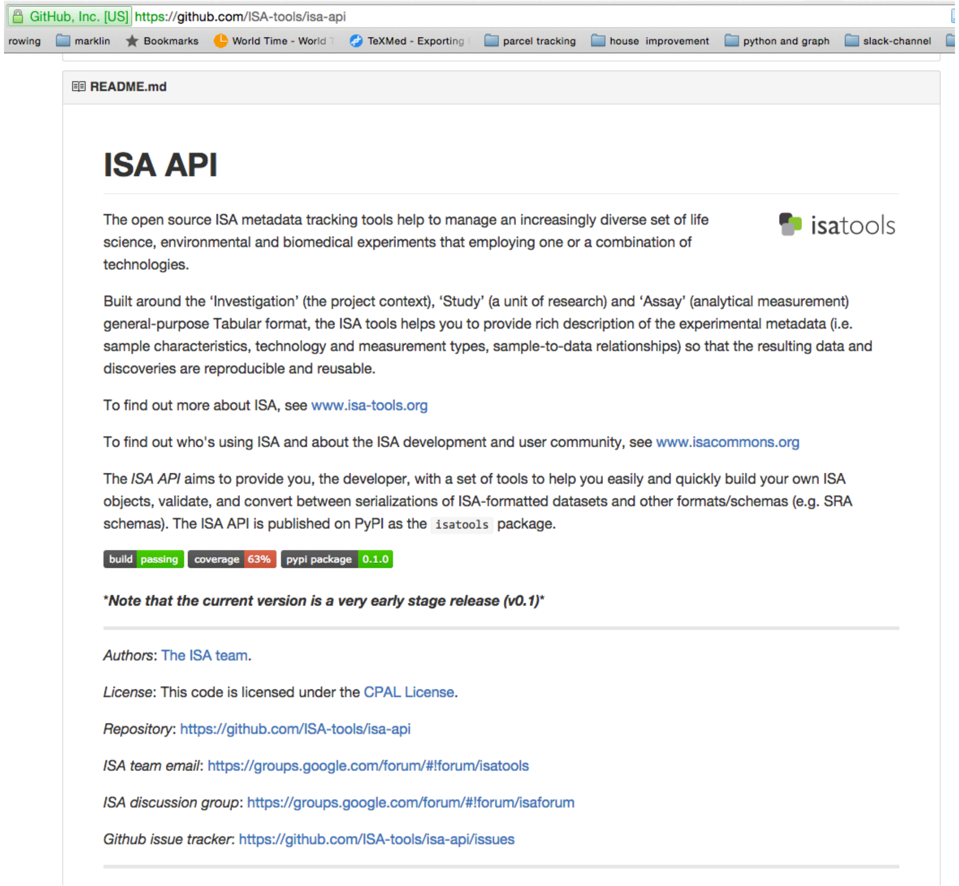


Validations marked with (*) are specially approved by the MetaboLights Curators

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

Showing 1 to 10 of 19 entries

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.

build passing | coverage 63% | pypi package 0.1.0

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.

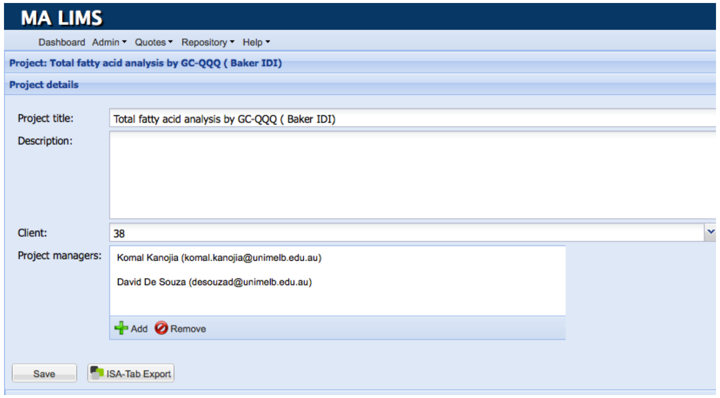
Repository: https://github.com/ISA-tools/isa-api

ISA team email: https://groups.google.com/forum/#!forum/isatools

ISA discussion group: https://groups.google.com/forum/#!forum/isaforum

Github issue tracker: https://github.com/ISA-tools/isa-api/issues

BIOCRATES
LIFE SCIENCES
The Deep Phenotyping Company



Project: Total fatty acid analysis by GC-QQQ (Baker IDI)

Project details

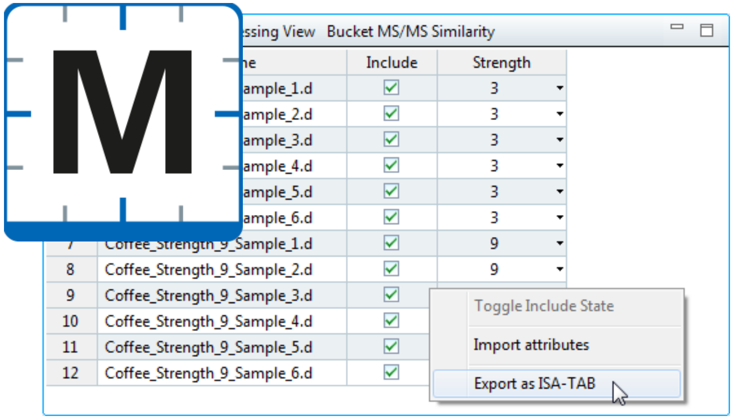
Project title: Total fatty acid analysis by GC-QQQ (Baker IDI)

Description:

Client: 38

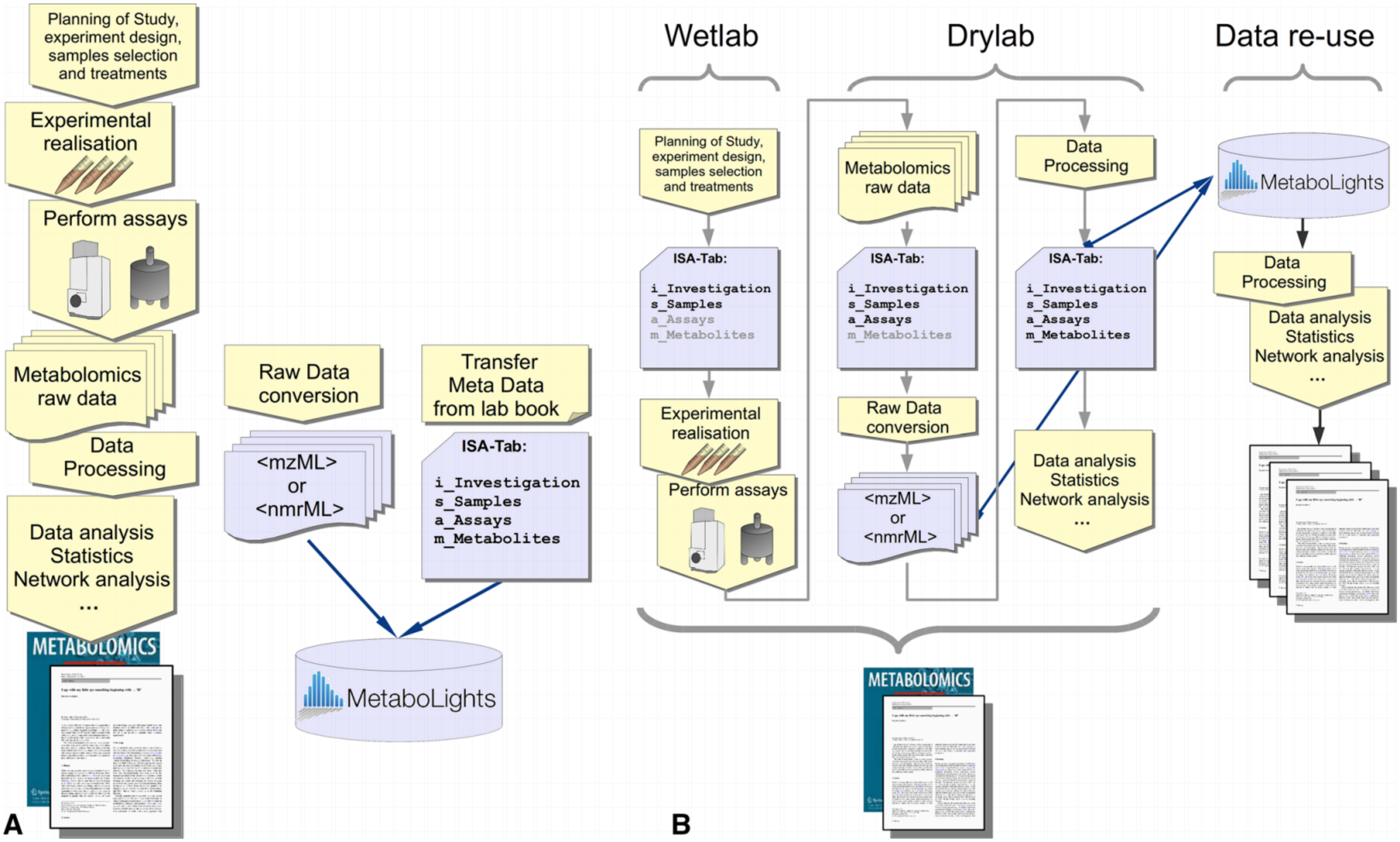
Project managers:
Komal Kanojia (komal.kanojia@unimelb.edu.au)
David De Souza (desouzad@unimelb.edu.au)

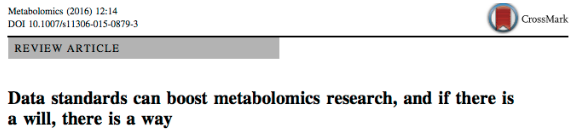
MZmine ~~2~~ 3



	ne	Include	Strength
	ample_1.d	☑	3
	ample_2.d	☑	3
	ample_3.d	☑	3
	ample_4.d	☑	3
	ample_5.d	☑	3
	ample_6.d	☑	3
7	Coffee_Strength_9_Sample_1.d	☑	9
8	Coffee_Strength_9_Sample_2.d	☑	9
9	Coffee_Strength_9_Sample_3.d	☑	
10	Coffee_Strength_9_Sample_4.d	☑	
11	Coffee_Strength_9_Sample_5.d	☑	
12	Coffee_Strength_9_Sample_6.d	☑	

Current way and ideal

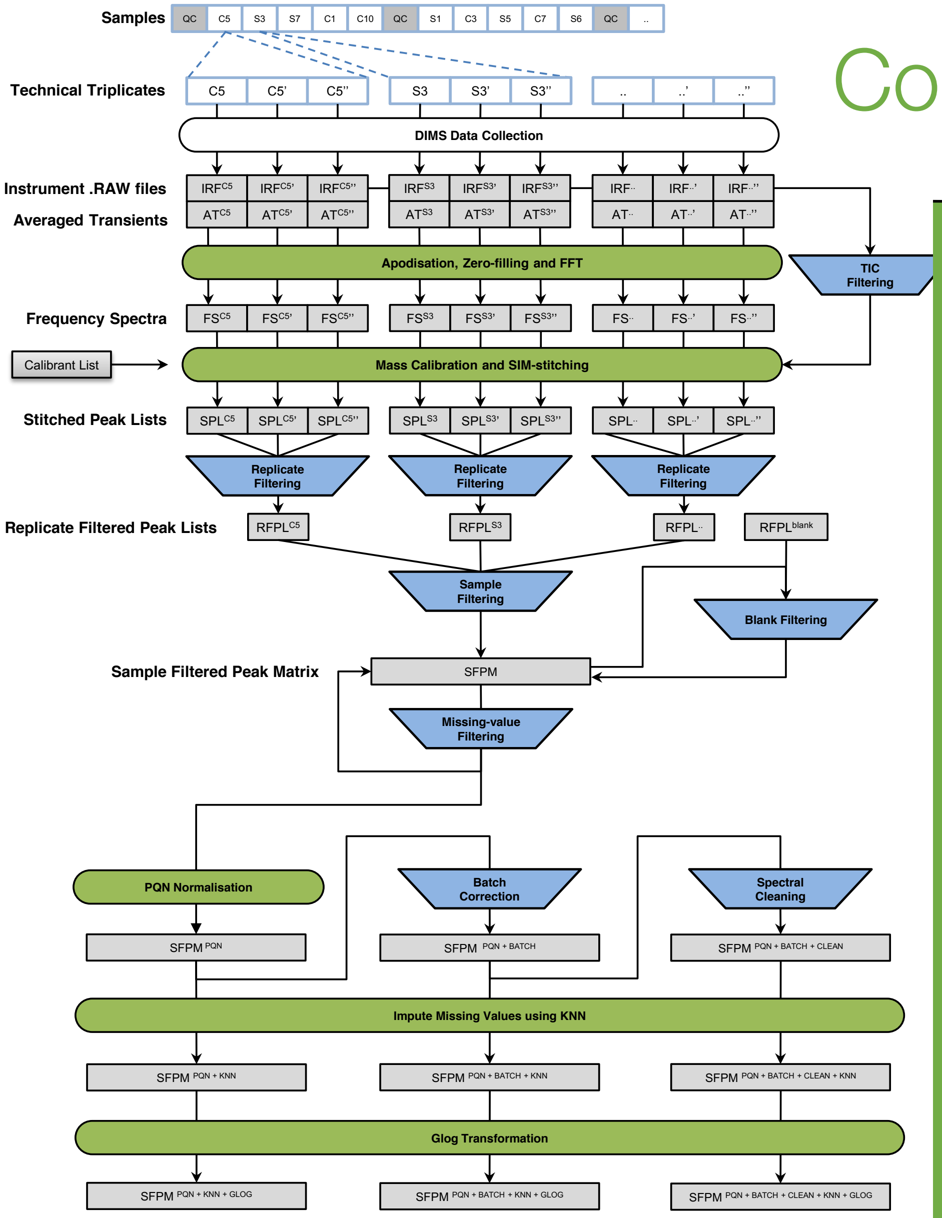


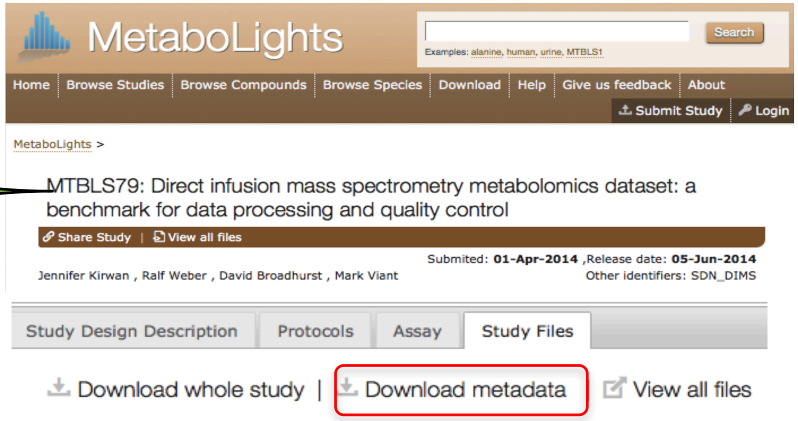

Metabolomics (2016) 12:14
DOI 10.1007/s11306-015-0879-3

REVIEW ARTICLE

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

Complex analysis pipelines





MTBLS79: Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control

Submited: 01-Apr-2014 ,Release date: 05-Jun-2014

Jennifer Kirwan , Ralf Weber , David Broadhurst , Mark Viant

Other identifiers: SDN_DIMS

PhenoMeNal - Goal

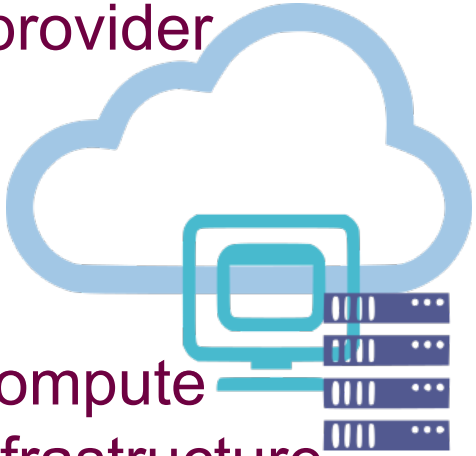
Data Producer

Data container

Tool maker

Packaged tool

Infrastructure provider



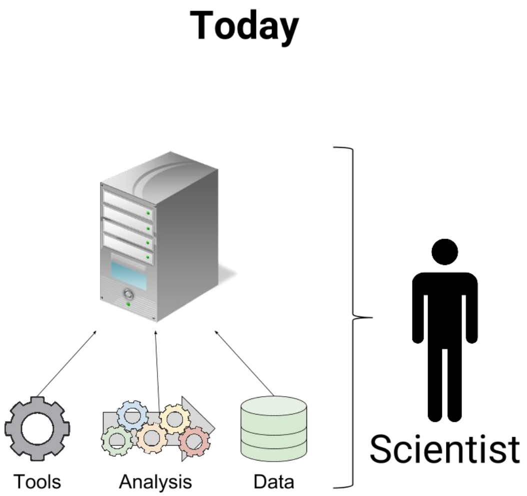
Compute Infrastructure

PhenoMeNal

VRE Portal

Key objectives

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



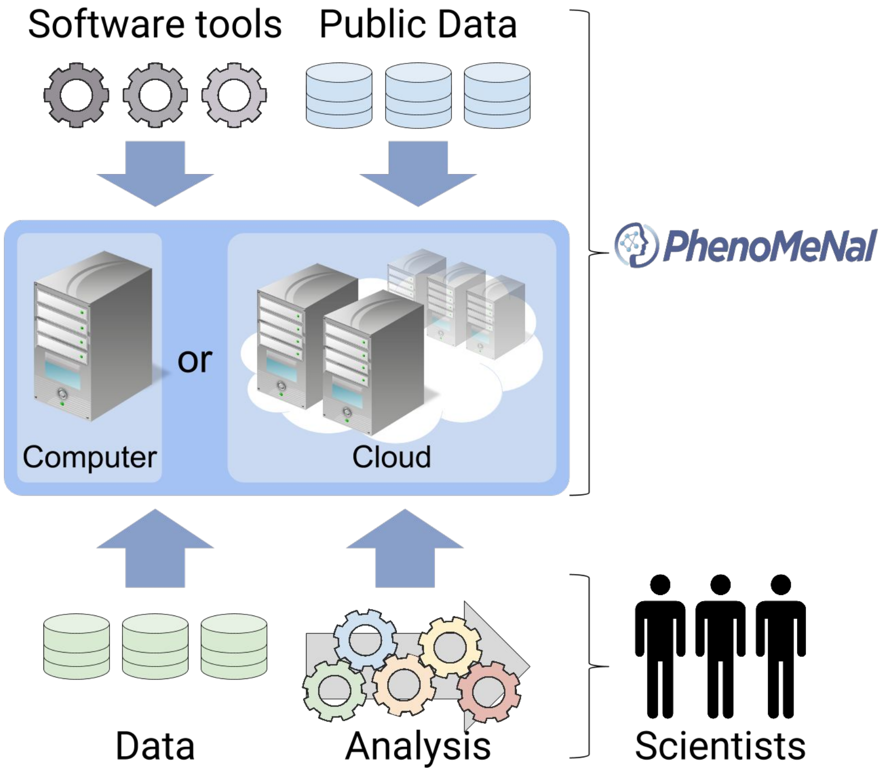
Today

Tools

Analysis

Data

Scientist



Software tools

Public Data

Computer

or

Cloud

PhenoMeNal

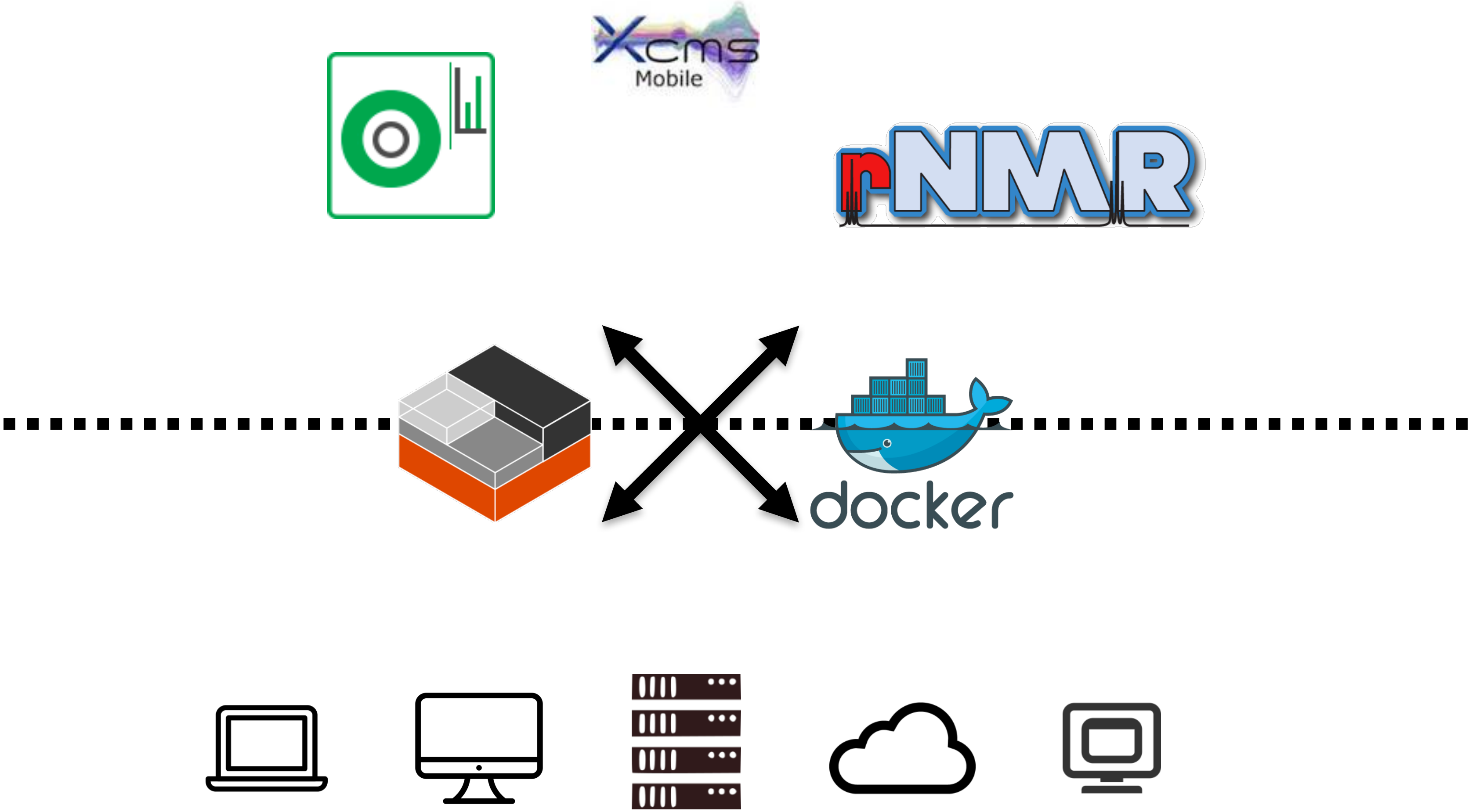
Data

Analysis

Scientists

Major revolution

Same in software

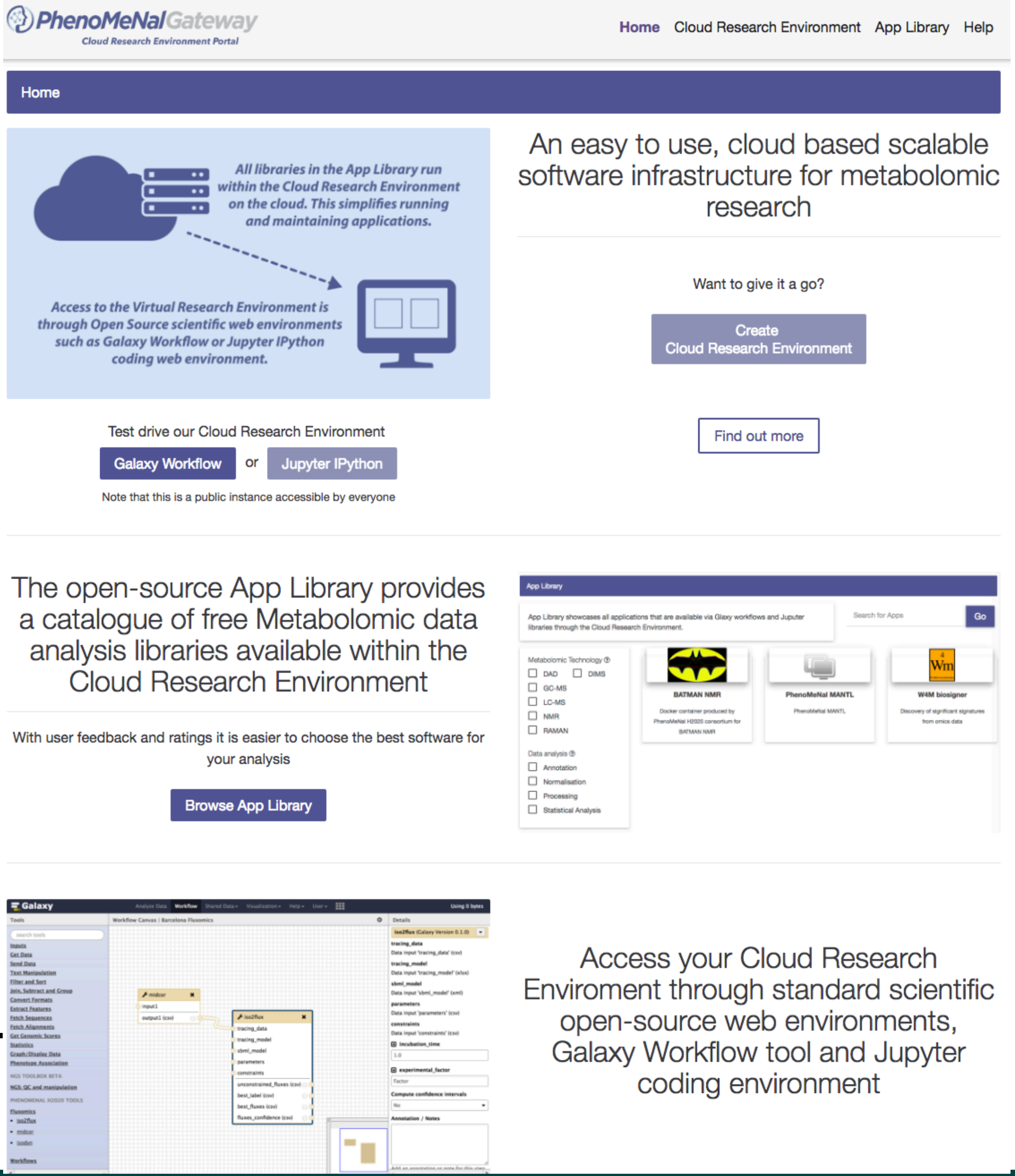


Developer's PI's Cluster Cloud Collaborator's

VRE Portal

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

http://portal.phenomenal-h2020.eu/



All libraries in the App Library run within the Cloud Research Environment on the cloud. This simplifies running and maintaining applications.

Access to the Virtual Research Environment is through Open Source scientific web environments such as Galaxy Workflow or Jupyter IPython coding web environment.

An easy to use, cloud based scalable software infrastructure for metabolomic research

Want to give it a go?

Test drive our Cloud Research Environment

Note that this is a public instance accessible by everyone

The open-source App Library provides a catalogue of free Metabolomic data analysis libraries available within the Cloud Research Environment

With user feedback and ratings it is easier to choose the best software for your analysis

App Library showcases all applications that are available via Glaxy workflows and Juputer libraries through the Cloud Research Environment.

Metabolomic Technology
☐ DAD ☐ DIMS
☐ GC-MS
☐ LC-MS
☐ NMR
☐ RAMAN

Data analysis
☐ Annotation
☐ Normalisation
☐ Processing
☐ Statistical Analysis

BATMAN NMR
Docker container produced by PhenoMeNal H2020 consortium for BATMAN NMR

PhenoMeNal MANTL
PhenoMeNal MANTL

W4M biosigner
Discovery of significant signatures from omics data

Access your Cloud Research Enviroment through standard scientific open-source web environments, Galaxy Workflow tool and Jupyter coding environment

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)

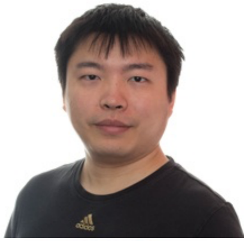
Previous: Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, Pablo Conesa

EBI PhenoMeNal – The team

Kenneth Haug

Reza Salek

Pablo Moreno



Sijin He

Christoph Steinbeck

Namrata Kale

COSMOS consortium

cosmos

SEVENTH FRAMEWORK PROGRAMME

PhenoMeNalconsortium

Funding and Collaborators

BBSRC
bioscience for the future

EMBL-EBI

isasoftwaresuite

UNIVERSITY OF
CAMBRIDGE

MRC Human Nutrition Research
Improving health through nutrition research

SEVENTH FRAMEWORK
PROGRAMME