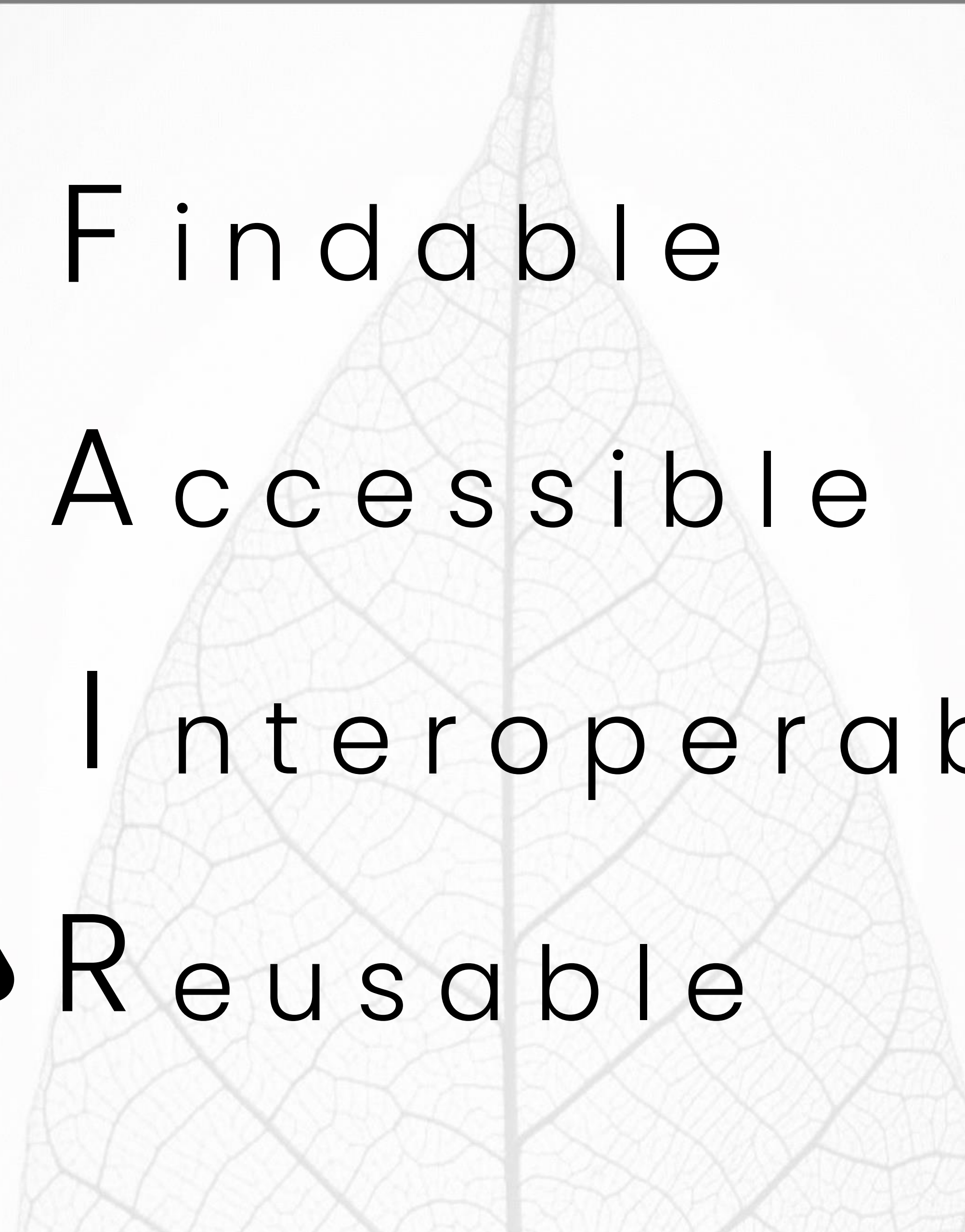
Bari, 29-04-2022

WORKSHOP
Metabolomics and Integrative omics:
from data production to analysis

F indable A ccessible I nteroperable R eusable

Data and Metadata Management

Stefania SAVOI



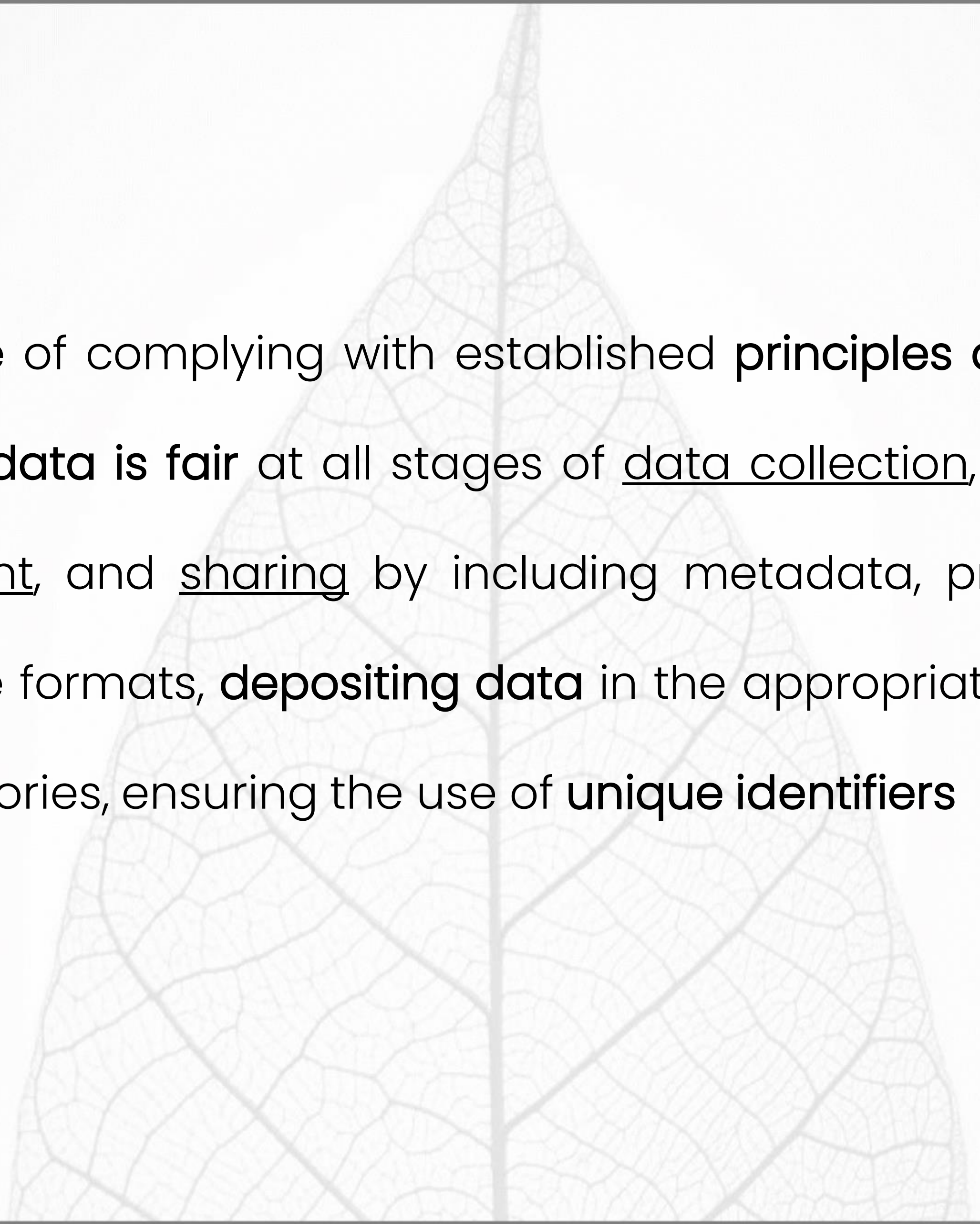
Findable

Accessible

Interoperable

Reusable

Wilkinson et al., 2016
Scientific Data



The practice of complying with established **principles and standards** so that the **data is fair** at all stages of data collection, data analysis, management, and sharing by including metadata, proving data in standard file formats, **depositing data** in the appropriate database or data repositories, ensuring the use of **unique identifiers**

FAIR Principles Compliance

approaching the wishlist ...

Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

✓ **F1**. Resource is uploaded to a public repository.

✓ **F2**. Metadata are assigned a globally unique and persistent identifier.

Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

✓ **A1**. Resource is accessible for download or manipulation by humans and is ideally also machine readable.

✓ **A2**. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

✓ **I1**. Resource is uploaded to a repository that is interoperable with other platforms.

✓ **I2**. Repository meta- data schema maps to or implements the CG Core metadata schema.

✓ **I3**. Metadata use standard vocabularies and/or ontologies.

Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

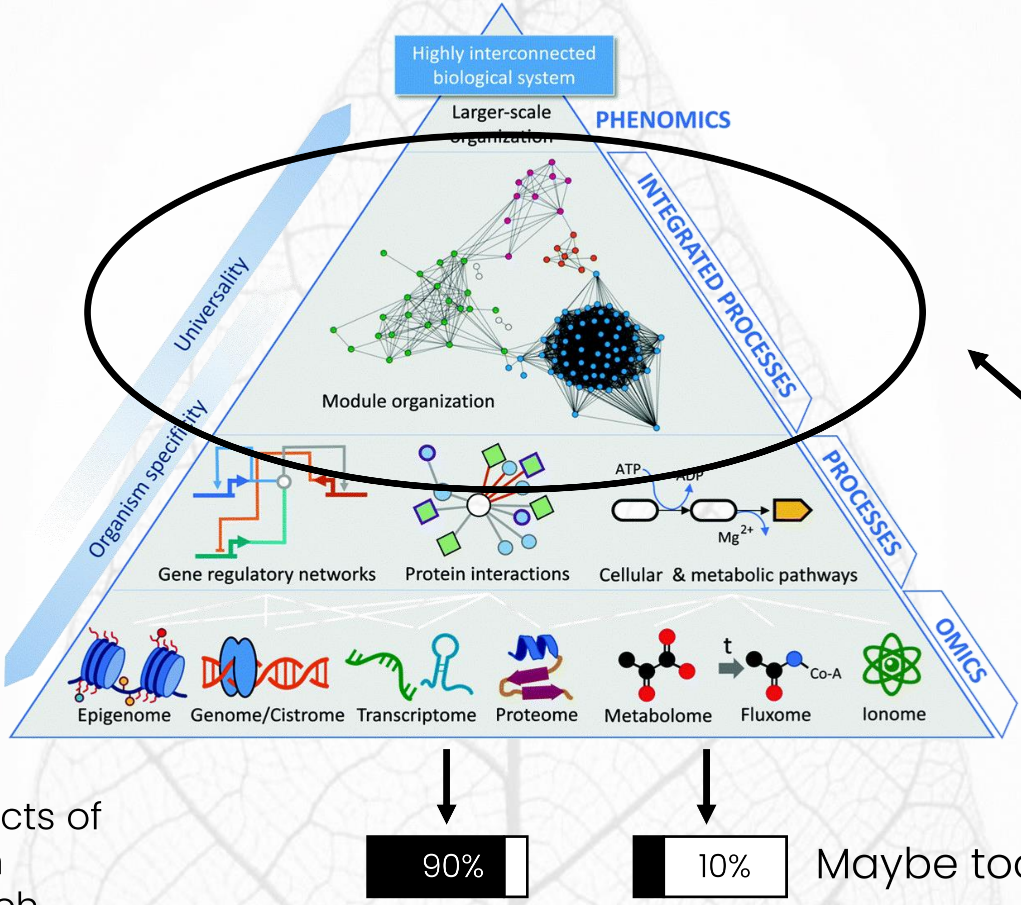
✓ **R1.** Metadata are released with a clear and accessible usage license.

✓ **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

“The ~~cheapest~~ experiment is the one already in the database”

the most valuable

ELIXIR EUROPE Tweet 11/03/21
Data-Driven Innovation in the Agritech Sector



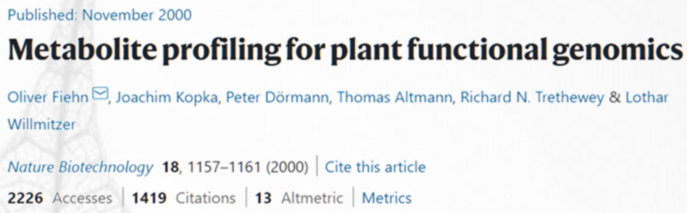
We need:
DATA
SHARING

Maybe too much optimistic?

Status and Prospects of Systems Biology in Grapevine Research, Matus et al, 2019

Why data sharing is so important?

- Knowledge is additive
- No one can measure everything
- Standing on the shoulder of giants
- Good science is reproducible
- Context is continuously changing



Published: November 2000

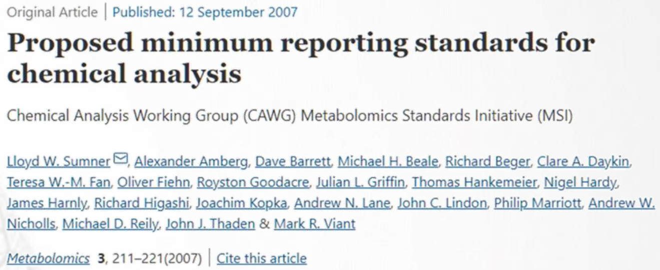
Metabolite profiling for plant functional genomics

Oliver Fiehn, Joachim Kopka, Peter Dörmann, Thomas Altmann, Richard N. Trethewey & Lothar Willmitzer

Nature Biotechnology **18**, 1157–1161 (2000) | Cite this article

2226 Accesses | **1419** Citations | **13** Altmetric | Metrics

2000 Metabol “omics”



Original Article | Published: 12 September 2007

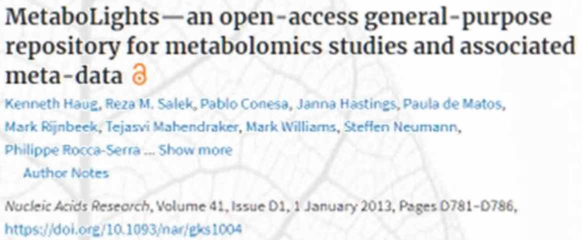
Proposed minimum reporting standards for chemical analysis

Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)

Lloyd W. Sumner, Alexander Amberg, Dave Barrett, Michael H. Beale, Richard Beger, Clare A. Daykin, Teresa W.-M. Fan, Oliver Fiehn, Royston Goodacre, Julian L. Griffin, Thomas Hankemeier, Nigel Hardy, James Harnly, Richard Higashi, Joachim Kopka, Andrew N. Lane, John C. Lindon, Philip Marriott, Andrew W. Nicholls, Michael D. Reily, John J. Thaden & Mark R. Viant

Metabolomics **3**, 211–221(2007) | Cite this article

2007 Standard



MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data

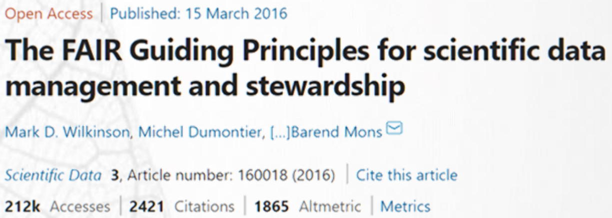
Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, Paula de Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra ... Show more

Author Notes

Nucleic Acids Research, Volume 41, Issue D1, 1 January 2013, Pages D781–D786, https://doi.org/10.1093/nar/gks1004

2013 MetaboLights

EMBL-EBI



Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

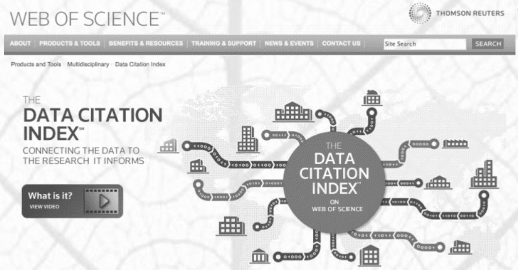
Mark D. Wilkinson, Michel Dumontier, [...]Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

212k Accesses | **2421** Citations | **1865** Altmetric | Metrics

2016 FAIR principles

2021 Rewards



WEB OF SCIENCE™

THOMSON REUTERS

ABOUT | PRODUCTS & TOOLS | BENEFITS & RESOURCES | TRAINING & SUPPORT | NEWS & EVENTS | CONTACT US
Site Search
SEARCH

Products and Tools › Multidisciplinary › Data Citation Index

THE DATA CITATION INDEX™

CONNECTING THE DATA TO THE RESEARCH IT INFORMS

What is it?
VIEW VIDEO
THE DATA CITATION INDEX™ ON WEB OF SCIENCE

Horizon 2020
European Union Funding
for Research & Innovation

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

COST

(European **Co**operation in **S**cience and **T**echnology)

is a funding organization for research and innovation networks

↓

COST Actions

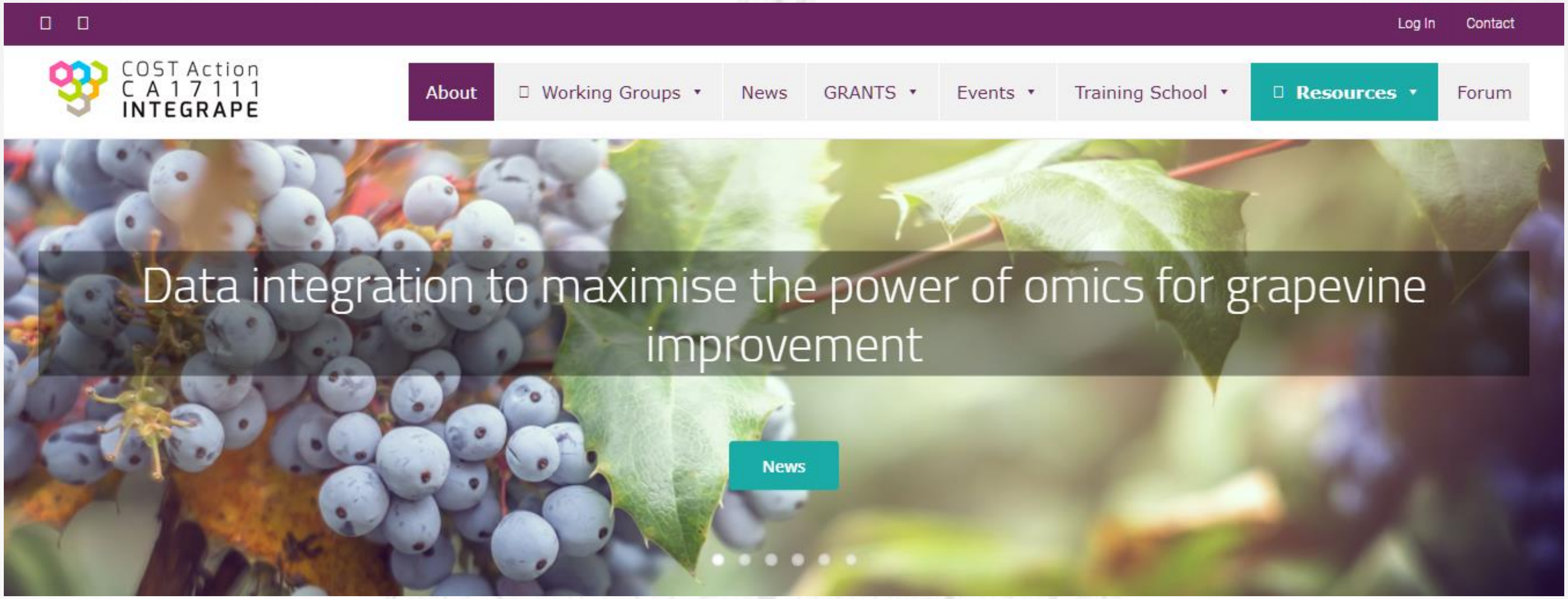
are network dedicated to scientific collaboration,

complementing national research funds

COST Action
CA17111
INTEGRAPE

Imagine having all data on grapevine in a single place ...

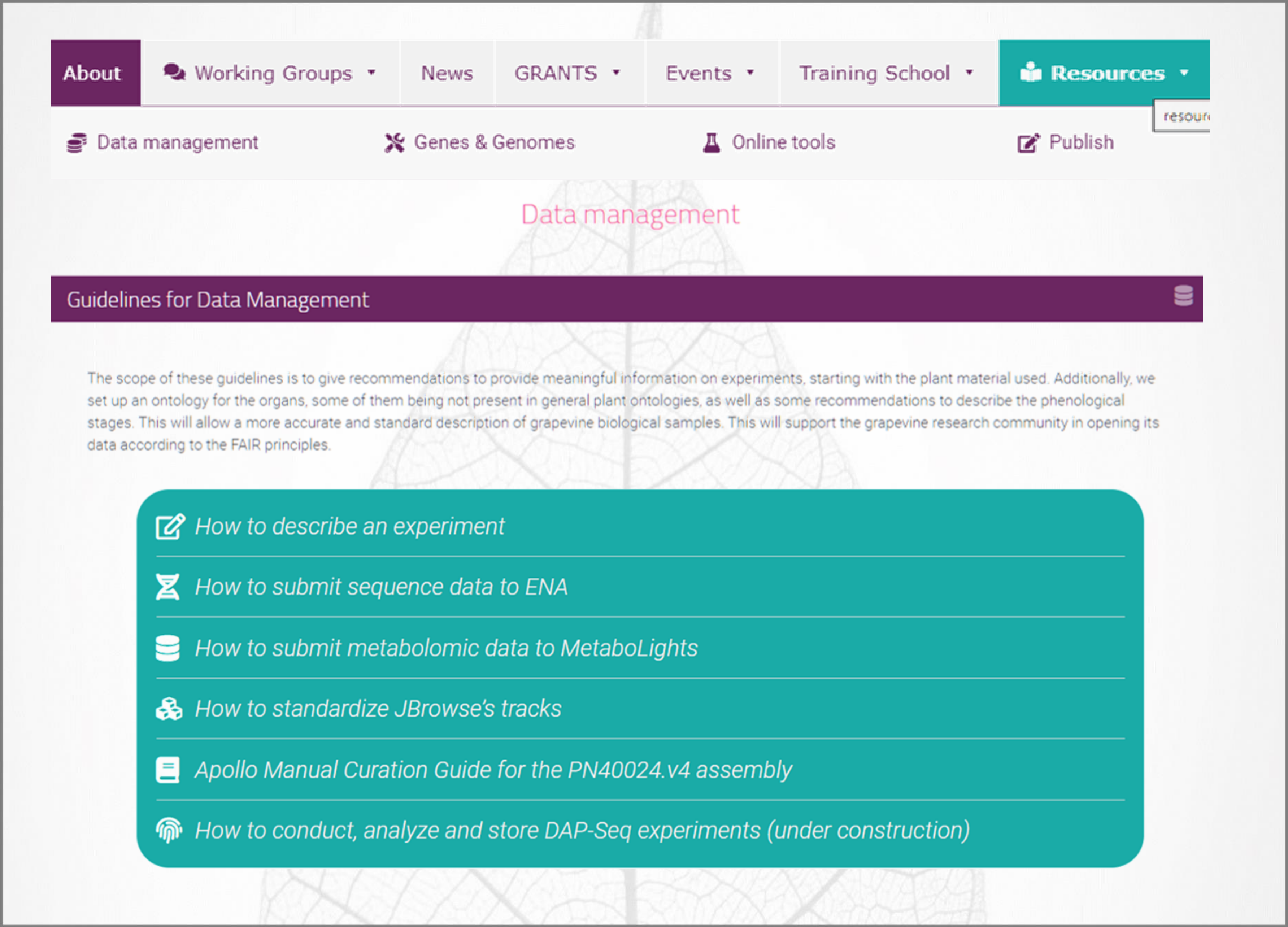
https://www.integrape.eu/



Data integration to maximise the power of omics for grapevine improvement

AIMS to develop **minimal data standards** and **good practices** in order **to integrate** data repositories and **improve** interoperability between datasets

- 4 YEARS: Sept 2018 - Sept 2022
- 22 European countries
- 3 COST Near Neighbour countries
- 2 COST International partner countries



Data management

Guidelines for Data Management

The scope of these guidelines is to give recommendations to provide meaningful information on experiments, starting with the plant material used. Additionally, we set up an ontology for the organs, some of them being not present in general plant ontologies, as well as some recommendations to describe the phenological stages. This will allow a more accurate and standard description of grapevine biological samples. This will support the grapevine research community in opening its data according to the FAIR principles.

- *How to describe an experiment*
- *How to submit sequence data to ENA*
- *How to submit metabolomic data to MetaboLights*
- *How to standardize JBrowse's tracks*
- *Apollo Manual Curation Guide for the PN40024.v4 assembly*
- *How to conduct, analyze and store DAP-Seq experiments (under construction)*

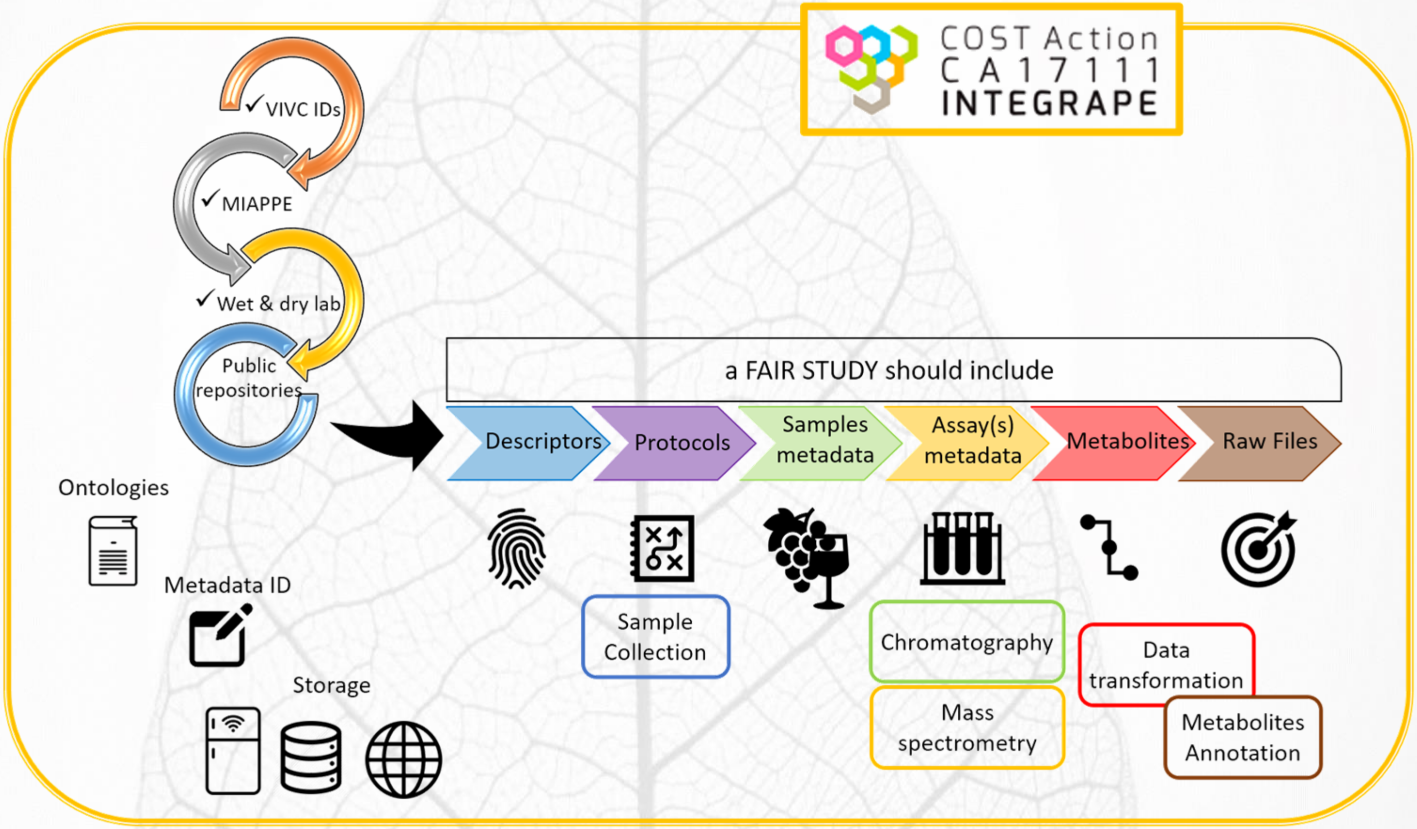
metabolites

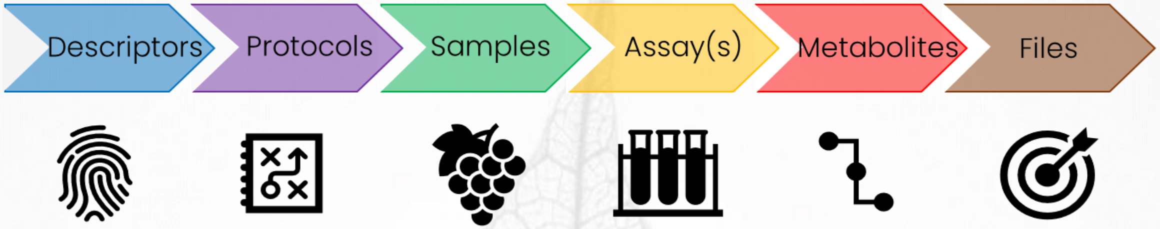
MDPI

Technical Note

Grapevine and Wine Metabolomics-Based Guidelines for FAIR Data and Metadata Management

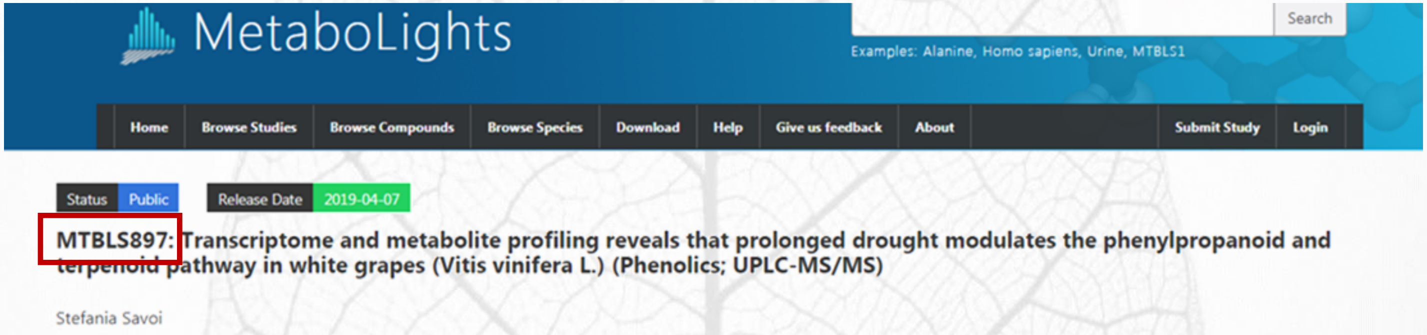
Stefania Savoi [1], Panagiotis Arapitsas [2,*], Éric Duchêne [3], Maria Nikolantonaki [4], Ignacio Ontañón [5], Silvia Carlin [2], Florian Schwander [6], Régis D. Gougeon [4], António César Silva Ferreira [7], Georgios Theodoridis [8], Reinhard Töpfer [6], Urska Vrhovsek [2], Anne-Francoise Adam-Blondon [9], Mario Pezzotti [10,*] and Fulvio Mattivi [2,11]





The fingerprint of the study

- Each study uploaded on a repository has a unique alphanumeric study ID (e.g., MTBLS897)
- Linked publication
- Keywords

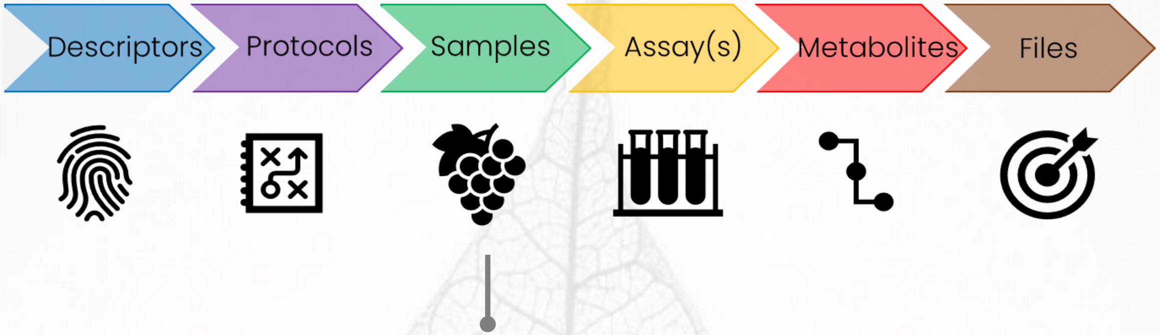


Status Public

Release Date 2019-04-07

MTBLS897: Transcriptome and metabolite profiling reveals that prolonged drought modulates the phenylpropanoid and terpenoid pathway in white grapes (Vitis vinifera L.) (Phenolics; UPLC-MS/MS)

Stefania Savoi



Descriptors

Protocols

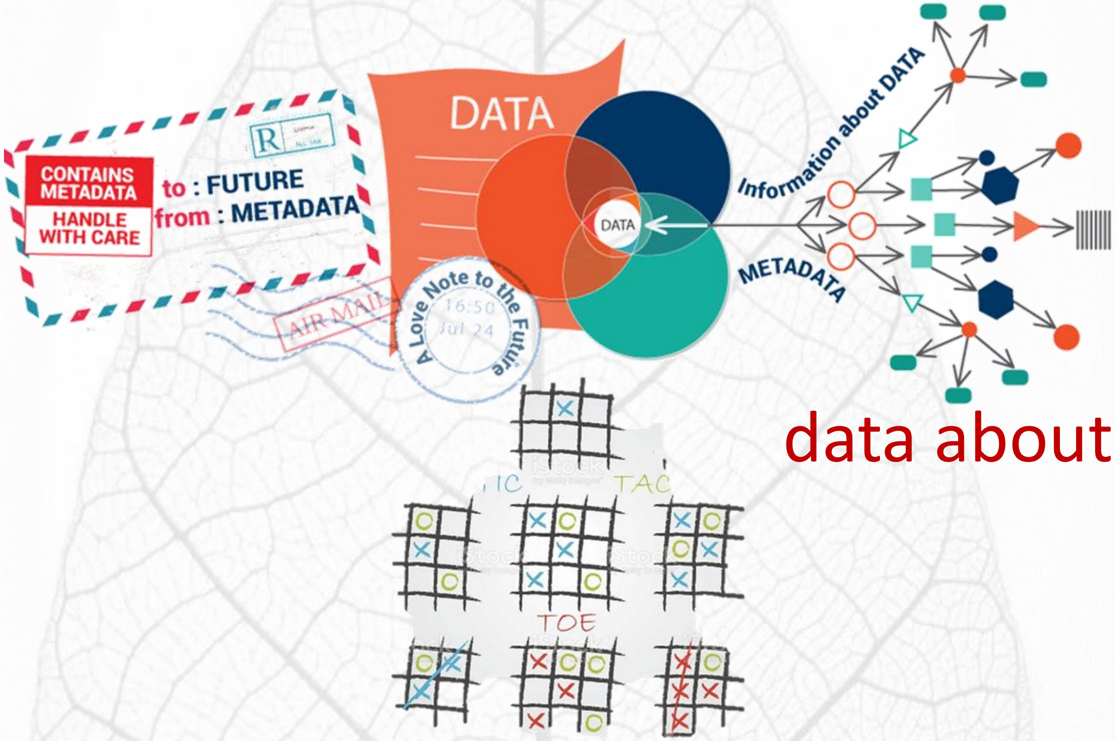
Samples

Assay(s)

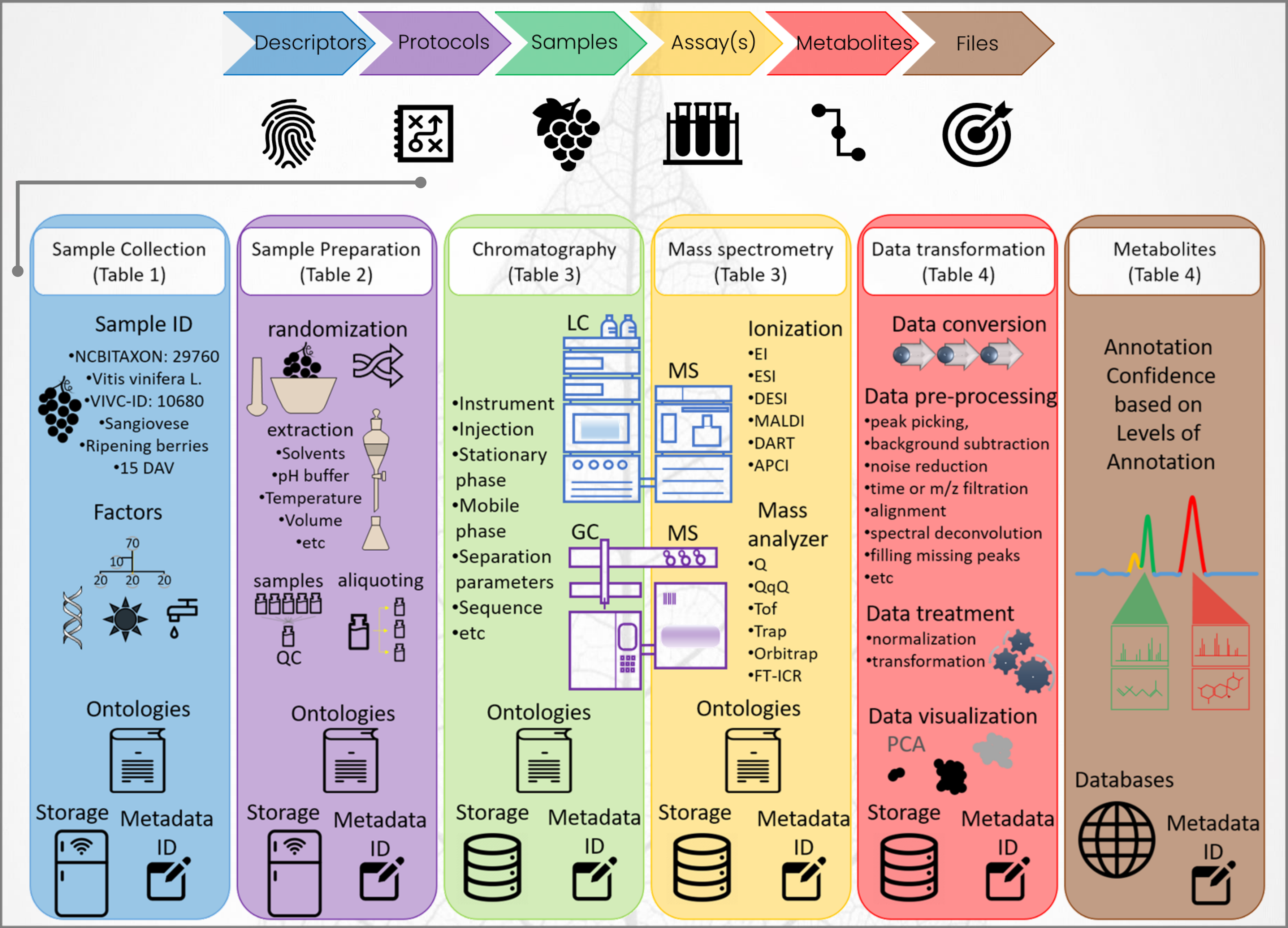
Metabolites

Files

Info about sample and all the metadata we can think of



data about data



Descriptors → Protocols → Samples → Assay(s) → Metabolites → Files

Sample Collection (Table 1)

Sample ID

- NCBITAXON: 29760
- Vitis vinifera L.
- VIVC-ID: 10680
- Sangiovese
- Ripening berries
- 15 DAV

Factors

Ontologies

Storage

Metadata ID

Sample Preparation (Table 2)

randomization

extraction

- Solvents
- pH buffer
- Temperature
- Volume
- etc

samples

QC

aliquoting

Ontologies

Storage

Metadata ID

Chromatography (Table 3)

- Instrument
- Injection
- Stationary phase
- Mobile phase
- Separation parameters
- Sequence
- etc

Ontologies

Storage

Metadata ID

Mass spectrometry (Table 3)

Ionization

- EI
- ESI
- DESI
- MALDI
- DART
- APCI

Mass analyzer

- Q
- QqQ
- Tof
- Trap
- Orbitrap
- FT-ICR

Ontologies

Storage

Metadata ID

Data transformation (Table 4)

Data conversion

Data pre-processing

- peak picking,
- background subtraction
- noise reduction
- time or m/z filtration
- alignment
- spectral deconvolution
- filling missing peaks
- etc

Data treatment

- normalization
- transformation

Data visualization

Storage

Metadata ID

Metabolites (Table 4)

Annotation Confidence based on Levels of Annotation

Databases

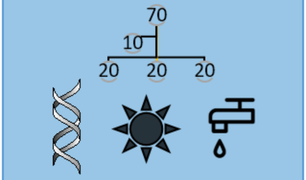
Metadata ID

Sample Collection
(Table 1)

Sample ID

- NCBITAXON: 29760
- Vitis vinifera L.
- VIVC-ID: 10680
- Sangiovese
- Ripening berries
- 15 DAV

Factors



Ontologies

Storage

Metadata ID

From Savoi et al., 2021
Metabolites

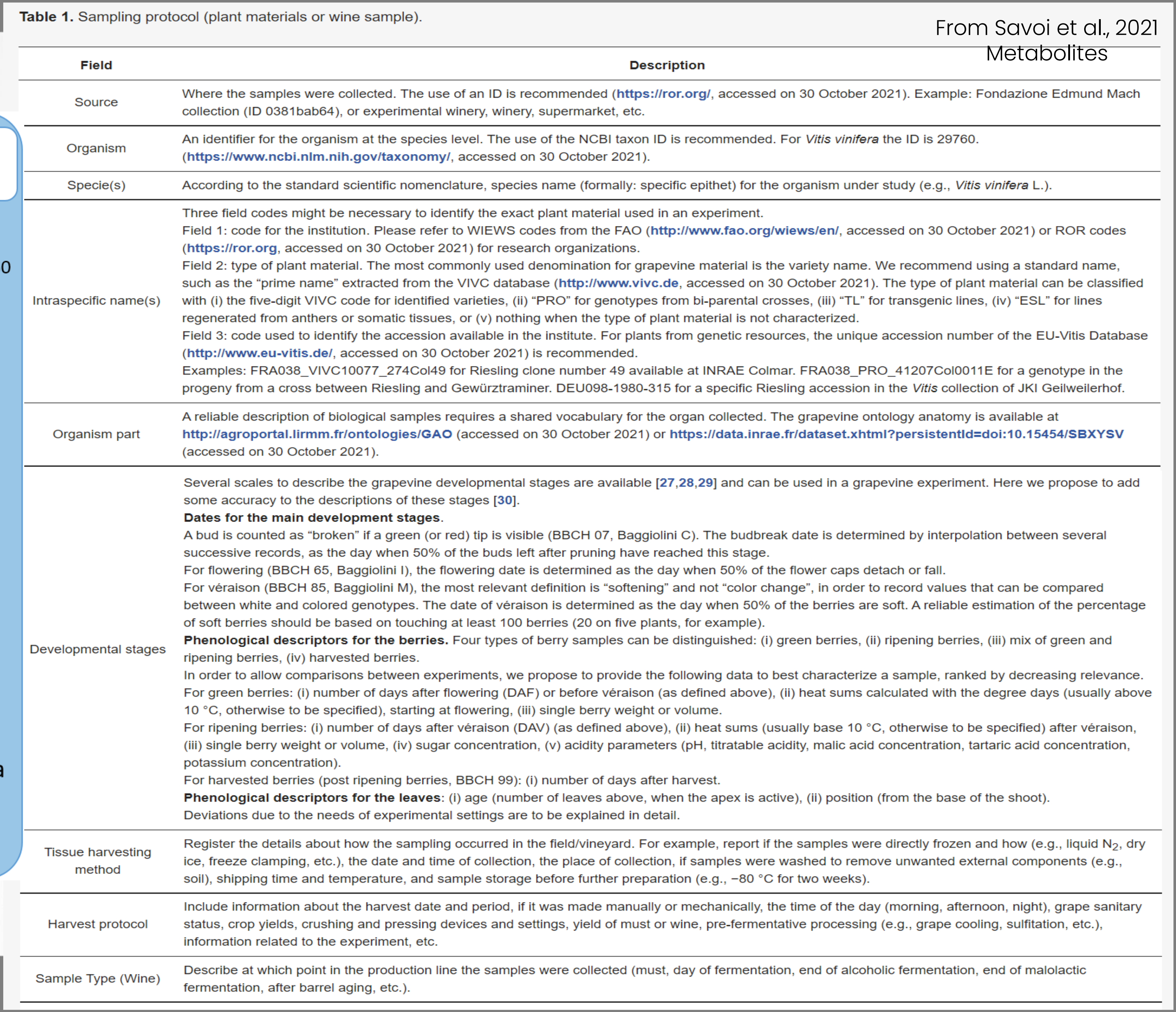
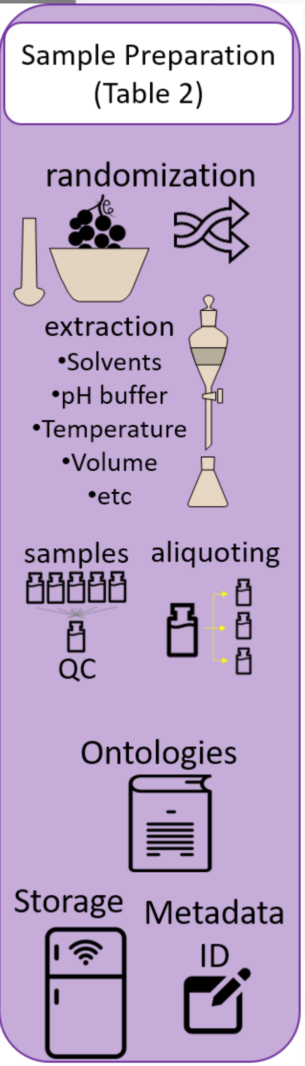


Table 1. Sampling protocol (plant materials or wine sample).

Field	Description
Source	Where the samples were collected. The use of an ID is recommended (**https://ror.org/**, accessed on 30 October 2021). Example: Fondazione Edmund Mach collection (ID 0381bab64), or experimental winery, winery, supermarket, etc.
Organism	An identifier for the organism at the species level. The use of the NCBI taxon ID is recommended. For *Vitis vinifera* the ID is 29760. (**https://www.ncbi.nlm.nih.gov/taxonomy/**, accessed on 30 October 2021).
Specie(s)	According to the standard scientific nomenclature, species name (formally: specific epithet) for the organism under study (e.g., *Vitis vinifera* L.).
Intraspecific name(s)	Three field codes might be necessary to identify the exact plant material used in an experiment. Field 1: code for the institution. Please refer to WIEWS codes from the FAO (**http://www.fao.org/wiews/en/**, accessed on 30 October 2021) or ROR codes (**https://ror.org**, accessed on 30 October 2021) for research organizations. Field 2: type of plant material. The most commonly used denomination for grapevine material is the variety name. We recommend using a standard name, such as the “prime name” extracted from the VIVC database (**http://www.vivc.de**, accessed on 30 October 2021). The type of plant material can be classified with (i) the five-digit VIVC code for identified varieties, (ii) “PRO” for genotypes from bi-parental crosses, (iii) “TL” for transgenic lines, (iv) “ESL” for lines regenerated from anthers or somatic tissues, or (v) nothing when the type of plant material is not characterized. Field 3: code used to identify the accession available in the institute. For plants from genetic resources, the unique accession number of the EU-Vitis Database (**http://www.eu-vitis.de/**, accessed on 30 October 2021) is recommended. Examples: FRA038_VIVC10077_274Col49 for Riesling clone number 49 available at INRAE Colmar. FRA038_PRO_41207Col0011E for a genotype in the progeny from a cross between Riesling and Gewürztraminer. DEU098-1980-315 for a specific Riesling accession in the *Vitis* collection of JKI Geilweilerhof.
Organism part	A reliable description of biological samples requires a shared vocabulary for the organ collected. The grapevine ontology anatomy is available at **http://agroportal.lirmm.fr/ontologies/GAO** (accessed on 30 October 2021) or **https://data.inrae.fr/dataset.xhtml?persistentId=doi:10.15454/SBXYSV** (accessed on 30 October 2021).
Developmental stages	Several scales to describe the grapevine developmental stages are available [27,28,29] and can be used in a grapevine experiment. Here we propose to add some accuracy to the descriptions of these stages [30]. **Dates for the main development stages**. A bud is counted as “broken” if a green (or red) tip is visible (BBCH 07, Baggiolini C). The budbreak date is determined by interpolation between several successive records, as the day when 50% of the buds left after pruning have reached this stage. For flowering (BBCH 65, Baggiolini I), the flowering date is determined as the day when 50% of the flower caps detach or fall. For véraison (BBCH 85, Baggiolini M), the most relevant definition is “softening” and not “color change”, in order to record values that can be compared between white and colored genotypes. The date of véraison is determined as the day when 50% of the berries are soft. A reliable estimation of the percentage of soft berries should be based on touching at least 100 berries (20 on five plants, for example). **Phenological descriptors for the berries.** Four types of berry samples can be distinguished: (i) green berries, (ii) ripening berries, (iii) mix of green and ripening berries, (iv) harvested berries. In order to allow comparisons between experiments, we propose to provide the following data to best characterize a sample, ranked by decreasing relevance. For green berries: (i) number of days after flowering (DAF) or before véraison (as defined above), (ii) heat sums calculated with the degree days (usually above 10 °C, otherwise to be specified), starting at flowering, (iii) single berry weight or volume. For ripening berries: (i) number of days after véraison (DAV) (as defined above), (ii) heat sums (usually base 10 °C, otherwise to be specified) after véraison, (iii) single berry weight or volume, (iv) sugar concentration, (v) acidity parameters (pH, titratable acidity, malic acid concentration, tartaric acid concentration, potassium concentration). For harvested berries (post ripening berries, BBCH 99): (i) number of days after harvest. **Phenological descriptors for the leaves**: (i) age (number of leaves above, when the apex is active), (ii) position (from the base of the shoot). Deviations due to the needs of experimental settings are to be explained in detail.
Tissue harvesting method	Register the details about how the sampling occurred in the field/vineyard. For example, report if the samples were directly frozen and how (e.g., liquid N_2, dry ice, freeze clamping, etc.), the date and time of collection, the place of collection, if samples were washed to remove unwanted external components (e.g., soil), shipping time and temperature, and sample storage before further preparation (e.g., −80 °C for two weeks).
Harvest protocol	Include information about the harvest date and period, if it was made manually or mechanically, the time of the day (morning, afternoon, night), grape sanitary status, crop yields, crushing and pressing devices and settings, yield of must or wine, pre-fermentative processing (e.g., grape cooling, sulfitation, etc.), information related to the experiment, etc.
Sample Type (Wine)	Describe at which point in the production line the samples were collected (must, day of fermentation, end of alcoholic fermentation, end of malolactic fermentation, after barrel aging, etc.).

From Savoi et al., 2021
Metabolites



Sample Preparation
(Table 2)

randomization

extraction
- Solvents
- pH buffer
- Temperature
- Volume
- etc

samples

QC

aliquoting

Ontologies

Storage

Metadata
ID

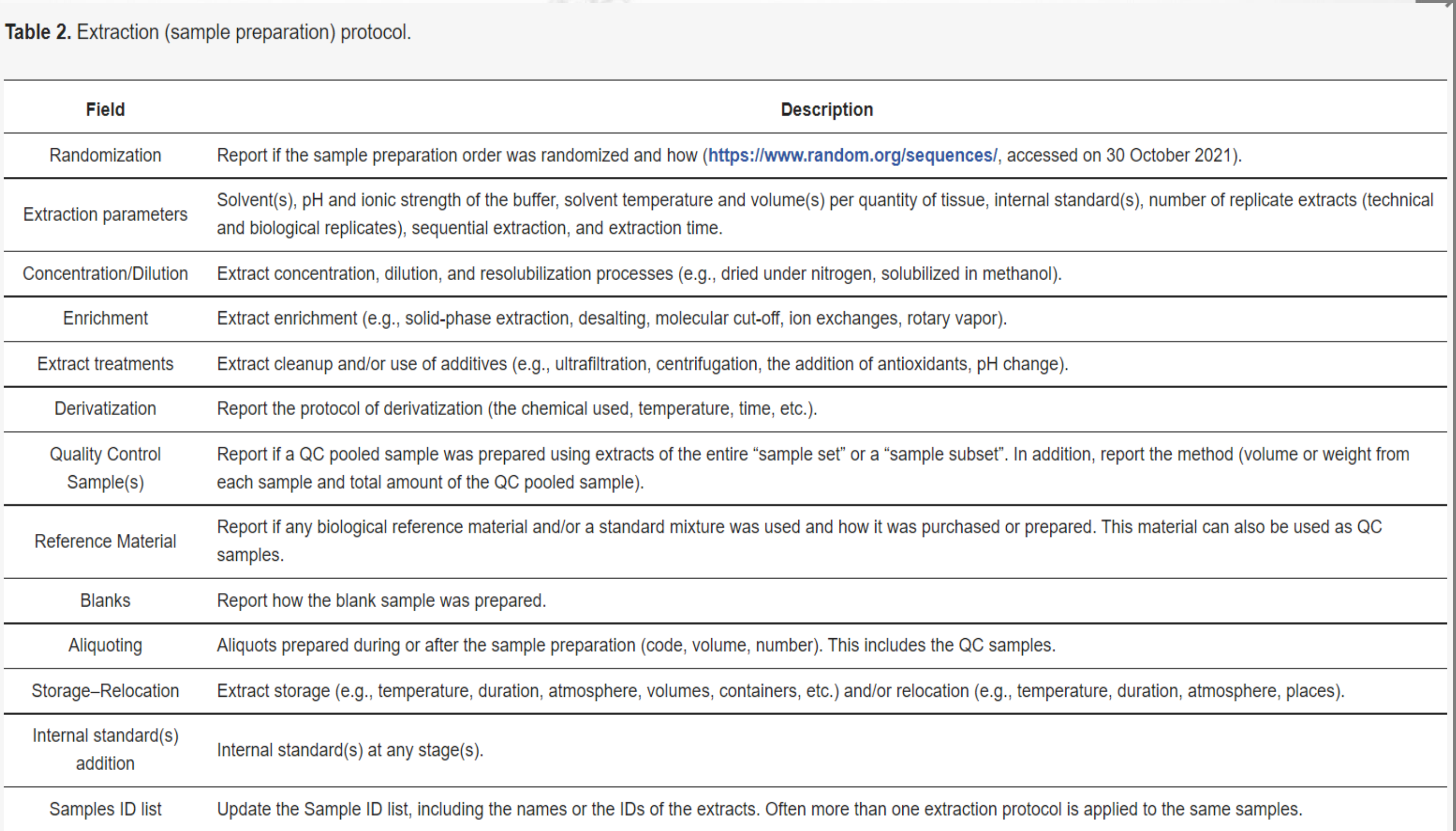
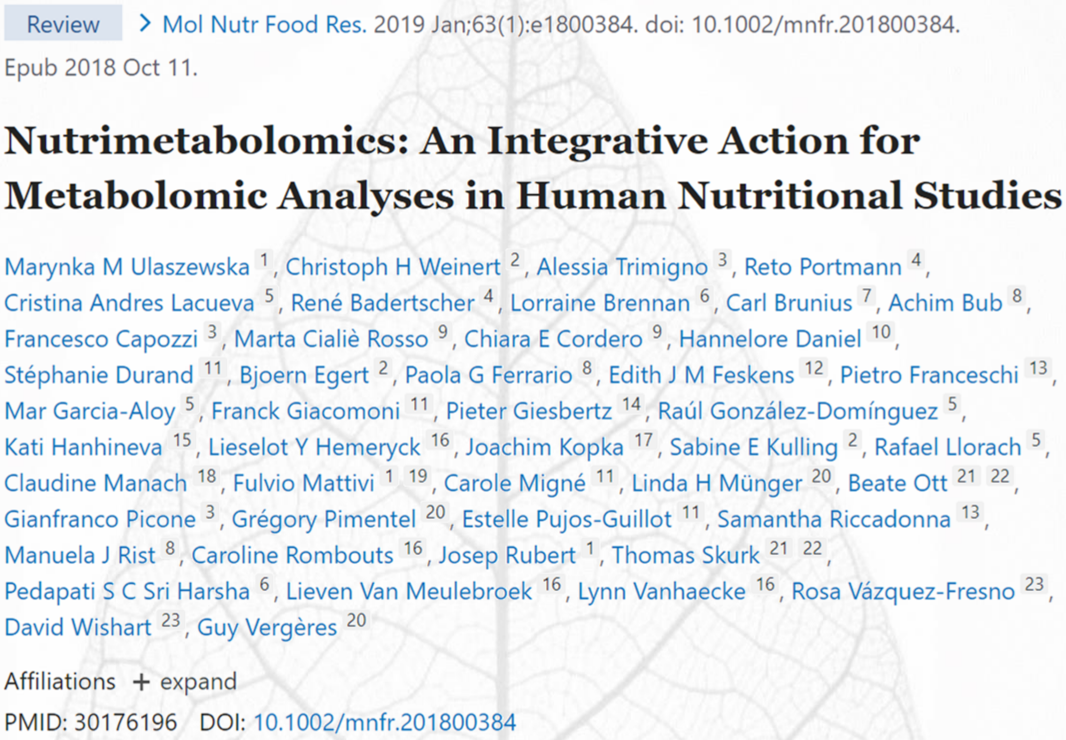


Table 2. Extraction (sample preparation) protocol.

Field	Description
Randomization	Report if the sample preparation order was randomized and how (https://www.random.org/sequences/, accessed on 30 October 2021).
Extraction parameters	Solvent(s), pH and ionic strength of the buffer, solvent temperature and volume(s) per quantity of tissue, internal standard(s), number of replicate extracts (technical and biological replicates), sequential extraction, and extraction time.
Concentration/Dilution	Extract concentration, dilution, and resolubilization processes (e.g., dried under nitrogen, solubilized in methanol).
Enrichment	Extract enrichment (e.g., solid-phase extraction, desalting, molecular cut-off, ion exchanges, rotary vapor).
Extract treatments	Extract cleanup and/or use of additives (e.g., ultrafiltration, centrifugation, the addition of antioxidants, pH change).
Derivatization	Report the protocol of derivatization (the chemical used, temperature, time, etc.).
Quality Control Sample(s)	Report if a QC pooled sample was prepared using extracts of the entire “sample set” or a “sample subset”. In addition, report the method (volume or weight from each sample and total amount of the QC pooled sample).
Reference Material	Report if any biological reference material and/or a standard mixture was used and how it was purchased or prepared. This material can also be used as QC samples.
Blanks	Report how the blank sample was prepared.
Aliquoting	Aliquots prepared during or after the sample preparation (code, volume, number). This includes the QC samples.
Storage–Relocation	Extract storage (e.g., temperature, duration, atmosphere, volumes, containers, etc.) and/or relocation (e.g., temperature, duration, atmosphere, places).
Internal standard(s) addition	Internal standard(s) at any stage(s).
Samples ID list	Update the Sample ID list, including the names or the IDs of the extracts. Often more than one extraction protocol is applied to the same samples.



Review > Mol Nutr Food Res. 2019 Jan;63(1):e1800384. doi: 10.1002/mnfr.201800384. Epub 2018 Oct 11.

Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies

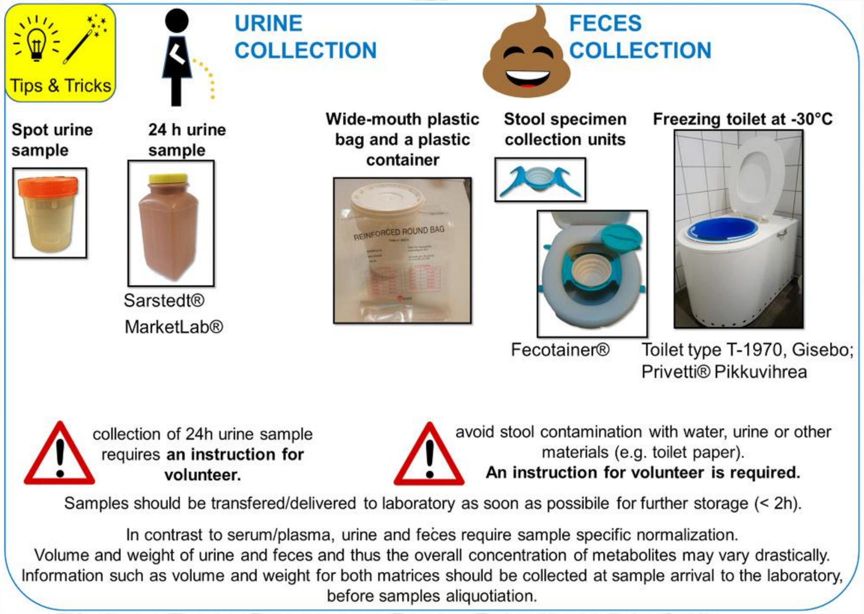
Marynka M Ulaszewska [1], Christoph H Weinert [2], Alessia Trimigno [3], Reto Portmann [4], Cristina Andres Lacueva [5], René Badertscher [4], Lorraine Brennan [6], Carl Brunius [7], Achim Bub [8], Francesco Capozzi [3], Marta Cialiè Rosso [9], Chiara E Cordero [9], Hannelore Daniel [10], Stéphanie Durand [11], Bjoern Egert [2], Paola G Ferrario [8], Edith J M Feskens [12], Pietro Franceschi [13], Mar Garcia-Aloy [5], Franck Giacomoni [11], Pieter Giesbertz [14], Raúl González-Domínguez [5], Kati Hanhineva [15], Lieselot Y Hemeryck [16], Joachim Kopka [17], Sabine E Kulling [2], Rafael Llorach [5], Claudine Manach [18], Fulvio Mattivi [1] [19], Carole Migné [11], Linda H Münger [20], Beate Ott [21] [22], Gianfranco Picone [3], Grégory Pimentel [20], Estelle Pujos-Guillot [11], Samantha Riccadonna [13], Manuela J Rist [8], Caroline Rombouts [16], Josep Rubert [1], Thomas Skurk [21] [22], Pedapati S C Sri Harsha [6], Lieven Van Meulebroek [16], Lynn Vanhaecke [16], Rosa Vázquez-Fresno [23], David Wishart [23], Guy Vergères [20]

Affiliations + expand

PMID: 30176196 DOI: 10.1002/mnfr.201800384

For human nutritional studies

From Ulaszewska et al., 2019
Molecolar Nutrition Food Research

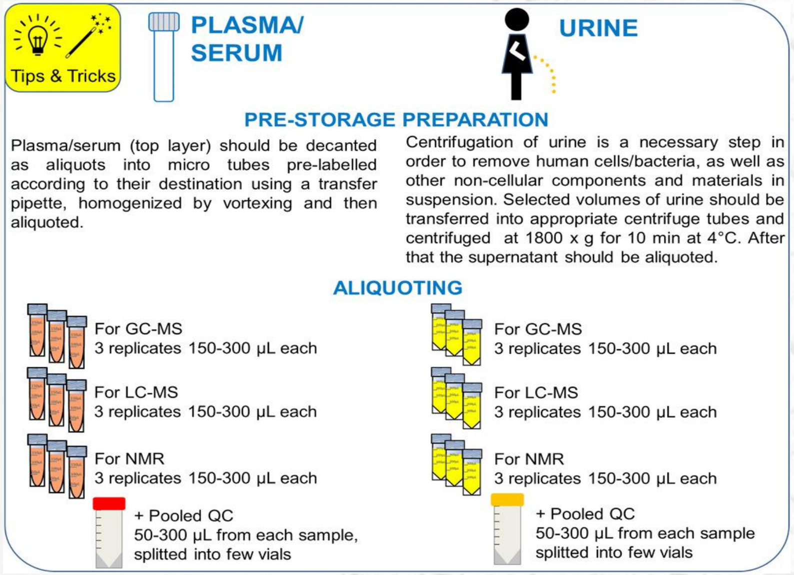


collection of 24h urine sample requires **an instruction for volunteer.**

avoid stool contamination with water, urine or other materials (e.g. toilet paper). **An instruction for volunteer is required.**

Samples should be transfered/delivered to laboratory as soon as possibile for further storage (< 2h).

In contrast to serum/plasma, urine and feces require sample specific normalization. Volume and weight of urine and feces and thus the overall concentration of metabolites may vary drastically. Information such as volume and weight for both matrices should be collected at sample arrival to the laboratory, before samples aliquotiation.



PRE-STORAGE PREPARATION

Plasma/serum (top layer) should be decanted as aliquots into micro tubes pre-labelled according to their destination using a transfer pipette, homogenized by vortexing and then aliquoted.

Centrifugation of urine is a necessary step in order to remove human cells/bacteria, as well as other non-cellular components and materials in suspension. Selected volumes of urine should be transferred into appropriate centrifuge tubes and centrifuged at 1800 x g for 10 min at 4°C. After that the supernatant should be aliquoted.

ALIQUOTING

For GC-MS
3 replicates 150-300 μL each

For LC-MS
3 replicates 150-300 μL each

For NMR
3 replicates 150-300 μL each

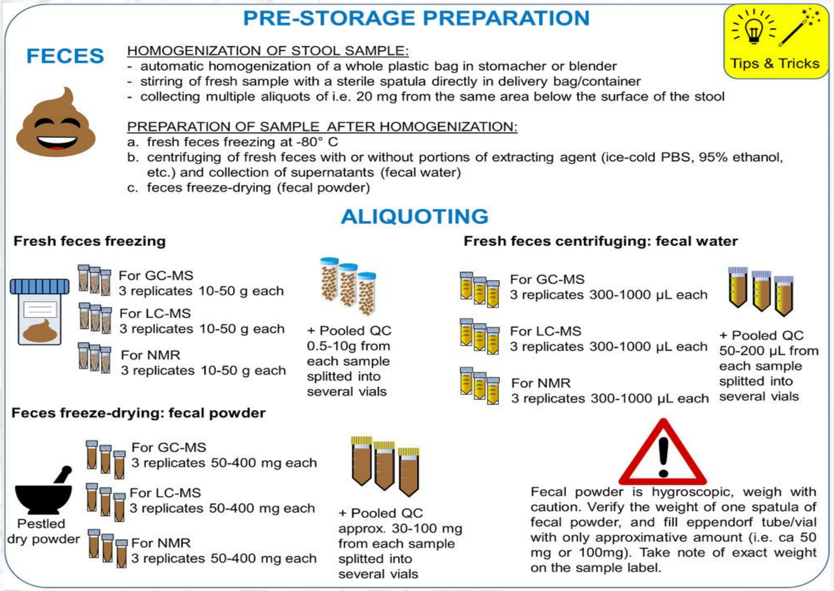
+ Pooled QC
50-300 μL from each sample, splitted into few vials

For GC-MS
3 replicates 150-300 μL each

For LC-MS
3 replicates 150-300 μL each

For NMR
3 replicates 150-300 μL each

+ Pooled QC
50-300 μL from each sample splitted into few vials



PRE-STORAGE PREPARATION

FECES

HOMOGENIZATION OF STOOL SAMPLE:
- automatic homogenization of a whole plastic bag in stomacher or blender
- stirring of fresh sample with a sterile spatula directly in delivery bag/container
- collecting multiple aliquots of i.e. 20 mg from the same area below the surface of the stool

PREPARATION OF SAMPLE AFTER HOMOGENIZATION:
a. fresh feces freezing at -80° C
b. centrifuging of fresh feces with or without portions of extracting agent (ice-cold PBS, 95% ethanol, etc.) and collection of supernatants (fecal water)
c. feces freeze-drying (fecal powder)

ALIQUOTING

Fresh feces freezing

For GC-MS
3 replicates 10-50 g each

For LC-MS
3 replicates 10-50 g each

For NMR
3 replicates 10-50 g each

+ Pooled QC
0.5-10g from each sample splitted into several vials

Fresh feces centrifuging: fecal water

For GC-MS
3 replicates 300-1000 μL each

For LC-MS
3 replicates 300-1000 μL each

For NMR
3 replicates 300-1000 μL each

+ Pooled QC
50-200 μL from each sample splitted into several vials

Feces freeze-drying: fecal powder

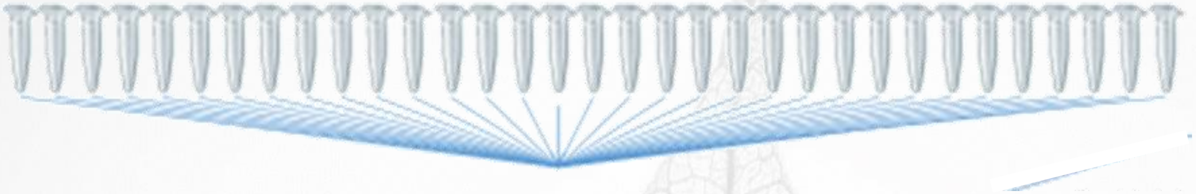
For GC-MS
3 replicates 50-400 mg each

For LC-MS
3 replicates 50-400 mg each

For NMR
3 replicates 50-400 mg each

+ Pooled QC
approx. 30-100 mg from each sample splitted into several vials

Fecal powder is hygroscopic, weigh with caution. Verify the weight of one spatula of fecal powder, and fill eppendorf tube/vial with only approximative amount (i.e. ca 50 mg or 100mg). Take note of exact weight on the sample label.

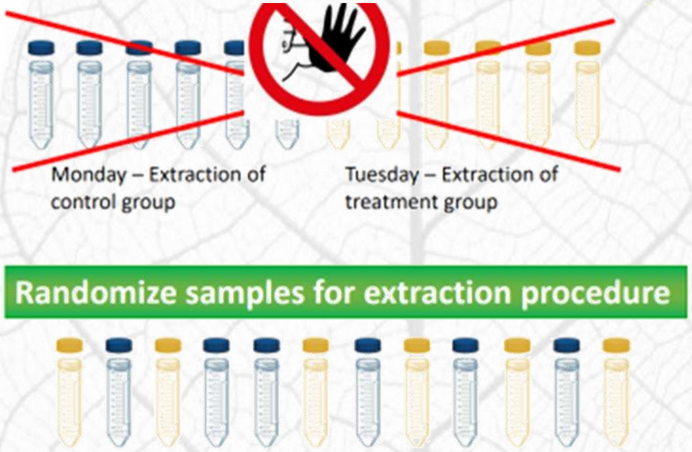


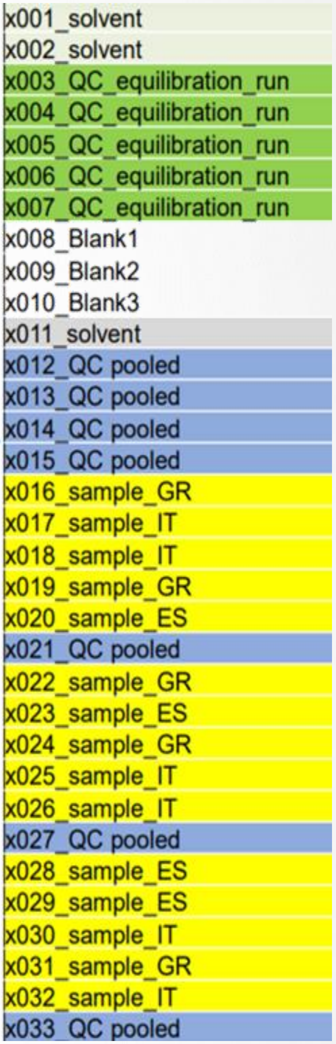
From Ulaszewska et al., 2019
Molecolar Nutrition Food
Research

QC pool sample

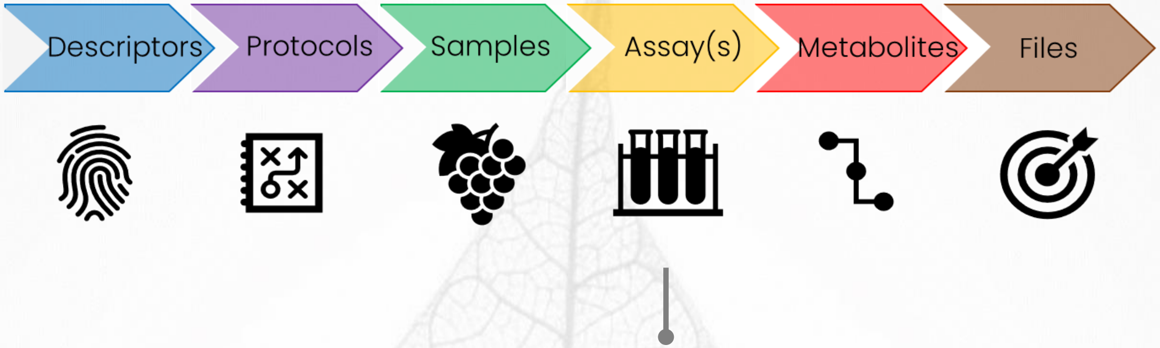
Consists of small aliquots (10uL or 50 uL) taken from each study samples pooled in one vial and injected along a queue many times

Depending on retention time duration it can be every 5, 10 samples


Monday – Extraction of control group
Tuesday – Extraction of treatment group
Randomize samples for extraction procedure



x001_solvent
x002_solvent
x003_QC_equilibration_run
x004_QC_equilibration_run
x005_QC_equilibration_run
x006_QC_equilibration_run
x007_QC_equilibration_run
x008_Blank1
x009_Blank2
x010_Blank3
x011_solvent
x012_QC pooled
x013_QC pooled
x014_QC pooled
x015_QC pooled
x016_sample_GR
x017_sample_IT
x018_sample_IT
x019_sample_GR
x020_sample_ES
x021_QC pooled
x022_sample_GR
x023_sample_ES
x024_sample_GR
x025_sample_IT
x026_sample_IT
x027_QC pooled
x028_sample_ES
x029_sample_ES
x030_sample_IT
x031_sample_GR
x032_sample_IT
x033_QC pooled



Descriptors

Protocols

Samples

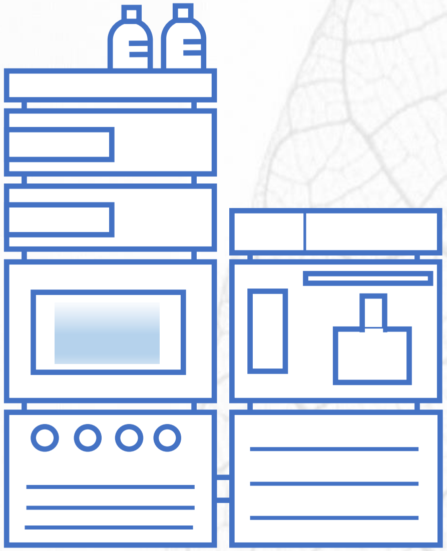
Assay(s)

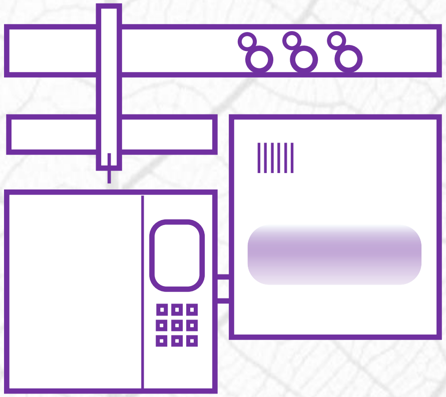
Metabolites

Files

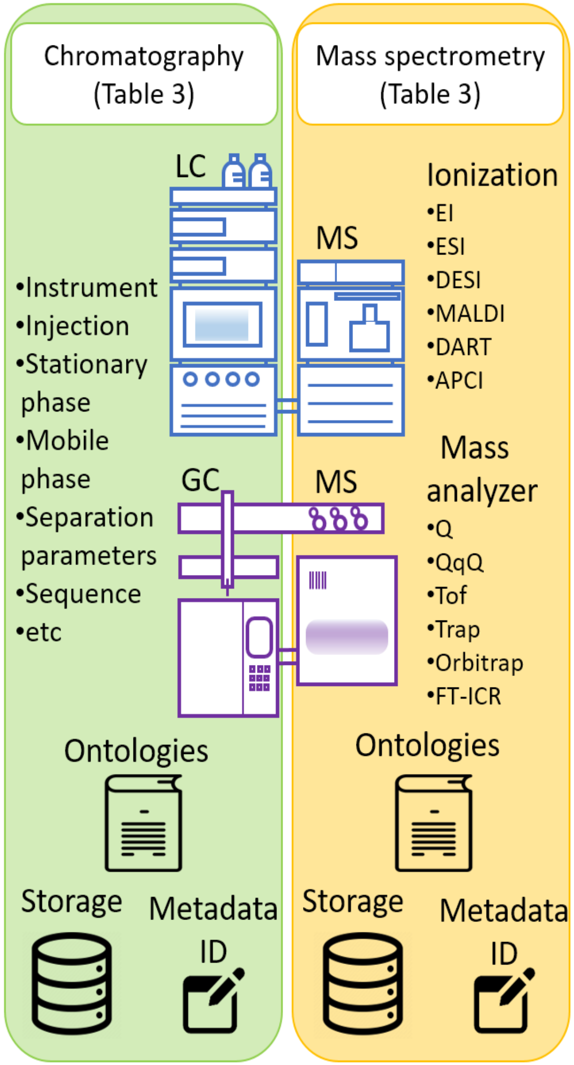
The core of the study

LC-MS	GC-MS	...





From Savoi et al., 2021
Metabolites



Chromatography (Table 3)

LC

GC

- Instrument
- Injection
- Stationary phase
- Mobile phase
- Separation parameters
- Sequence
- etc

Ontologies

Storage

Metadata ID

Mass spectrometry (Table 3)

MS

MS

Ionization
- EI
- ESI
- DESI
- MALDI
- DART
- APCI

Mass analyzer
- Q
- QqQ
- Tof
- Trap
- Orbitrap
- FT-ICR

Ontologies

Storage

Metadata ID

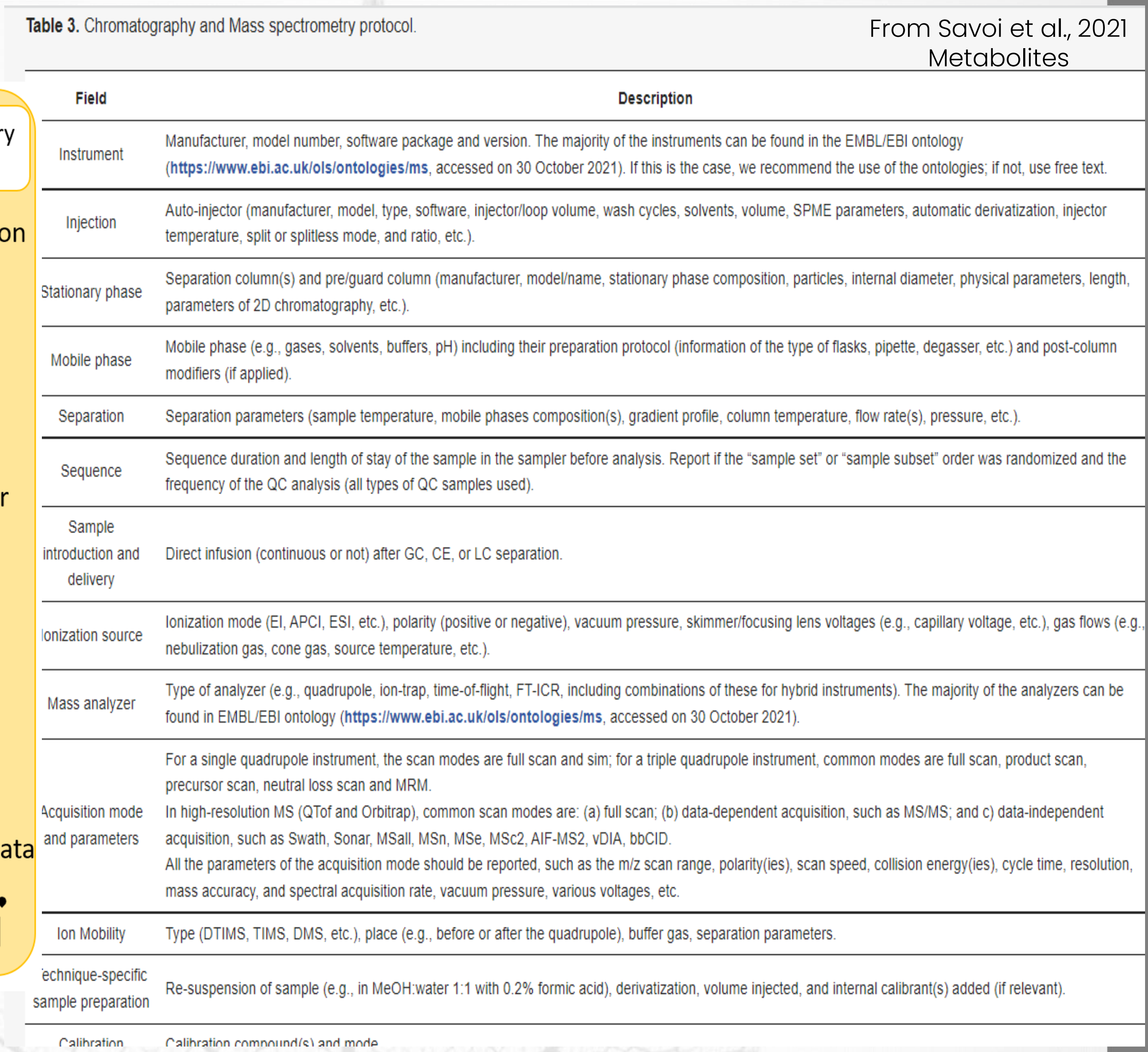


Table 3. Chromatography and Mass spectrometry protocol.

Field	Description
Instrument	Manufacturer, model number, software package and version. The majority of the instruments can be found in the EMBL/EBI ontology (**https://www.ebi.ac.uk/ols/ontologies/ms**, accessed on 30 October 2021). If this is the case, we recommend the use of the ontologies; if not, use free text.
Injection	Auto-injector (manufacturer, model, type, software, injector/loop volume, wash cycles, solvents, volume, SPME parameters, automatic derivatization, injector temperature, split or splitless mode, and ratio, etc.).
Stationary phase	Separation column(s) and pre/guard column (manufacturer, model/name, stationary phase composition, particles, internal diameter, physical parameters, length, parameters of 2D chromatography, etc.).
Mobile phase	Mobile phase (e.g., gases, solvents, buffers, pH) including their preparation protocol (information of the type of flasks, pipette, degasser, etc.) and post-column modifiers (if applied).
Separation	Separation parameters (sample temperature, mobile phases composition(s), gradient profile, column temperature, flow rate(s), pressure, etc.).
Sequence	Sequence duration and length of stay of the sample in the sampler before analysis. Report if the "sample set" or "sample subset" order was randomized and the frequency of the QC analysis (all types of QC samples used).
Sample introduction and delivery	Direct infusion (continuous or not) after GC, CE, or LC separation.
Ionization source	Ionization mode (EI, APCI, ESI, etc.), polarity (positive or negative), vacuum pressure, skimmer/focusing lens voltages (e.g., capillary voltage, etc.), gas flows (e.g., nebulization gas, cone gas, source temperature, etc.).
Mass analyzer	Type of analyzer (e.g., quadrupole, ion-trap, time-of-flight, FT-ICR, including combinations of these for hybrid instruments). The majority of the analyzers can be found in EMBL/EBI ontology (**https://www.ebi.ac.uk/ols/ontologies/ms**, accessed on 30 October 2021).
Acquisition mode and parameters	For a single quadrupole instrument, the scan modes are full scan and sim; for a triple quadrupole instrument, common modes are full scan, product scan, precursor scan, neutral loss scan and MRM. In high-resolution MS (QTof and Orbitrap), common scan modes are: (a) full scan; (b) data-dependent acquisition, such as MS/MS; and c) data-independent acquisition, such as Swath, Sonar, MSall, MSn, MSe, MSc2, AIF-MS2, vDIA, bbCID. All the parameters of the acquisition mode should be reported, such as the m/z scan range, polarity(ies), scan speed, collision energy(ies), cycle time, resolution, mass accuracy, and spectral acquisition rate, vacuum pressure, various voltages, etc.
Ion Mobility	Type (DTIMS, TIMS, DMS, etc.), place (e.g., before or after the quadrupole), buffer gas, separation parameters.
echnique-specific sample preparation	Re-suspension of sample (e.g., in MeOH:water 1:1 with 0.2% formic acid), derivatization, volume injected, and internal calibrant(s) added (if relevant).
Calibration	Calibration compound(s) and mode

Data transformation (Table 4)

Data conversion

Data pre-processing

- peak picking,
- background subtraction
- noise reduction
- time or m/z filtration
- alignment
- spectral deconvolution
- filling missing peaks
- etc

Data treatment

- normalization
- transformation

Data visualization

PCA

Storage

Metadata ID

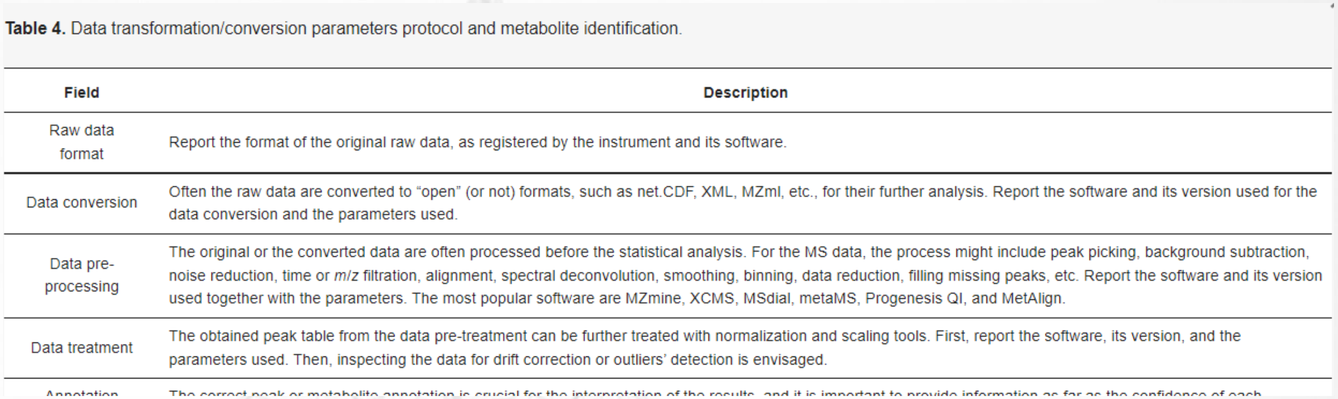
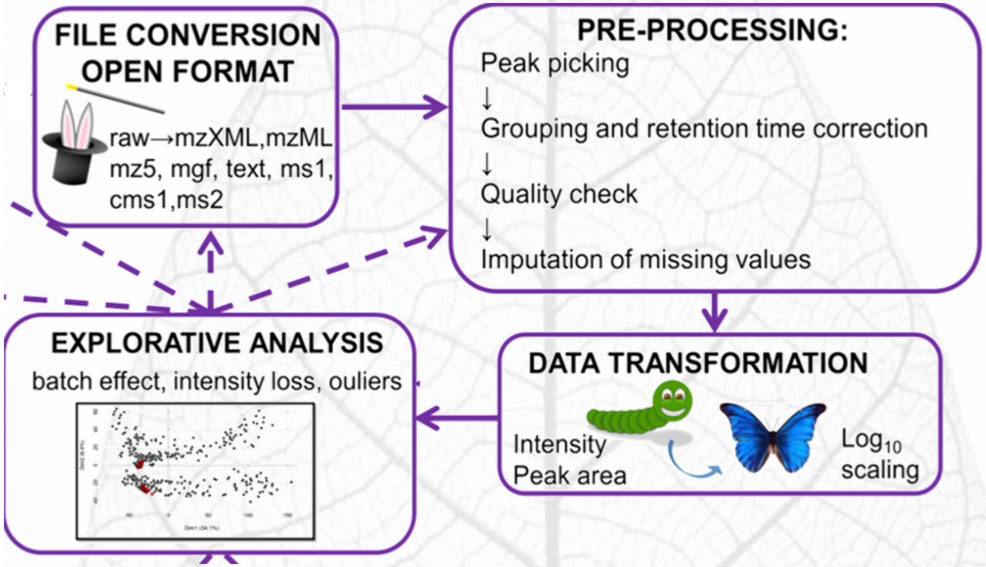


Table 4. Data transformation/conversion parameters protocol and metabolite identification.

Field	Description
Raw data format	Report the format of the original raw data, as registered by the instrument and its software.
Data conversion	Often the raw data are converted to "open" (or not) formats, such as net.CDF, XML, MZml, etc., for their further analysis. Report the software and its version used for the data conversion and the parameters used.
Data pre-processing	The original or the converted data are often processed before the statistical analysis. For the MS data, the process might include peak picking, background subtraction, noise reduction, time or *m*/*z* filtration, alignment, spectral deconvolution, smoothing, binning, data reduction, filling missing peaks, etc. Report the software and its version used together with the parameters. The most popular software are MZmine, XCMS, MSdial, metaMS, Progenesis QI, and MetAlign.
Data treatment	The obtained peak table from the data pre-treatment can be further treated with normalization and scaling tools. First, report the software, its version, and the parameters used. Then, inspecting the data for drift correction or outliers' detection is envisaged.
Annotation	The correct peak or metabolite annotation is crucial for the interpretation of the results, and it is important to provide information as far as the confidence of each



FILE CONVERSION OPEN FORMAT

raw→mzXML,mzML mz5, mgf, text, ms1, cms1,ms2

PRE-PROCESSING:

Peak picking ↓ Grouping and retention time correction ↓ Quality check ↓ Imputation of missing values

DATA TRANSFORMATION

Intensity Peak area

Log_{10} scaling

EXPLORATIVE ANALYSIS

batch effect, intensity loss, ouliers

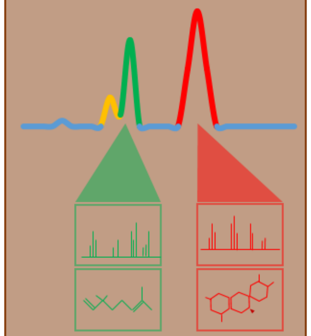
Some data pre-processing software

Schemes of Mar Garcia Aloy

Descriptors → Protocols → Samples → Assay(s) → Metabolites → Files

Metabolites (Table 4)

Annotation Confidence based on Levels of Annotation



Databases

Metadata ID

Metabolite ID

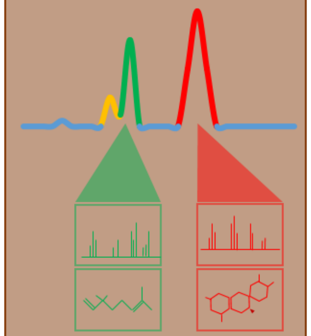
ChEBI
hmdb
The Human Metabolome Database
FOODB
ChemSpider
The free chemical database
KEGG
Kyoto Encyclopedia of Genes and Genomes
PubChem
OPEN CHEMISTRY DATABASE
CAS
A DIVISION OF THE AMERICAN CHEMICAL SOCIETY
COCONUT

Schemes of
Mar Garcia Aloy

Descriptors | Protocols | Samples | Assay(s) | Metabolites | Files

Metabolites
(Table 4)

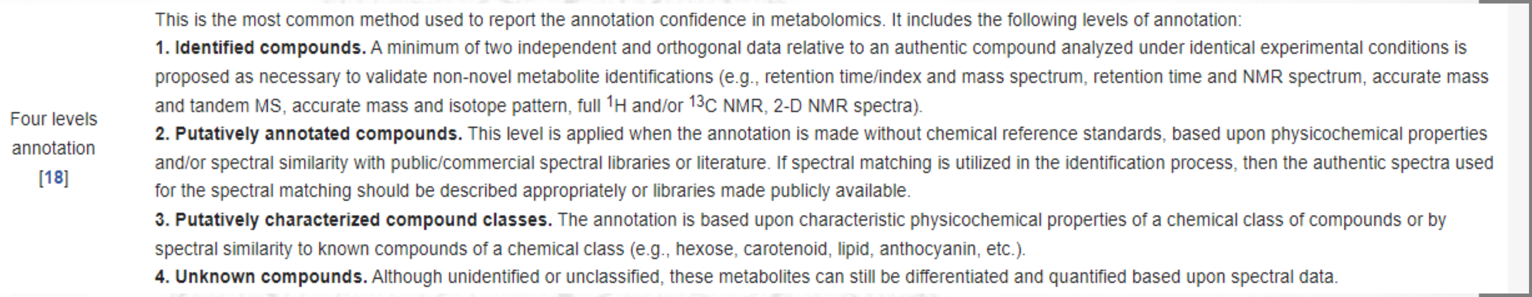
Annotation Confidence based on Levels of Annotation



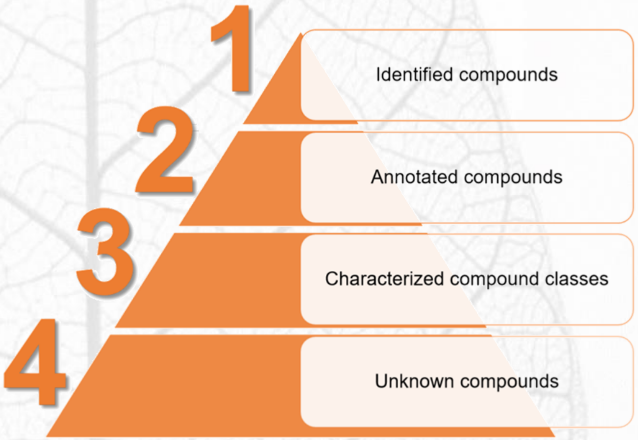
Databases

Metadata ID

FOUR LEVELS OF ANNOTATION CONFIDENCE

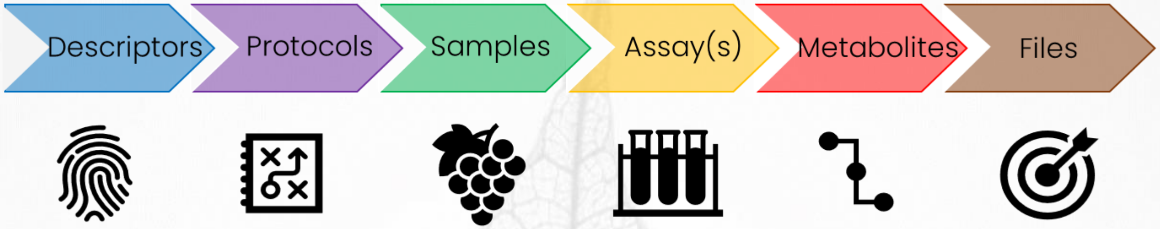


Four levels annotation [18]	This is the most common method used to report the annotation confidence in metabolomics. It includes the following levels of annotation: **1. Identified compounds.** A minimum of two independent and orthogonal data relative to an authentic compound analyzed under identical experimental conditions is proposed as necessary to validate non-novel metabolite identifications (e.g., retention time/index and mass spectrum, retention time and NMR spectrum, accurate mass and tandem MS, accurate mass and isotope pattern, full ^{1}H and/or ^{13}C NMR, 2-D NMR spectra). **2. Putatively annotated compounds.** This level is applied when the annotation is made without chemical reference standards, based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries or literature. If spectral matching is utilized in the identification process, then the authentic spectra used for the spectral matching should be described appropriately or libraries made publicly available. **3. Putatively characterized compound classes.** The annotation is based upon characteristic physicochemical properties of a chemical class of compounds or by spectral similarity to known compounds of a chemical class (e.g., hexose, carotenoid, lipid, anthocyanin, etc.). **4. Unknown compounds.** Although unidentified or unclassified, these metabolites can still be differentiated and quantified based upon spectral data.



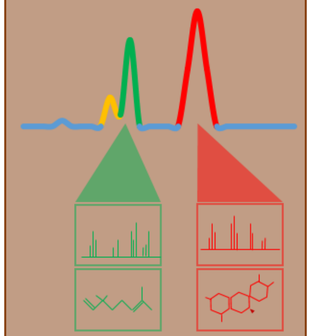
Sumner et al. 2007
Metabolomics

Schemes of
Mar Garcia Aloy



Metabolites
(Table 4)

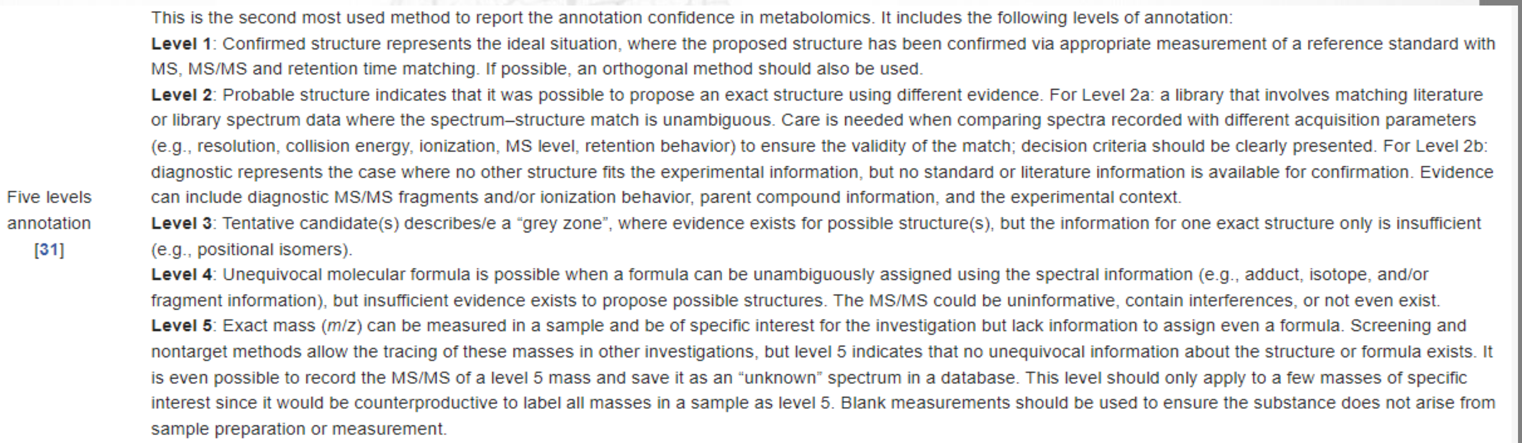
Annotation Confidence based on Levels of Annotation



Databases

Metadata ID

FIVE LEVELS OF ANNOTATION CONFIDENCE



Five levels annotation [31]

This is the second most used method to report the annotation confidence in metabolomics. It includes the following levels of annotation:

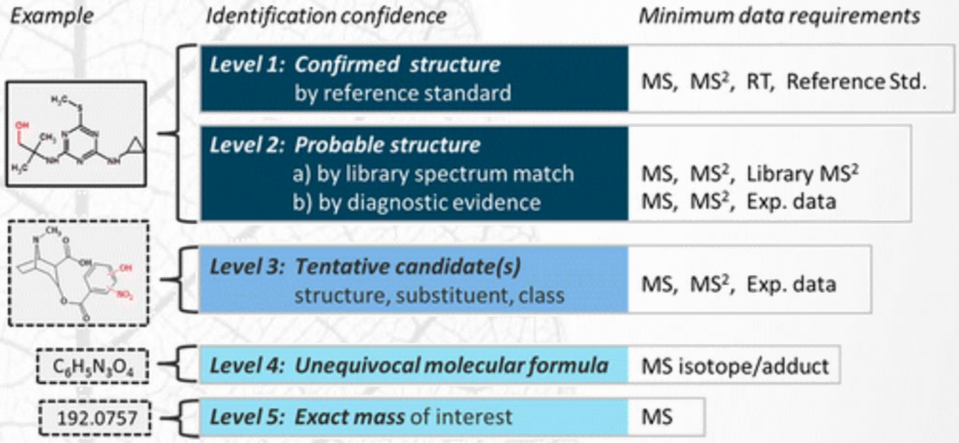
Level 1: Confirmed structure represents the ideal situation, where the proposed structure has been confirmed via appropriate measurement of a reference standard with MS, MS/MS and retention time matching. If possible, an orthogonal method should also be used.

Level 2: Probable structure indicates that it was possible to propose an exact structure using different evidence. For Level 2a: a library that involves matching literature or library spectrum data where the spectrum–structure match is unambiguous. Care is needed when comparing spectra recorded with different acquisition parameters (e.g., resolution, collision energy, ionization, MS level, retention behavior) to ensure the validity of the match; decision criteria should be clearly presented. For Level 2b: diagnostic represents the case where no other structure fits the experimental information, but no standard or literature information is available for confirmation. Evidence can include diagnostic MS/MS fragments and/or ionization behavior, parent compound information, and the experimental context.

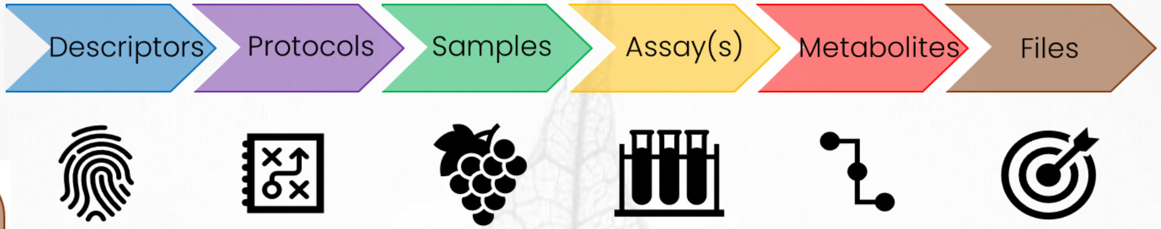
Level 3: Tentative candidate(s) describes/e a "grey zone", where evidence exists for possible structure(s), but the information for one exact structure only is insufficient (e.g., positional isomers).

Level 4: Unequivocal molecular formula is possible when a formula can be unambiguously assigned using the spectral information (e.g., adduct, isotope, and/or fragment information), but insufficient evidence exists to propose possible structures. The MS/MS could be uninformative, contain interferences, or not even exist.

Level 5: Exact mass (*m*/*z*) can be measured in a sample and be of specific interest for the investigation but lack information to assign even a formula. Screening and nontarget methods allow the tracing of these masses in other investigations, but level 5 indicates that no unequivocal information about the structure or formula exists. It is even possible to record the MS/MS of a level 5 mass and save it as an "unknown" spectrum in a database. This level should only apply to a few masses of specific interest since it would be counterproductive to label all masses in a sample as level 5. Blank measurements should be used to ensure the substance does not arise from sample preparation or measurement.


Example
Identification confidence
Minimum data requirements
Level 1: Confirmed structure by reference standard
MS, MS², RT, Reference Std.
Level 2: Probable structure
a) by library spectrum match
b) by diagnostic evidence
MS, MS², Library MS²
MS, MS², Exp. data
Level 3: Tentative candidate(s) structure, substituent, class
MS, MS², Exp. data
C6H5N3O4
Level 4: Unequivocal molecular formula
MS isotope/adduct
192.0757
Level 5: Exact mass of interest
MS

Schymansky et al. 2014
Environmental Science & Technology



Descriptors

Protocols

Samples

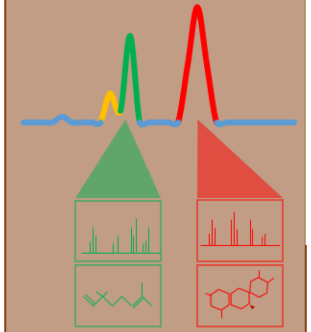
Assay(s)

Metabolites

Files

Metabolites
(Table 4)

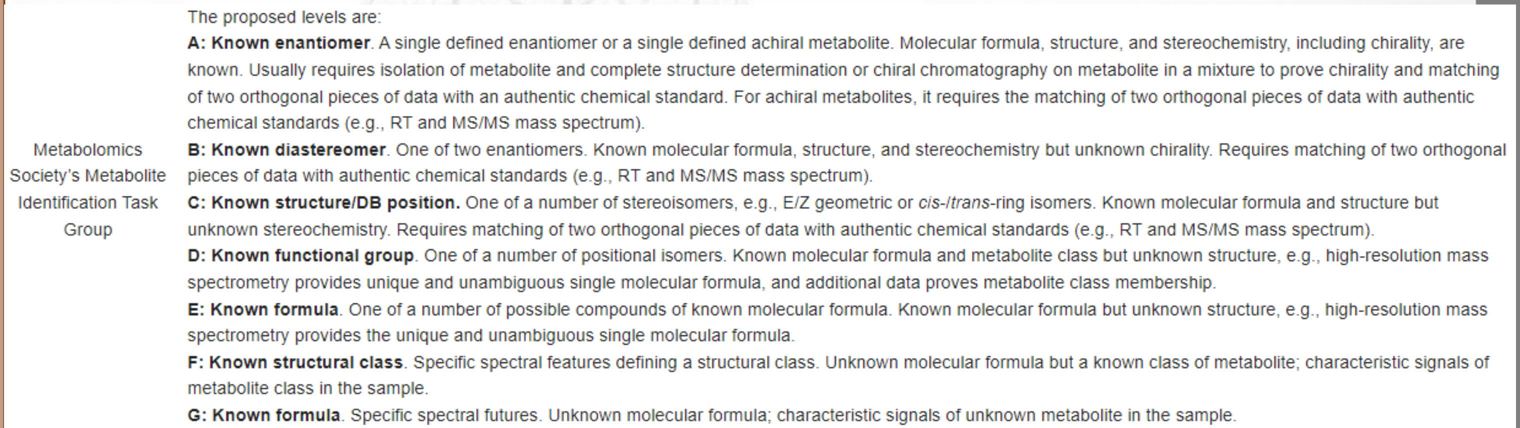
Annotation
Confidence
based on
Levels of
Annotation



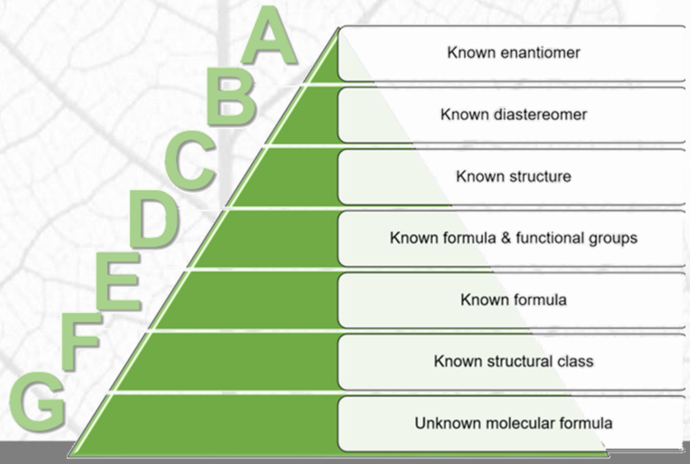
Databases

Metadata
ID

METABOLOMICA SOCIETY'S METABOLITE IDENTIFICATION TASK GROUP



Metabolomics Society's Metabolite Identification Task Group	The proposed levels are: **A: Known enantiomer**. A single defined enantiomer or a single defined achiral metabolite. Molecular formula, structure, and stereochemistry, including chirality, are known. Usually requires isolation of metabolite and complete structure determination or chiral chromatography on metabolite in a mixture to prove chirality and matching of two orthogonal pieces of data with an authentic chemical standard. For achiral metabolites, it requires the matching of two orthogonal pieces of data with authentic chemical standards (e.g., RT and MS/MS mass spectrum). **B: Known diastereomer**. One of two enantiomers. Known molecular formula, structure, and stereochemistry but unknown chirality. Requires matching of two orthogonal pieces of data with authentic chemical standards (e.g., RT and MS/MS mass spectrum). **C: Known structure/DB position.** One of a number of stereoisomers, e.g., E/Z geometric or *cis-/trans*-ring isomers. Known molecular formula and structure but unknown stereochemistry. Requires matching of two orthogonal pieces of data with authentic chemical standards (e.g., RT and MS/MS mass spectrum). **D: Known functional group**. One of a number of positional isomers. Known molecular formula and metabolite class but unknown structure, e.g., high-resolution mass spectrometry provides unique and unambiguous single molecular formula, and additional data proves metabolite class membership. **E: Known formula**. One of a number of possible compounds of known molecular formula. Known molecular formula but unknown structure, e.g., high-resolution mass spectrometry provides the unique and unambiguous single molecular formula. **F: Known structural class**. Specific spectral features defining a structural class. Unknown molecular formula but a known class of metabolite; characteristic signals of metabolite class in the sample. **G: Known formula**. Specific spectral futures. Unknown molecular formula; characteristic signals of unknown metabolite in the sample.



Known enantiomer

Known diastereomer

Known structure

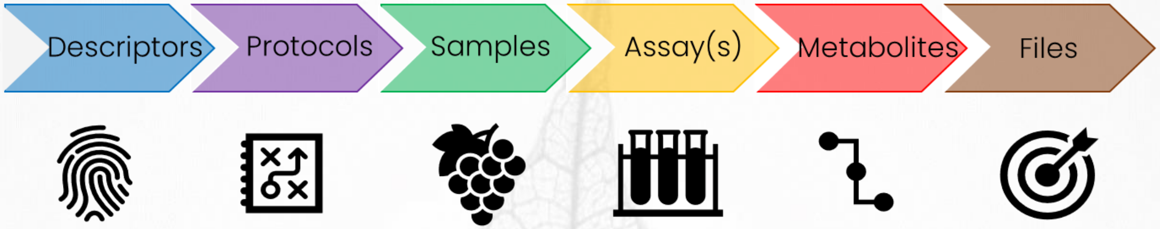
Known formula & functional groups

Known formula

Known structural class

Unknown molecular formula

Schemes of
Mar Garcia Aloy



Descriptors

Protocols

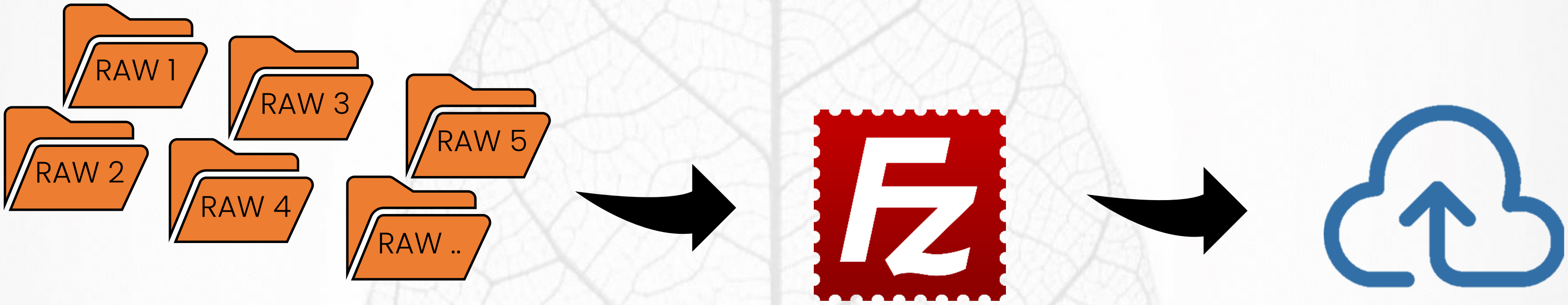
Samples

Assay(s)

Metabolites

Files

Target achieved



Thanks for your attention

CONTACT:
stefania.savoi@unito.it
savoi.stefania@gmail.com