LC-MS metabolomics: from data extraction to system level integration

Pietro Franceschi

Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Trento

pietro.franceschi@fmach.it

Duccio Cavalieri

Dipartimento di Biologia. Università degli Studi di Firenze

FONDAZIONE EDMUND MACH

Metabolomics @ FEM

- Food, plants (grapevine, apple, soft fruits)
- Untargeted LC/GC-HR-MS, Targeted LC/GC - MS
- NMR (600, 400)
- Data processing and statistical analysis (stat modeling, chemometrics, machine learning)

WHAT is Metabolomics

The objective of metabolomics is to characterize, in the most complete and comprehensive way, the pool of small molecules which are the end product of the metabolism.
The pool of these molecules is known as metabolome.
Metabolomics aims at measuring the metabolites in a quantitative way and to characterize their relations and associations.

Targeted

Untargeted

FONDAZIONE EDMUND MACH

WHY Metabolomics

- For the quest of molecular markers (e.g. nutrition, health, ...)
- To perform molecular phenotyping (e.g. personalized medicine)
- To associate a gene to its function (e.g. to support breeding)
- To study the chemical interaction in complex systems (e.g. ecological interaction)
- To enforce our understanding of metabolism
- ...

FONDAZIONE EDMUND MACH

HOW Metabolomics

We need an analytical technique:

- Sensitive to the chemical structure
- Universal - *able to see almost all classes of molecules*
- Sensitive - *able to see metabolites with low concentrations*
- With an high dynamic range - *able to measure at the same time trace and high abundant compounds*

XY-MS(MS)

NMR

FONDAZIONE EDMUND MACH

Chemical Challenges

- Size of the metabolic chemical space

  The molecules included in the metabolome can be extremely diverse in particular if plants/microorganisms are concerned

- Diversity in the chemical properties of the metabolites

  The chemical diversity results in different properties which would require different methods of analysis

- Huge differences in concentrations

  Some of the metabolites are present in high concentrations, while other highly relevant compounds are present only in small traces. Our analytical method should be able to see them at the same time

FONDAZIONE EDMUND MACH

Practical Challenges

- Pre-analytical

  Sample collection, storage, handling in particular outside a research environment
- Analytical

  Analytical drifts, analytical compromise
- Data treatment/analysis

  Large datasets (1GB/sample), huge number of variables (~10000)
- Annotation

  In untargeted metabolomics we do not measure metabolites, but we want them …
- (Statistical) Validation of the outcomes

  Can I trust in what I see?

FONDAZIONE EDMUND MACH

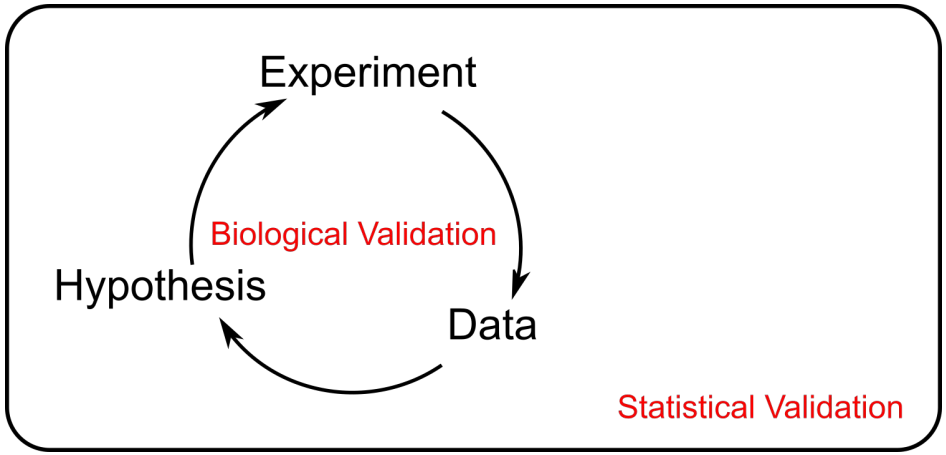
Thoughts on validation

Statistical Validation

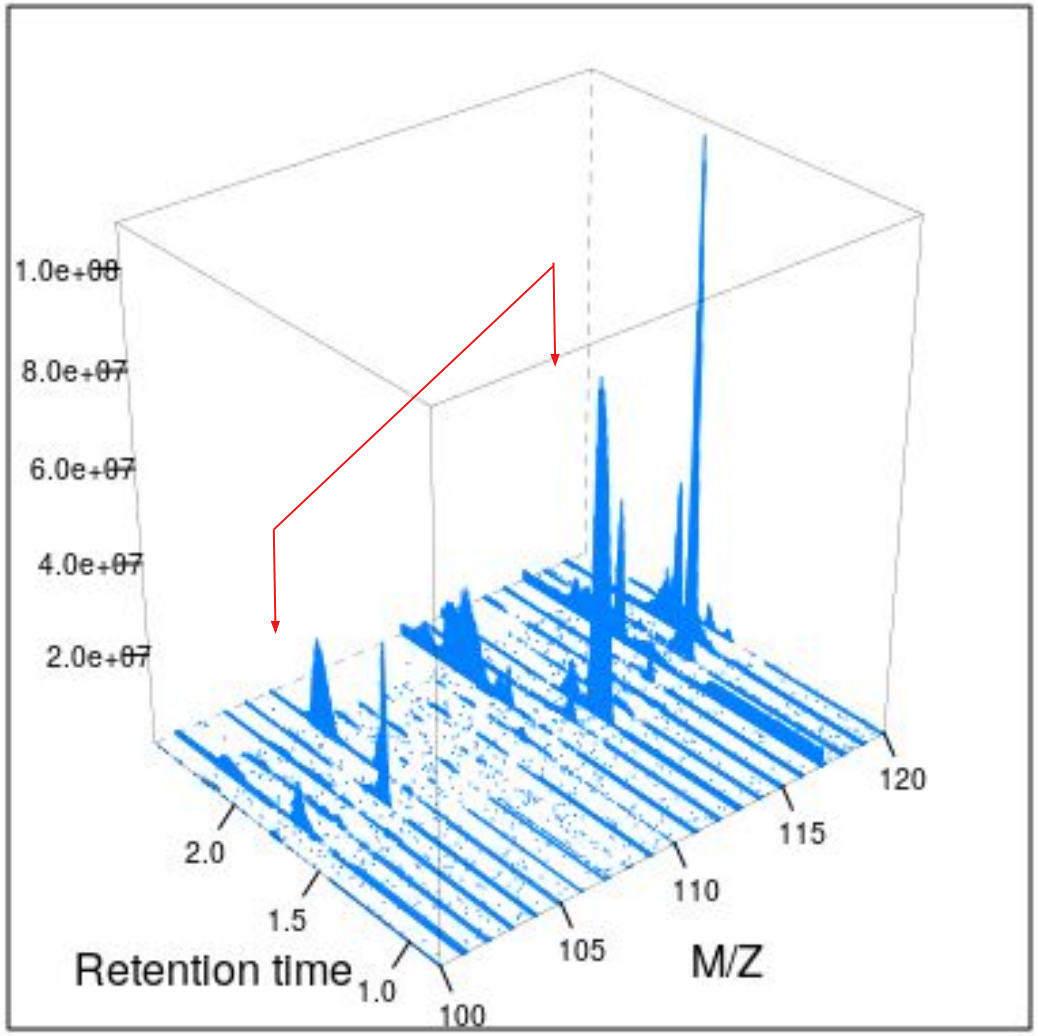
Can I draw (induce!) reliable general rules from the result of my study?

Biological Validation

Is what I'm getting reasonably fitting within the established body of knowledge?



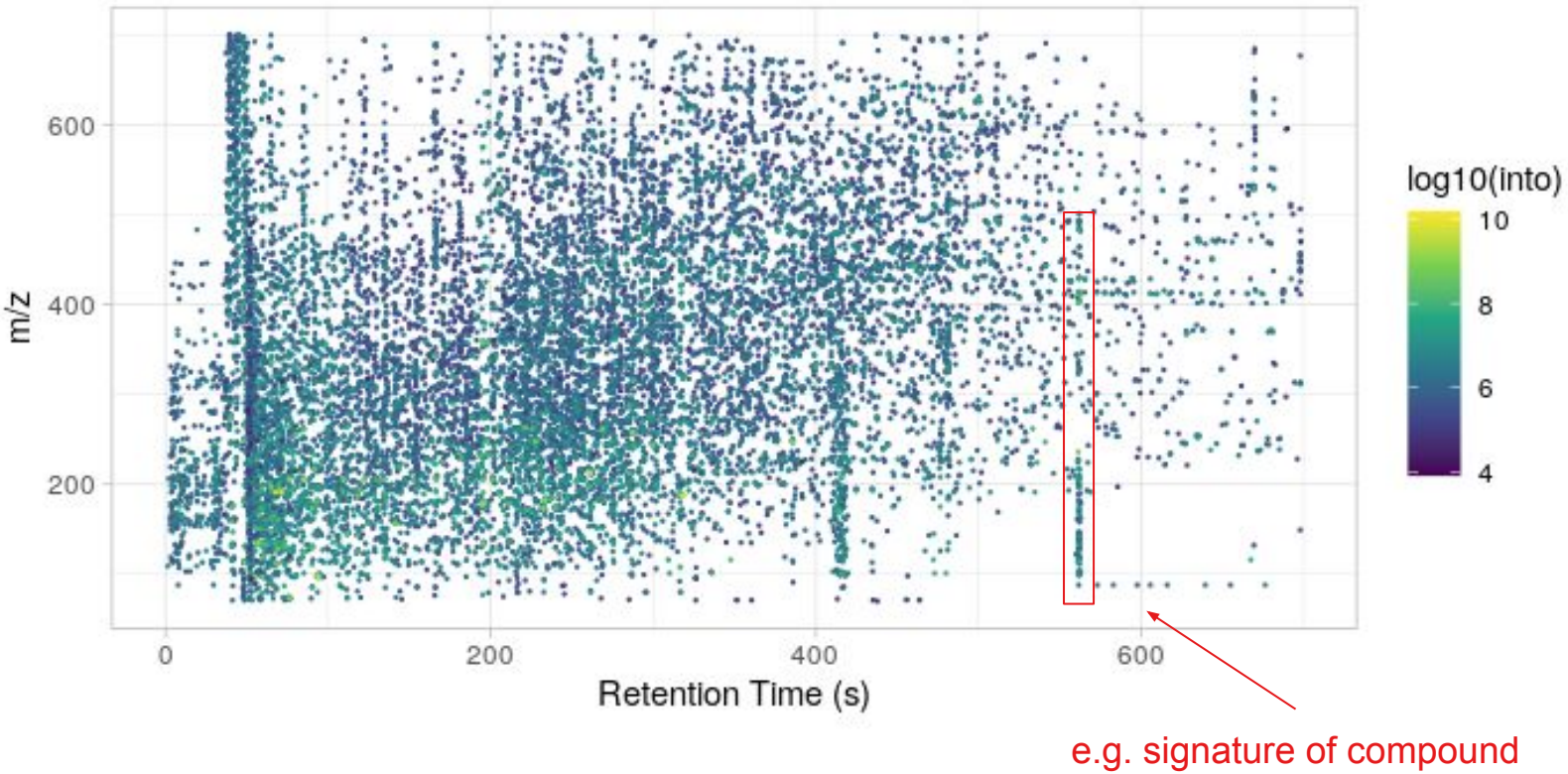
What we measure in LC-MS



- Peaks on the ionic traces
- Two peaks for the same molecule/metabolite

FONDAZIONE EDMUND MACH

What we measure in LC-MS



- One sample peak map
- Each dot is a peak (red arrows of the previous plot)
- ~16000 peaks

FONDAZIONE EDMUND MACH

Important points

- Each dot is an ion showing a chromatographic peak.
  In its extracted ion trace
- Each dot is not a metabolite.
  A neutral gives more ions
- The dots associated to the same metabolite shows up as vertical lines.
- Coelution blurs the vertical lines.
  We do not see only vertical structures ...
- Each sample gives a slightly different map.
  Samples are different for analytical and biological reasons

FONDAZIONE EDMUND MACH

Data Analysis workflow

1. For each sample, extract the map of dots
2. Align the maps of the different samples in a consensus map of features
3. Correct for analytical drifts
4. Create a data matrix
5. Mine that matrix highlighting the most relevant information (univariate statistics, multivariate analysis, machine learning)

FONDAZIONE EDMUND MACH

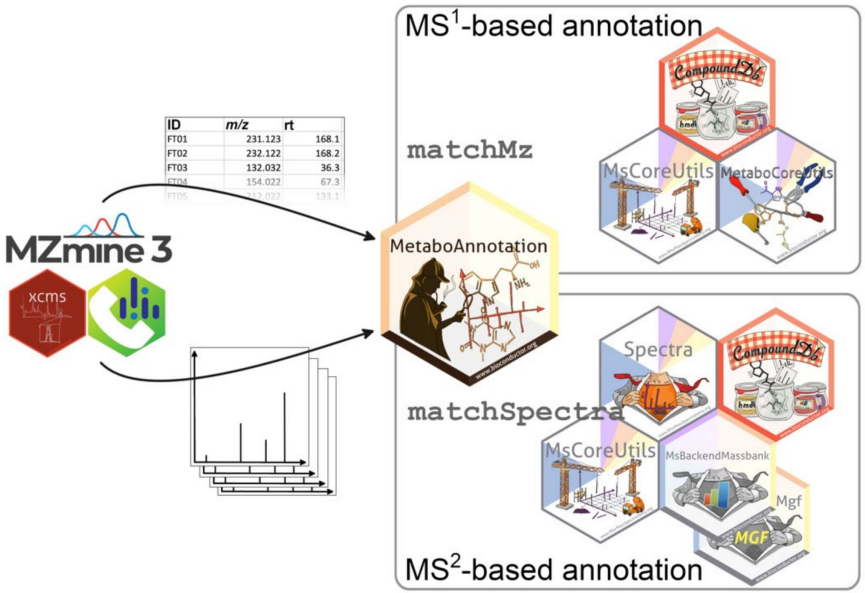
Preprocessing solutions

1. Commercial solutions (e.g. vendor software)
2. Open-source scripts and packages
   a. xcms (R)
   b. OpenMS (Python)
3. Open Source Desktop Applications
   a. OpenMS
   b. MzMine
4. Web-based applications
   a. MetaboAnalyst
   b. Workflow4Metabolomics

Perspective …

FONDAZIONE EDMUND MACH

Focus on `xcms`



Open Access Article

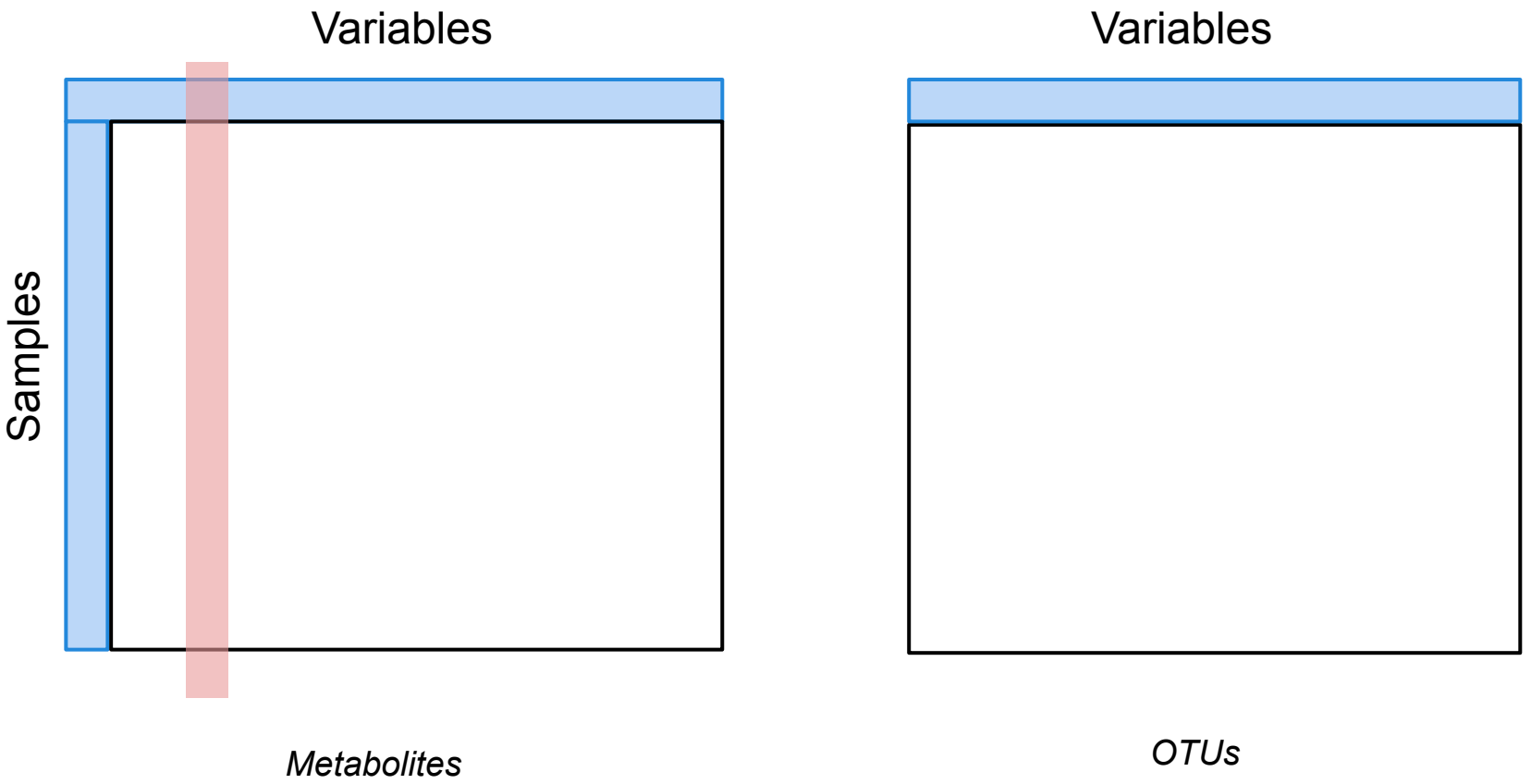
A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R

by Johannes Rainer [1,*], Andrea Vicini [1], Liesa Salzer [2], Jan Stanstrup [3], Josep M. Badia [4,5], Steffen Neumann [6,7], Michael A. Stravs [8,9], Vinicius Verri Hernandes [1,10], Laurent Gatto [11], Sebastian Gibb [12] and Michael Witting [13,14]

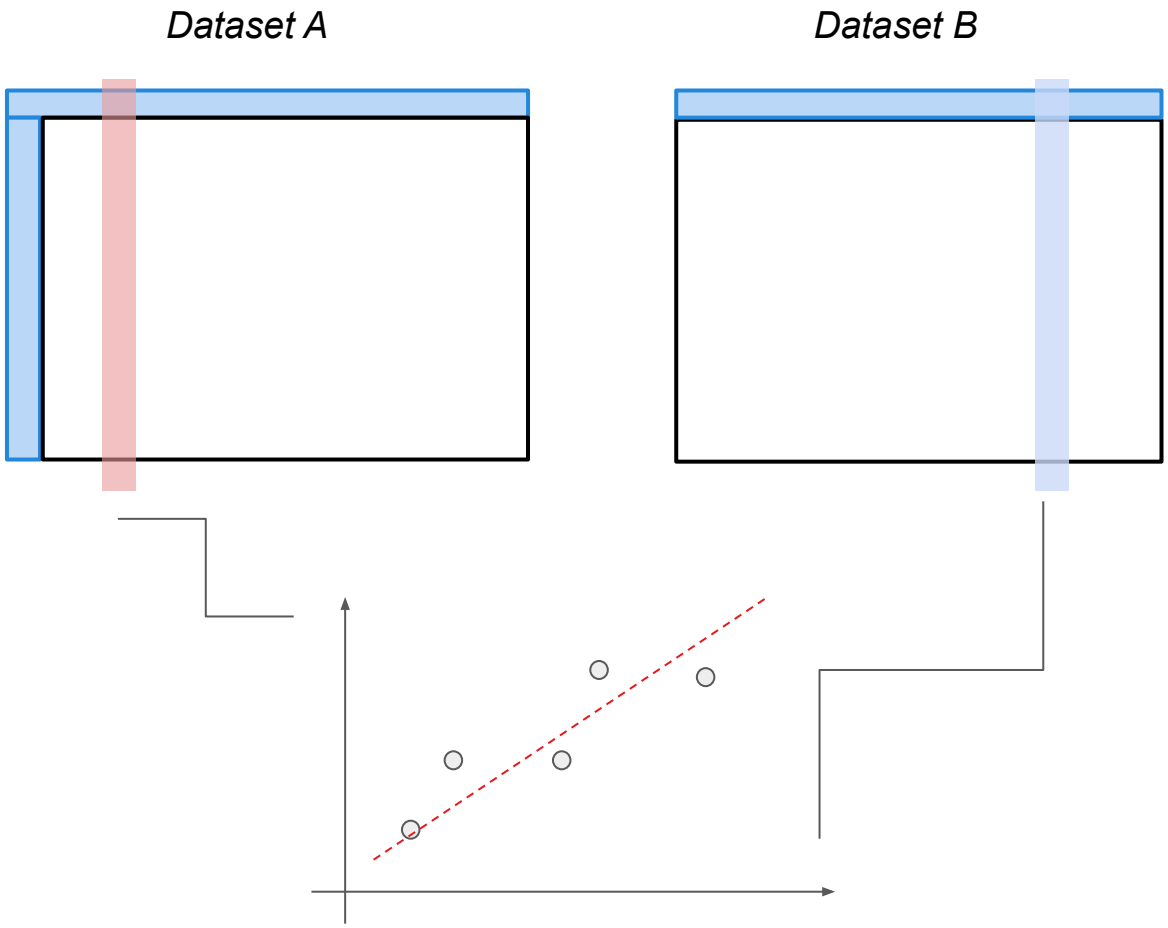
Rainer, J et al. . A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites* **2022**, *12*, 173. https://doi.org/10.3390/metabo12020173

FONDAZIONE EDMUND MACH

Data fusion:
more than one data matrix



Statistical Dependence

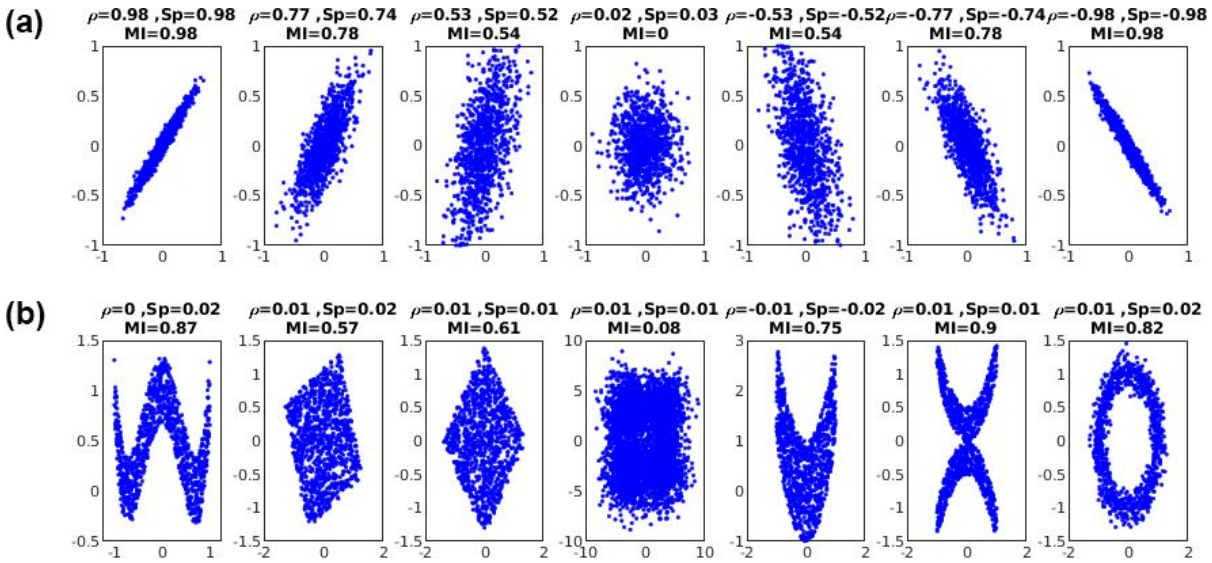


Pearson Correlation, Rank Correlation, Mutual Information

Flexibility

FONDAZIONE EDMUND MACH

different degrees of Flexibility

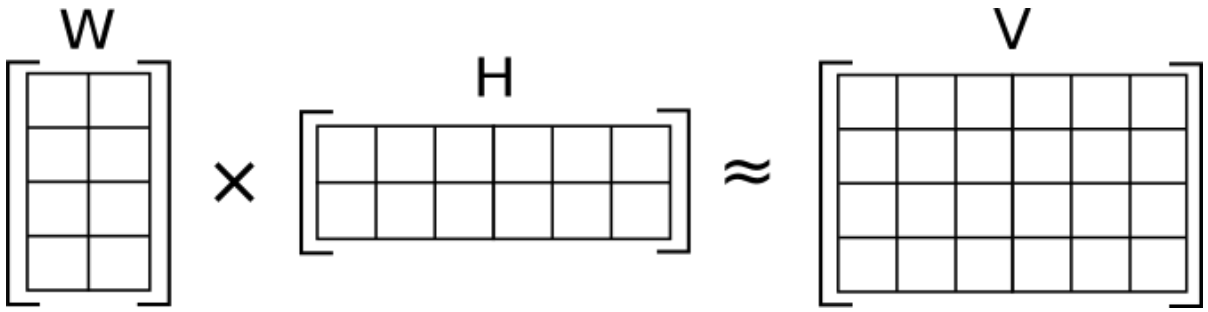


Atmos. Chem. Phys., 18, 12699–12714, 2018 https://doi.org/10.5194/acp-18-12699-2018

Davide Albanese, Samantha Riccadonna, Claudio Donati, Pietro Franceschi, A practical tool for maximal information coefficient analysis, *GigaScience*, Volume 7, Issue 4, April 2018, giy032, https://doi.org/10.1093/gigascience/giy032

Linear associations

Methods relying on Matrix algebra and decomposition rely on the presence of linear associations between the variables

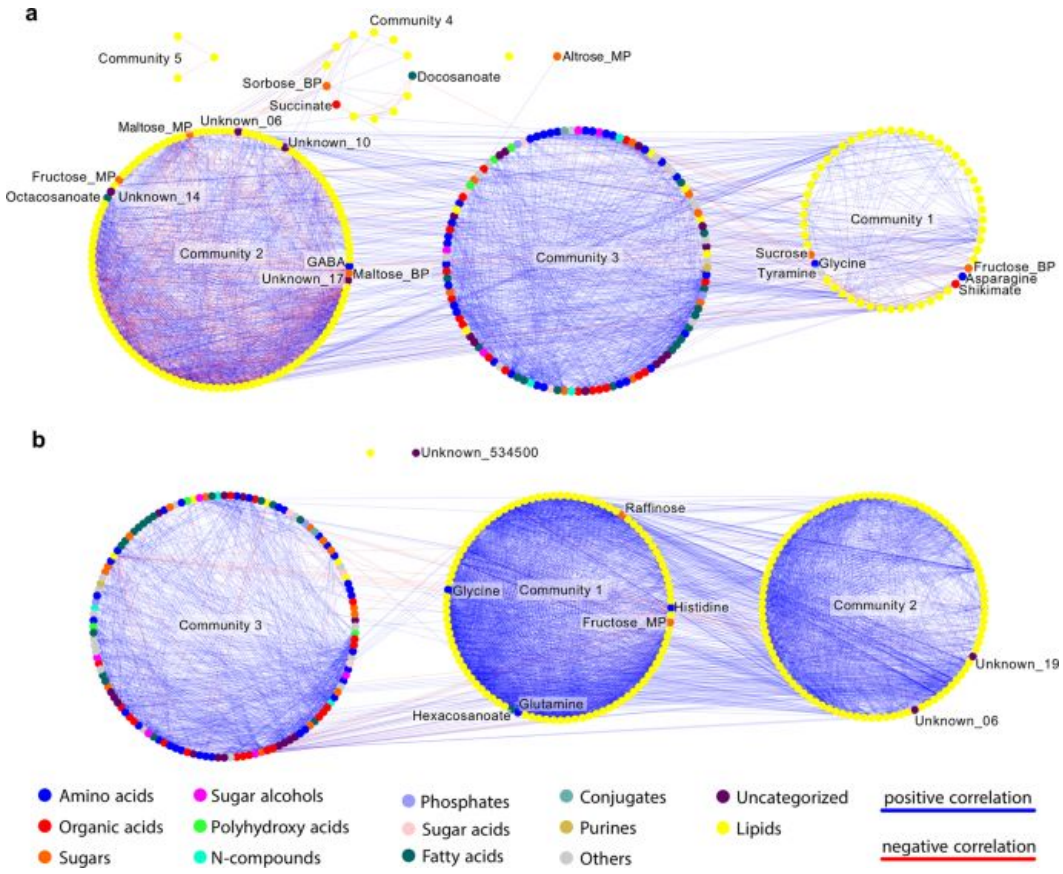


1. PLS (Partial Least Squares Regression)
2. Co-Inertia analysis
3. Multi block methods
4. JIVE (Joint and Individual Variation Explained)
5. …

FONDAZIONE EDMUND MACH

Association Networks

All measures of association can be used to construct association networks



Toubiana, D., Sade, N., Liu, L. *et al.* C *Sci Rep* 10, 4489 (2020). https://doi.org/10.1038/s41598-020-61081-4

FONDAZIONE EDMUND MACH

Critical Points

- "Being associated" is a transitive relationship. If A is associated with B, and B with C … A is also associated with C
- Some of the association we find can be false positives, i.e. the results of the chance alone
- Association is not causation
- Many "genomic" (or ecological) are compositional (relative abundances) …
- When we can call an association "significant"

FONDAZIONE EDMUND MACH

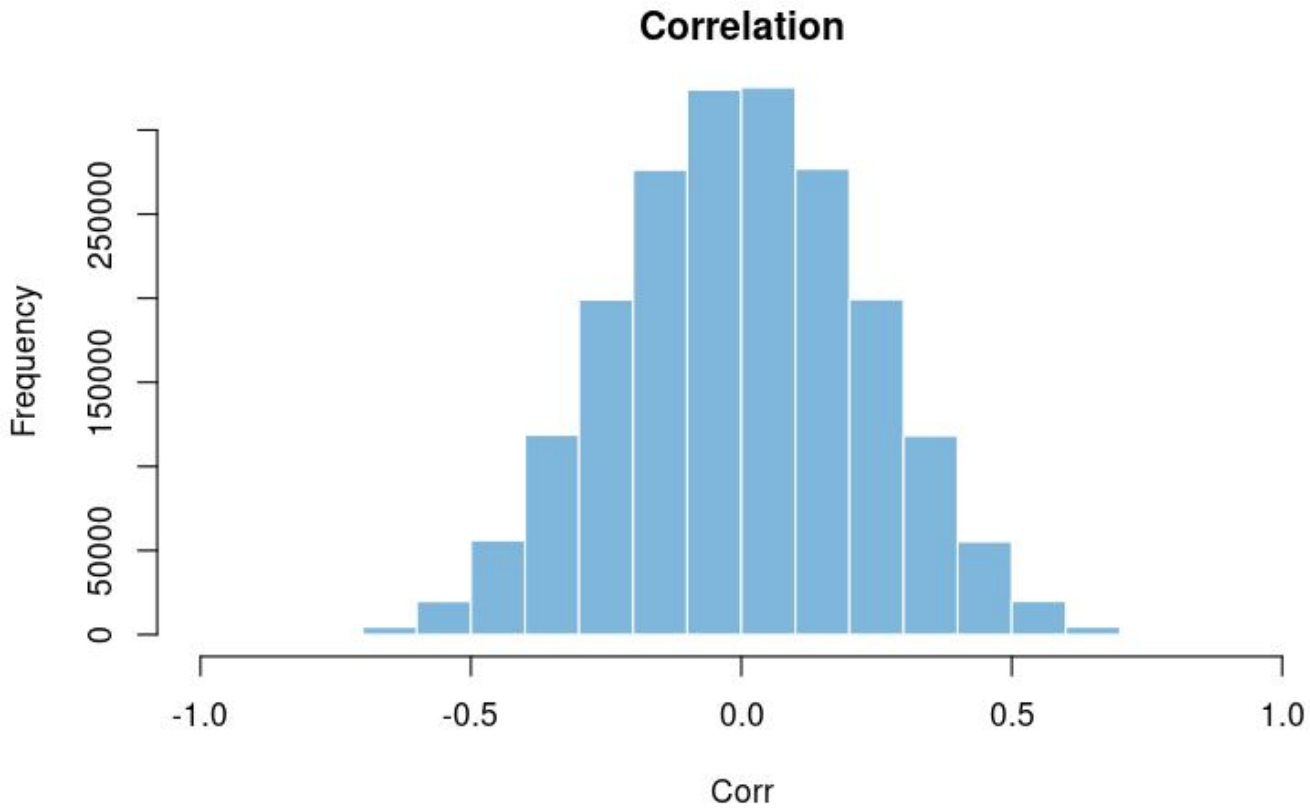
Critical Points

- Do we expect different omics layers to be in synchrony (e.g. metabolome and transcriptome) … time course
- Complex association patterns have to be characterised with large sample cohorts… non trivial for untargeted MS based metabolomics
- “Analytical” correlation is stronger than the biological or ecological one …

FONDAZIONE EDMUND MACH

Random Associations

20 samples, 1000 variables … random numbers



range: -0.9 - 0.86

FONDAZIONE EDMUND MACH

On causation …

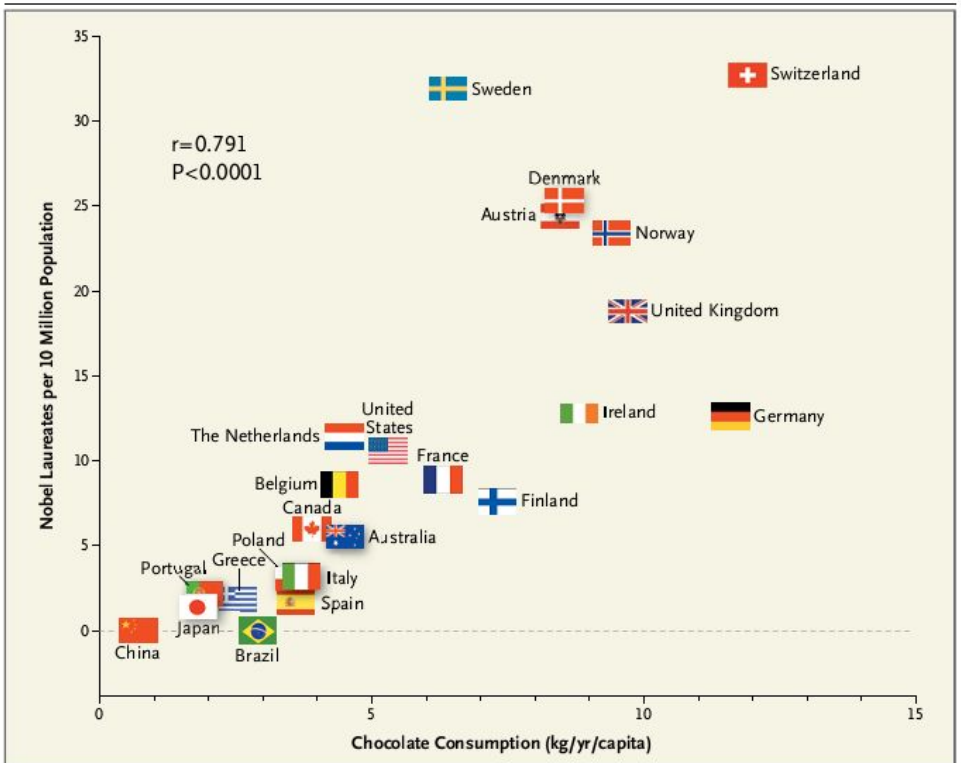


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

The slope of the regression line allows us to estimate that it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1.

N Engl J Med 2012; 367:1562-1564 DOI: 10.1056/NEJMon1211064

FONDAZIONE EDMUND MACH

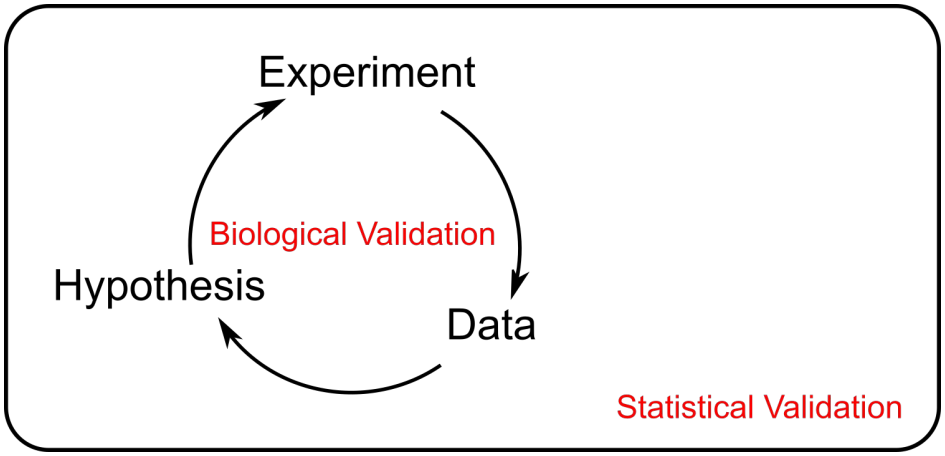
Validation is Critical

Statistical Validation

New samples, new experiments

Knowledge driven validation

Is what I'm getting reasonably fitting within the established body of knowledge?



FONDAZIONE EDMUND MACH

Metabolomics is …

1. Fun!
2. Challenging, but powerful
3. Multidisciplinary
4.