Reproducibility and Standardization

Pietro Franceschi
pietro.franceschi@fmach.it

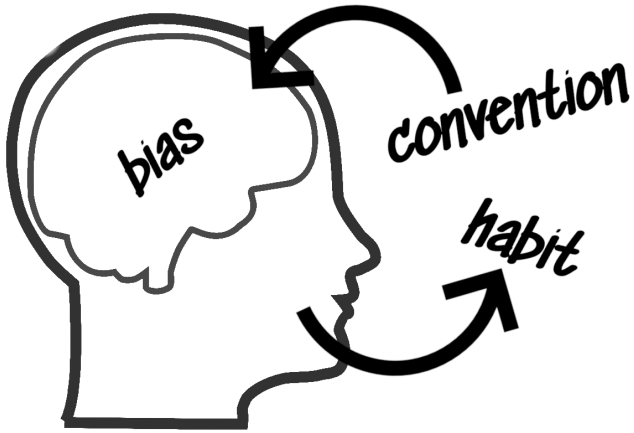
FEM - UBC

Science and Reproducibility

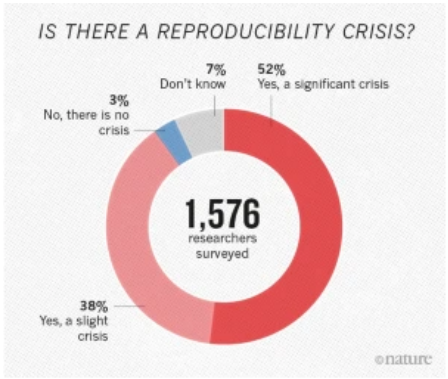
Every scientific result should be reproducible

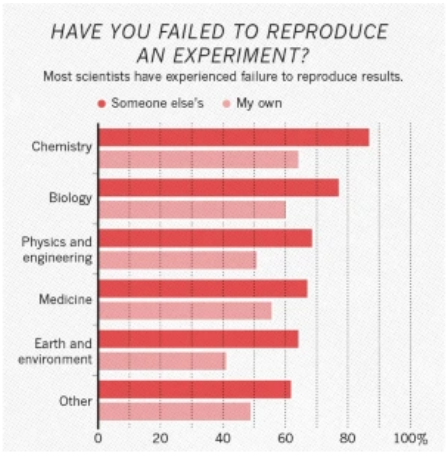
Sources of non-reproducibility

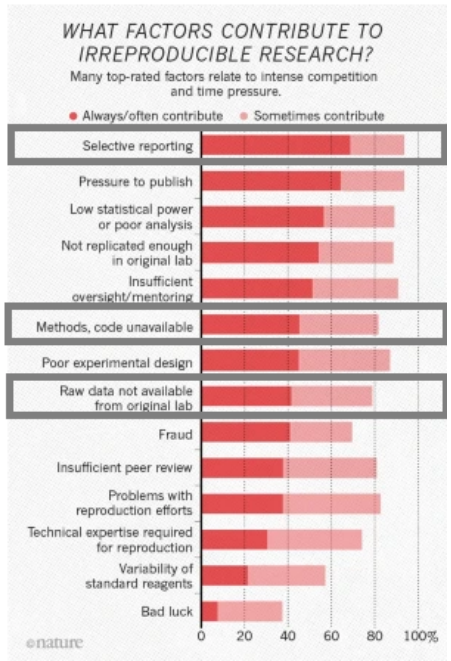
- Errors
- Frauds
- Incomplete or bad reporting
- False positives (sampling)
- ...



Reproducibility in Science







Nature. 2016;533:452-454. doi: 10.1038/533452a

"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments"

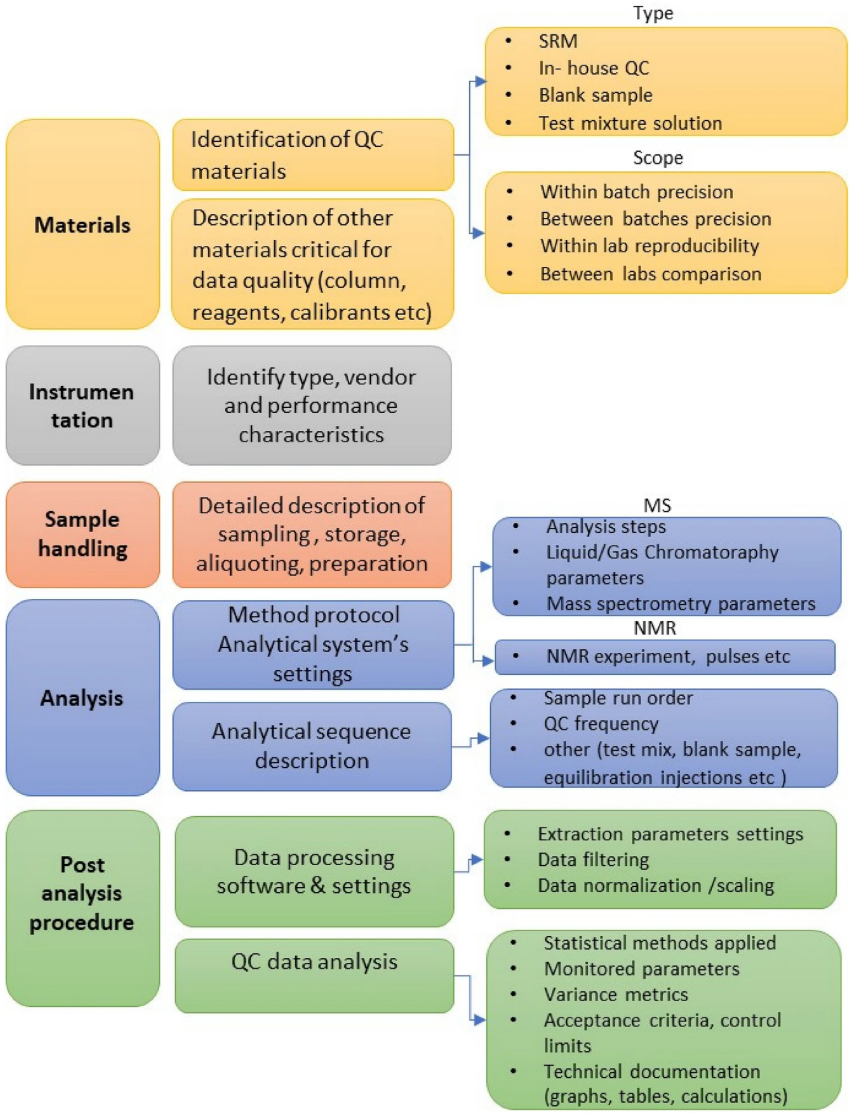
Sharing and Standardization

- The impact of all the abovementioned problems can be reduced by **sharing** data and methods
- Sharing requires **standardization**

`If I have seen further, it is by standing on the shoulders of giants`

Isaac Newton

Minimum Reporting Standards

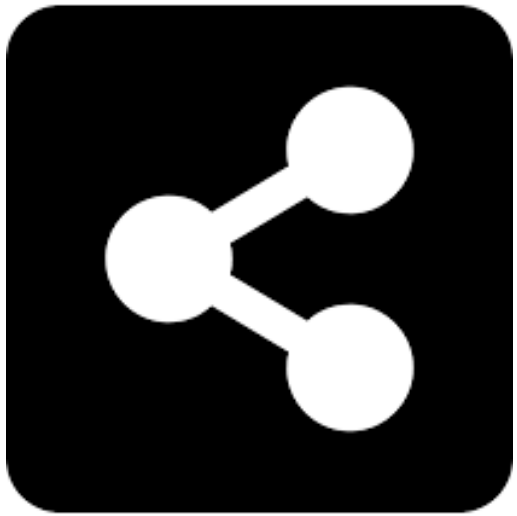


What can we standardize in metabolomics?

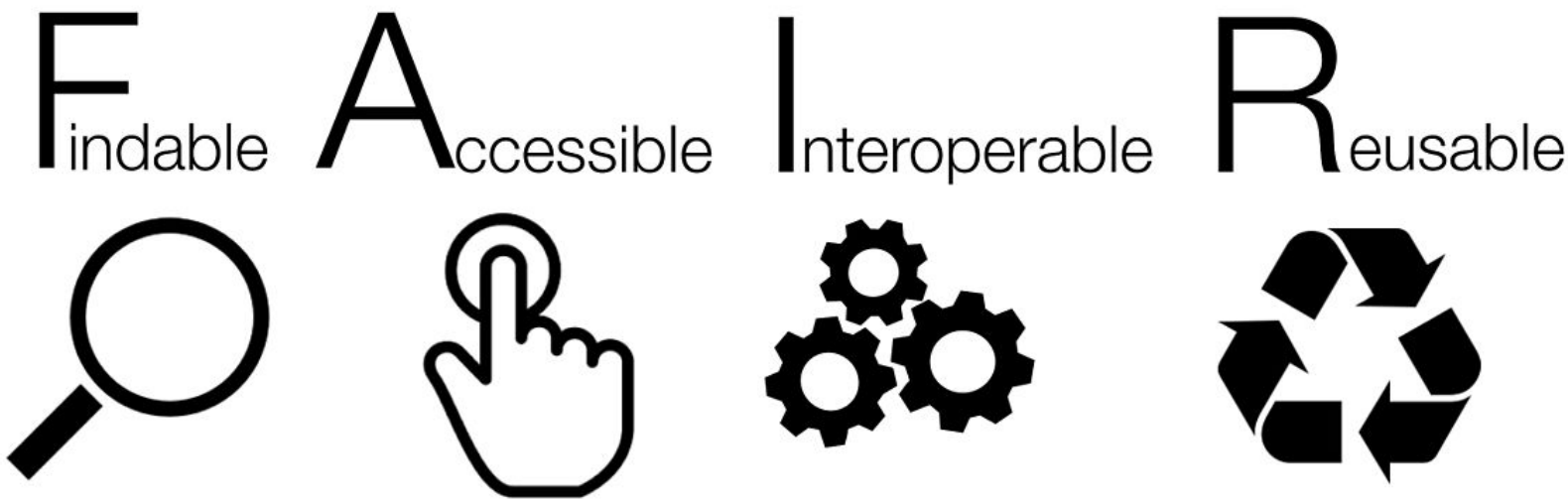
- Analytical protocols
- Data analysis pipelines
- Description of samples and experiments
- Names and onthologies
- Names of chemicals
- Annotation levels (untargeted)

What can we share in metabolomics?

- Sample metadata
- Raw experimental data
- Data Analysis Scripts
- MS and NMR spectra



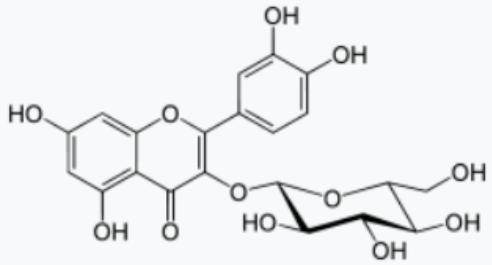
Please . . . be FAIR



Sharing is Value

Standardization can be tricky: chemical names

Isoquercetin



Names	
IUPAC name	3-(β-D-Glucopyranosyloxy)-3′,4′,5,7-tetrahydroxyflavone
Preferred IUPAC name	2-(3,4-Dihydroxyphenyl)-5,7-dihydroxy-3-{[(2*S*,3*R*,4*S*,5*S*,6*R*)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxy}-4*H*-1-benzopyran-4-one
Other names	Isoquercitroside Isoquercitin Isoquercetin Trifoliin Isotrifolin Trifoliin A Isohyperoside Isotrifoliin Quercetin-3-glucoside Quercetin-3-*O*-glucoside Quercetin 3-*O*-β-D-glucopyranoside

Identifiers	
CAS Number	482-35-9 ✓
3D model (JSmol)	Interactive image
ChemSpider	4444361
ECHA InfoCard	100.123.856
PubChem CID	5280804
UNII	6HN2PC637T ✓
CompTox Dashboard (EPA)	DTXSID3041110

InChI [hide]

InChI=1S/C21H20O12/c22-6-13-15(27)17(29)18(30)21(32-13)33-20-16(28)14-11(26)4-8(23)5-12(14)31-19(20)7-1-2-9(24)10(25)3-7/h1-5,13,15,17-18,21-27,29-30H,6H2/t13-,15-,17+,18-,21+/m1/s1

Key: OVSQVDMCBVZWGM-QSOFNFLRSA-N

InChI=1/C21H20O12/c22-6-13-15(27)17(29)18(30)21(32-13)33-20-16(28)14-11(26)4-8(23)5-12(14)31-19(20)7-1-2-9(24)10(25)3-7/h1-5,13,15,17-18,21-27,29-30H,6H2/t13-,15-,17+,18-,21+/m1/s1

Key: OVSQVDMCBVZWGM-QSOFNFLRBX

SMILES [hide]

C1=CC(=C(C=C1C2=C(C(=O)C3=C(C=C(C=C3O2)O)O)O[C@H]4[C@@H]([C@H]([C@@H]([C@H](O4)CO)O)O)O)O)O

What is good for a computer is not necessarily good for a man . . .

Isa Tab

Standardizing metadata for scientific experiments

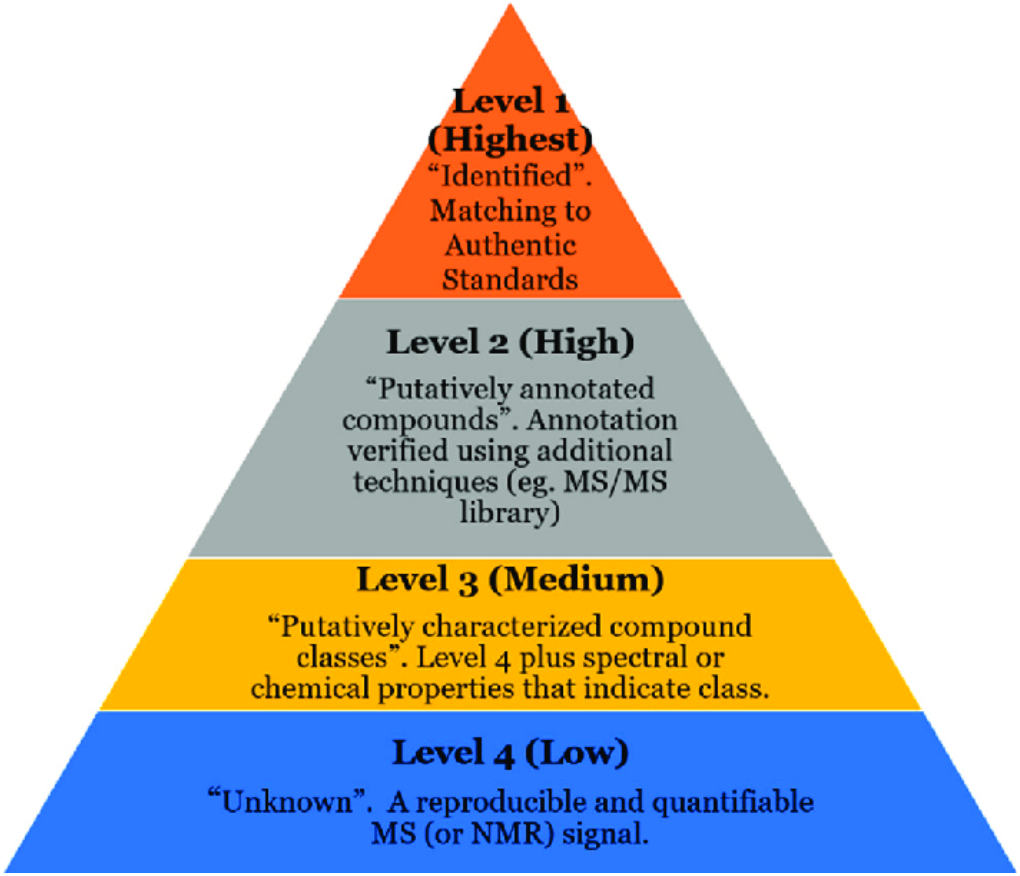
https://isa-tools.org/

MIAPPE

Minimum Information About Plant Phenotyping Experiments

https://www.miappe.org/

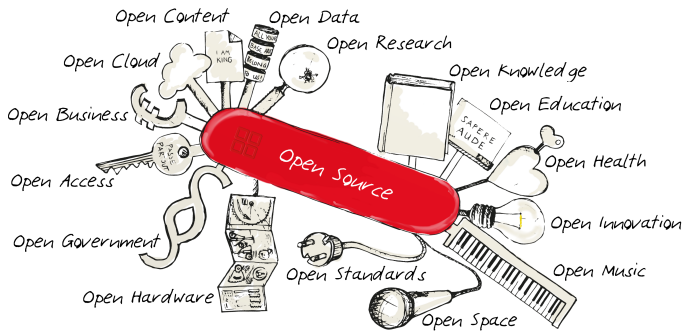
Annotation Levels



Metabolomics Standard Initiative

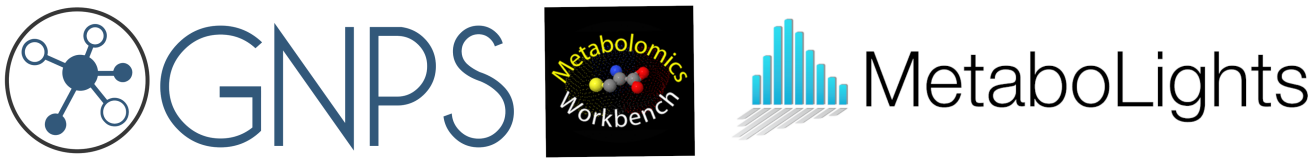
Sharing Raw data

- The **machine specific** data can be more difficult to use but they are the most informative
- **Open Sources Formats** (mzML) boost interoperability and community analysis
- For targeted assays coherent tables are sufficient
- Untargeted MS data can be converted with *proteowizard*



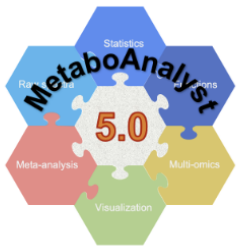
Metabolomics Data Repositories

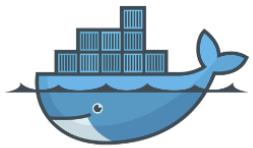
- **Metabolights**:https://www.ebi.ac.uk/metabolights/
- **Metabolomics Workbench**:https://www.metabolomicsworkbench.org/
- **GNPS**:https://gnps.ucsd.edu
- ...



Sharing data analysis pipelines

- Scripting Language (R,Python,Matlab) . . . not Excel ;-)
- Workflow Managers *Online*
  - Workflow4Metabolomics: https://workflow4metabolomics.org/
  - Metaboanalyst: https://www.metaboanalyst.ca/
  - `xcms` online: https://xcmsonline.scripps.edu/
- Workflow Managers *Offline*
  - Galaxy: https://usegalaxy.org/
  - Knime: https://www.knime.com/
- Containers
  - Docker: https://www.docker.com/
  - Singularity: https://apptainer.org/





Sharing Spectra

Databases of spectra are useful for **annotation** and for the training of **machine based** annotation approaches

Reference databases

Biological Magnetic Resonance Data Bank: a repository for data from NMR spectroscopy on proteins, peptides, nucleic acids, and other biomolecules. Developed by the University of Wisconsin.

Birmingham Metabolite Library: contains >3000 experimental 1D and 2D J-resolved NMR spectra of 208 metabolite standards. This resource was established by the University of Birmingham, UK, and was funded by the BBSRC.

Glycan Mass Spectral Database:database of MS spectra data for N- and O-linked glycans and glycolipids glycans along with their partial chemical structures.

Golm Metabolome Database: GC-MS metabolomics library developed and maintained in a collaboration between the Root Metabolism Group and the Bioinformatics Group of the Max Planck Institute for Molecular Plant Physiology, Germany.

Human Metabolome Database: a comprehensive, high-quality, freely accessible, online database of small molecule metabolites found in the human body.

MassBank Japan: one of the largest open databases of mass spectral data and covers numerous different instrument types. MassBank was initiated by the Institute for Advanced Biosciences, in Keio University, Tsuruoka City, Yamagata, Japan.

MassBank Bank: Europe Mirror:one of the largest open databases of mass spectral data and covers numerous different instrument types

MassBank of North America:a metadata-centric, auto-curating repository designed for efficient storage and querying of mass spectral records.

METLIN: is a repository of metabolite information as well as tandem mass spectrometry data.

MzCloud: a web-based mass spectral database that comprises a curated collection of high and low resolution tandem mass spectra acquired under a number of experimental conditions which address the problem of spectra reproducibility.

NIST Standard Reference Database: extensive collection of data sets (EI MS, MS/MS, Replicate spectra, Retention index).

Phenol-Explorer: a database on natural phenols and polyphenols including food composition, food processing, and polyphenol metabolites in human and experimental animals.

Spectral Database for Organic Compounds: repository of spectral database of organic compound; Variety of data sets (MS, NMR, IR, Raman, ESR).

https://wiki.metabolomicssociety.org/index.php/Databases