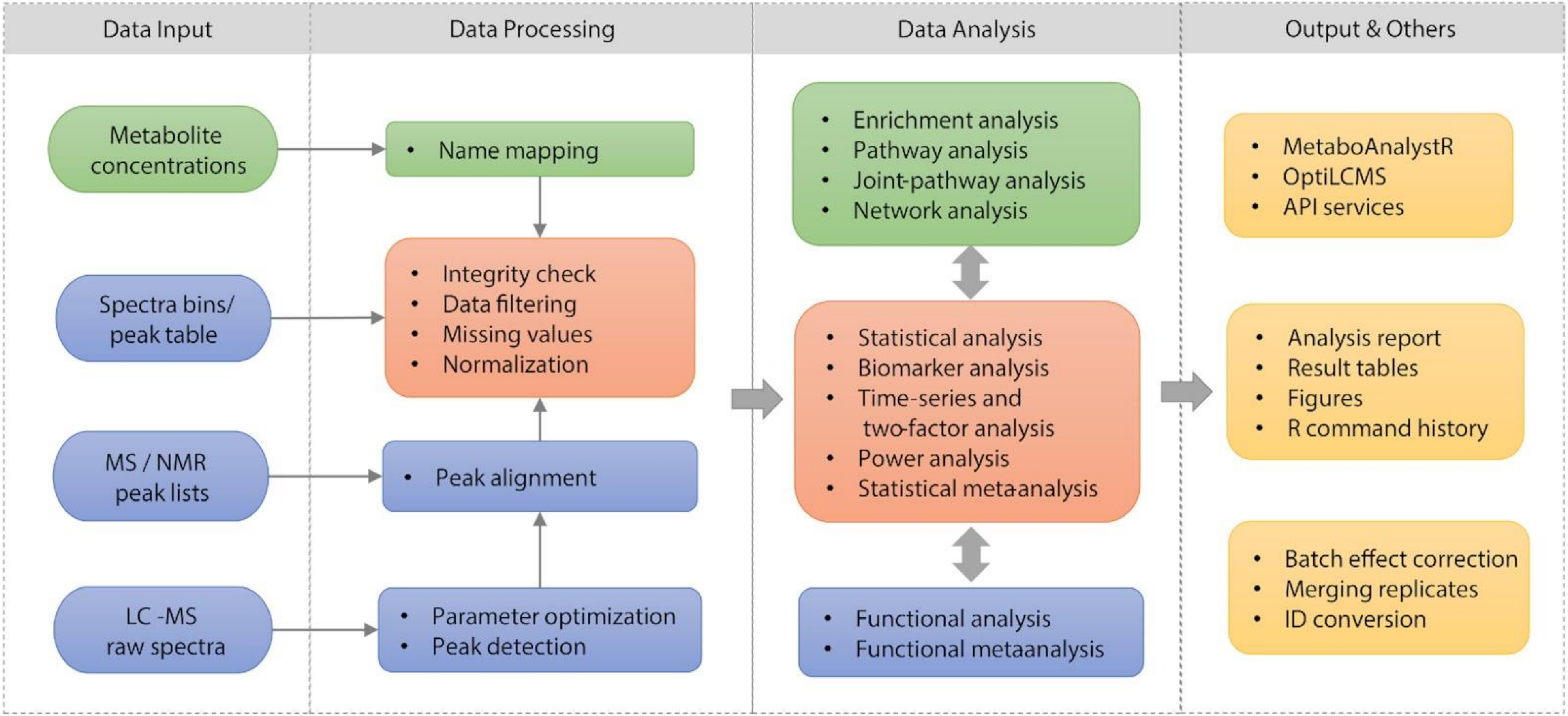
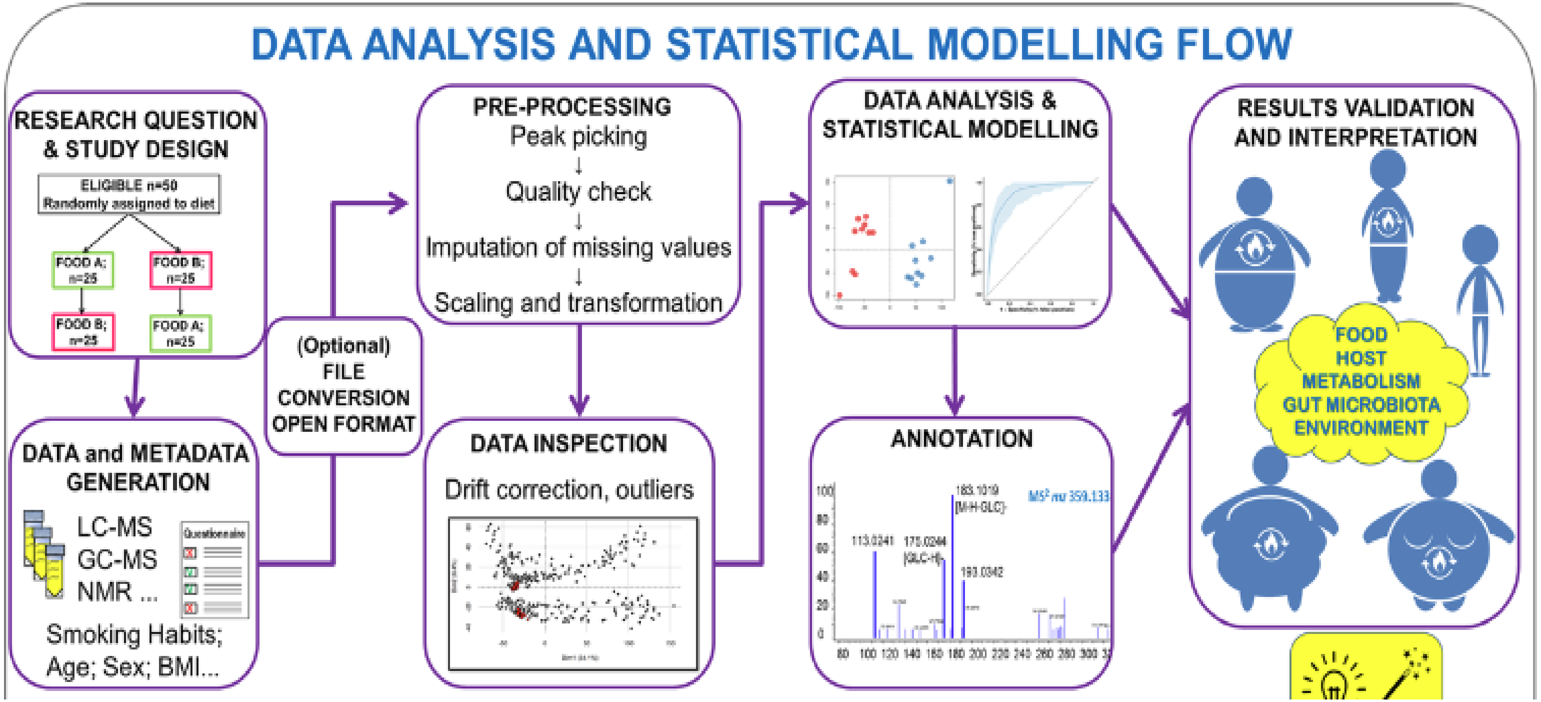
Many Data Analysis Pipelines

Anna Guadall
Alex Sánchez

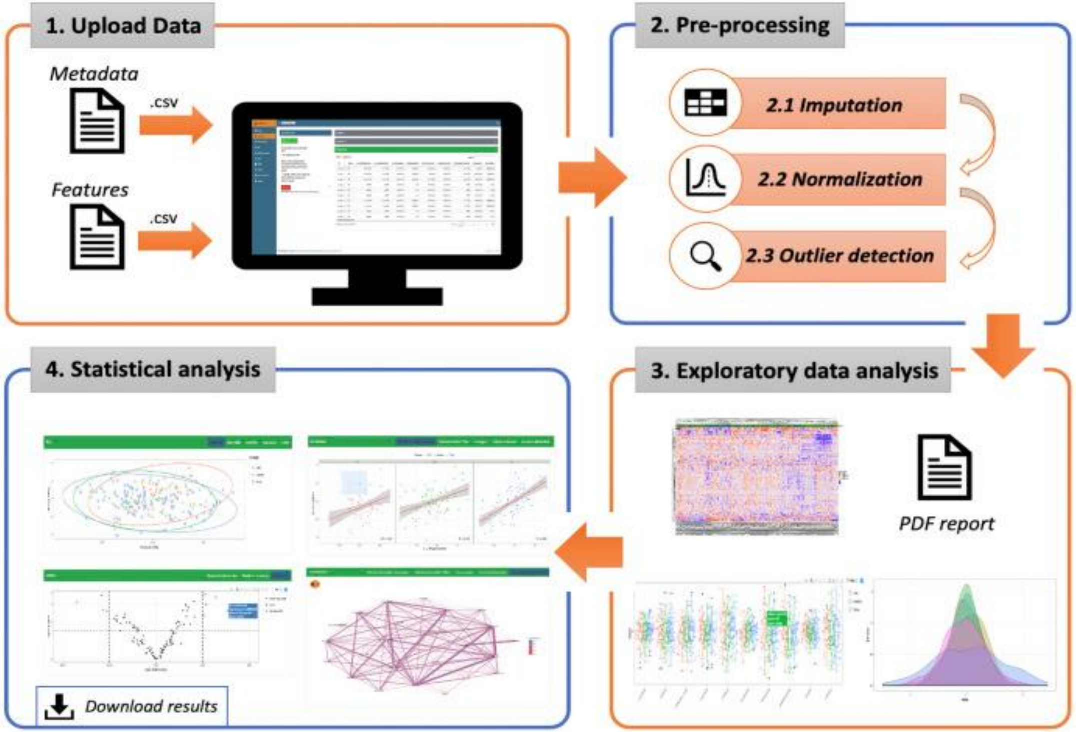


Overview of MetaboAnalyst v5.0 workflows. Steps for targeted metabolomics are indicated by boxes in green, steps for untargeted metabolomics are in blue, and those in orange can be used for both. Experienced users can use various utility functions or install the corresponding R packages (yellow boxes) to perform analysis beyond those pre-defined regular workflows.


DATA ANALYSIS AND STATISTICAL MODELLING FLOW
RESEARCH QUESTION & STUDY DESIGN
ELIGIBLE n=50
Randomly assigned to diet
FOOD A; n=25
FOOD B; n=25
FOOD B; n=25
FOOD A; n=25
DATA and METADATA GENERATION
LC-MS
GC-MS
NMR ...
Questionnaire
Smoking Habits; Age; Sex; BMI...
(Optional) FILE CONVERSION OPEN FORMAT
PRE-PROCESSING
Peak picking
Quality check
Imputation of missing values
Scaling and transformation
DATA INSPECTION
Drift correction, outliers
DATA ANALYSIS & STATISTICAL MODELLING
ANNOTATION
MS² m/z 359.133
183.1019
[M-H-GLC]-
113.0241
175.0244
[GLC-H]-
193.0342
RESULTS VALIDATION AND INTERPRETATION
FOOD
HOST
METABOLISM
GUT MICROBIOTA
ENVIRONMENT

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

Marynka M. Ulaszewska, Christoph H. Weinert, Alessia Trimigno, Reto Portmann,



POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis

Pol Castellano-Escuder, Conceptualization, Software, Writing – original draft,[1,2,3,*] Raúl González-Domínguez, Conceptualization, Writing – review & editing,[1,3] Francesc Carmona-Pontaque, Conceptualization, Writing – review & editing,[2,3] Cristina Andrés-Lacueva, Funding acquisition, Supervision, Writing – review & editing,[1,3] and Alex Sánchez-Pla, Conceptualization, Supervision, Writing – review & editing[2,3,*]

Centering, scaling, and transformations: improving the biological information content of metabolomics data

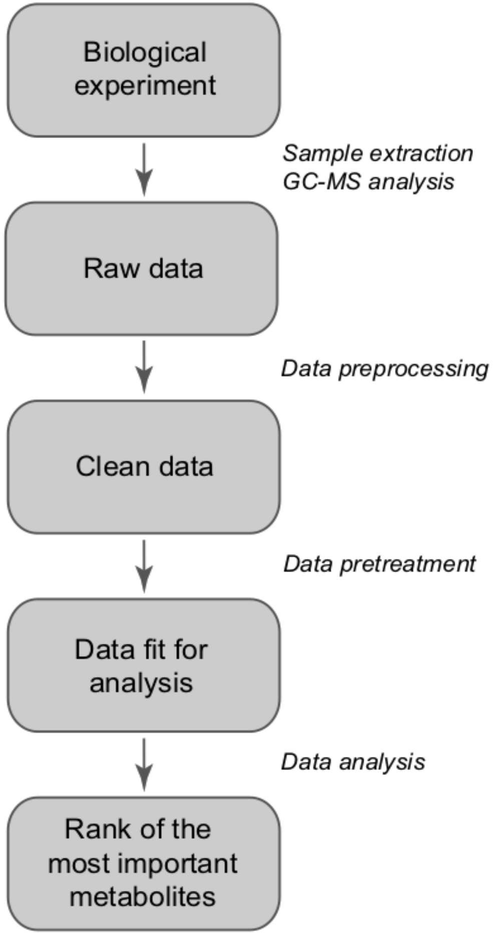
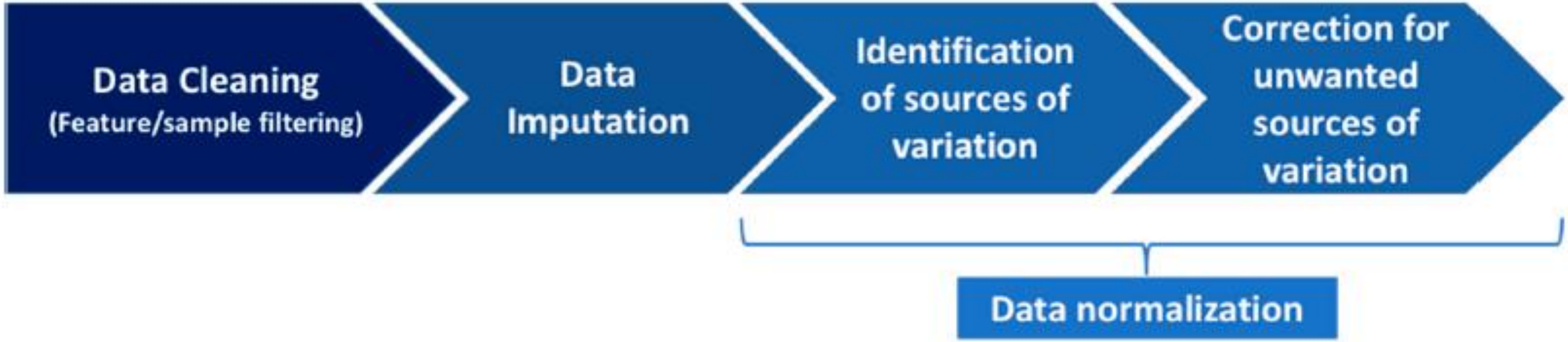


Figure 1
The different steps between biological sampling and ranking of the most important metabolites.



Data Cleaning (Feature/sample filtering)

Data Imputation

Identification of sources of variation

Correction for unwanted sources of variation

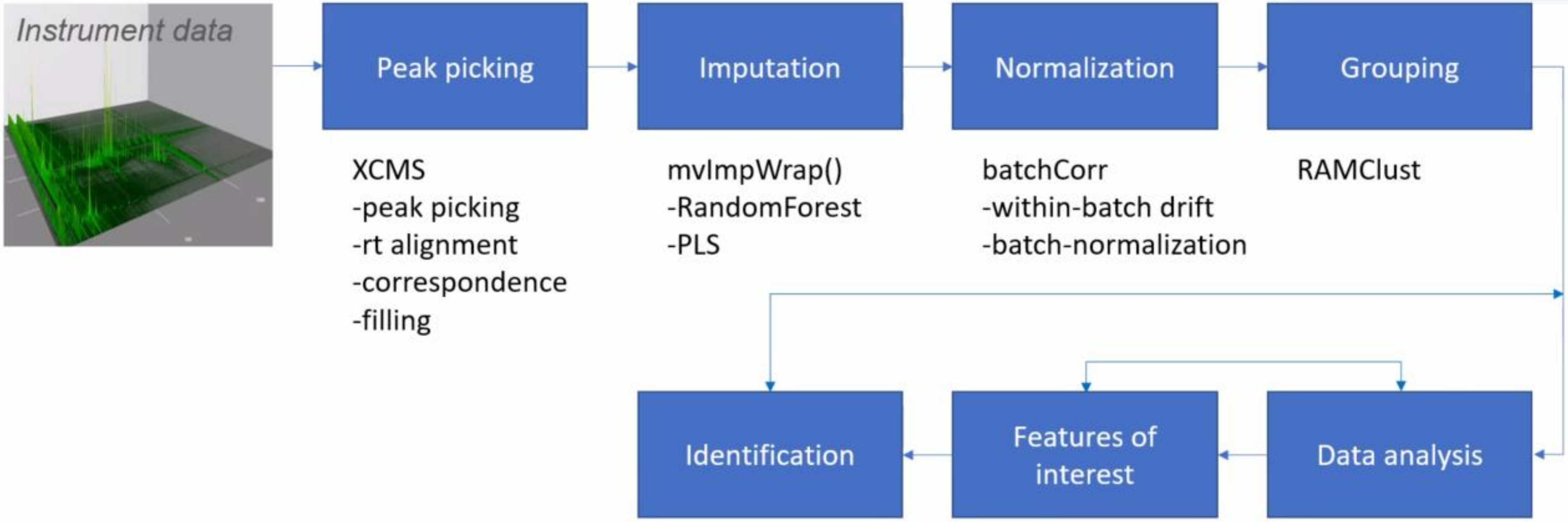
Data normalization

A New Pipeline for the Normalization and Pooling of Metabolomics Data

Vivian Viallon [1,*], Mathilde His [1], Sabina Rinaldi [1], Marie Breeur [1], Audrey Gicquiau [1], Bertrand Hemon [1],

Processing pipeline

Hablando: Carl Brunius



XCMS
-peak picking
-rt alignment
-correspondence
-filling

mvImpWrap()
-RandomForest
-PLS

batchCorr
-within-batch drift
-batch-normalization

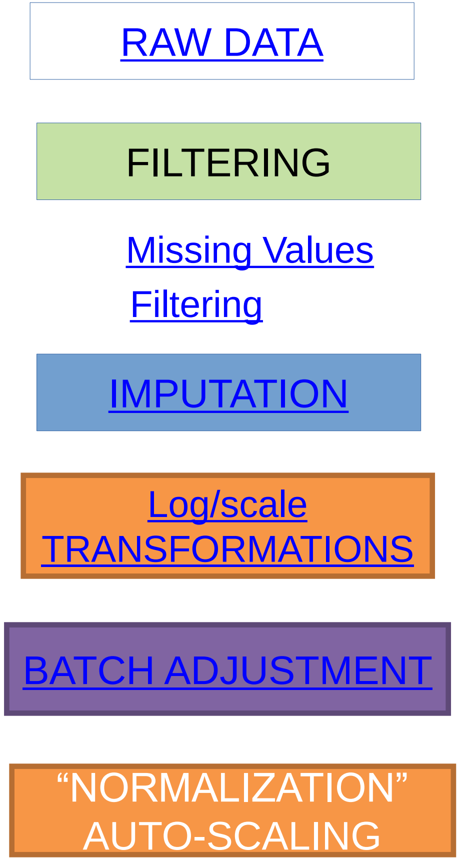
RAMClust

-MSMS
-auth standards
-database matching
-in silico

-minimal-optimal
-all-relevant

-multivariate
-"univariate"
-network
-epidemiology

Se parecen pero no son lo mismo



RAW DATA

FILTERING

Missing Values Filtering

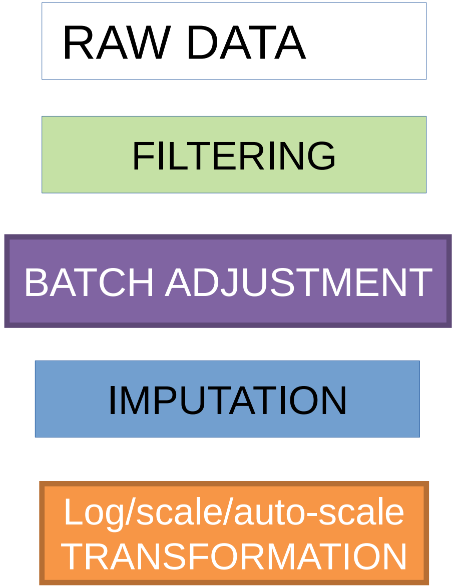
IMPUTATION

Log/scale TRANSFORMATIONS

BATCH ADJUSTMENT

“NORMALIZATION” AUTO-SCALING

DiGuMet



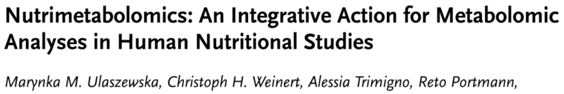
RAW DATA

FILTERING

BATCH ADJUSTMENT

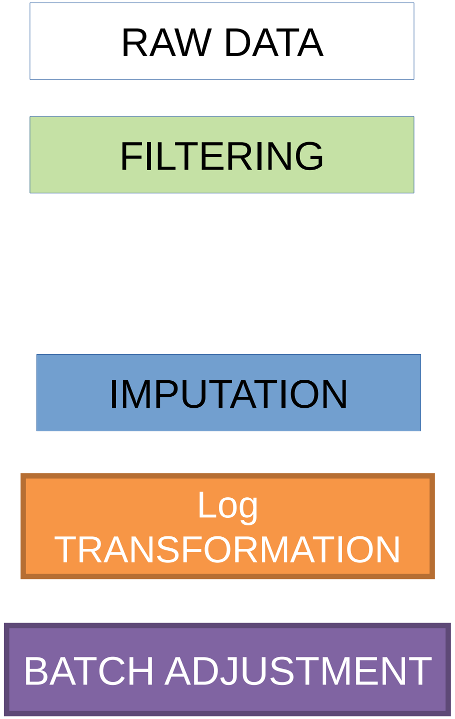
IMPUTATION

Log/scale/auto-scale TRANSFORMATION



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

Marynka M. Ulaszewska, Christoph H. Weinert, Alessia Trimigno, Reto Portmann,



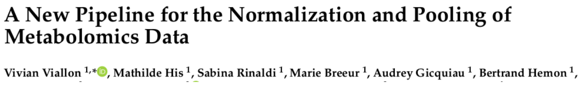
RAW DATA

FILTERING

IMPUTATION

Log TRANSFORMATION

BATCH ADJUSTMENT



A New Pipeline for the Normalization and Pooling of Metabolomics Data

Vivian Viallon [1,*], Mathilde His [1], Sabina Rinaldi [1], Marie Breeur [1], Audrey Gicquiau [1], Bertrand Hemon [1],

Raw data

- Concentrations
- Approx. Concentrations (relative to)
- Peak ratios?

⌂

Missing values

- No result
- Values out of limits of detection

Filtering

- Missing values
  - Missigness threshold?
  - Missings to be considered?
    - "Fully missing values"?
    - Values out of limits of detection?

- Less informative metabolites/samples
  - Variance?

- Outliers
  - IQR?
  - PCA?

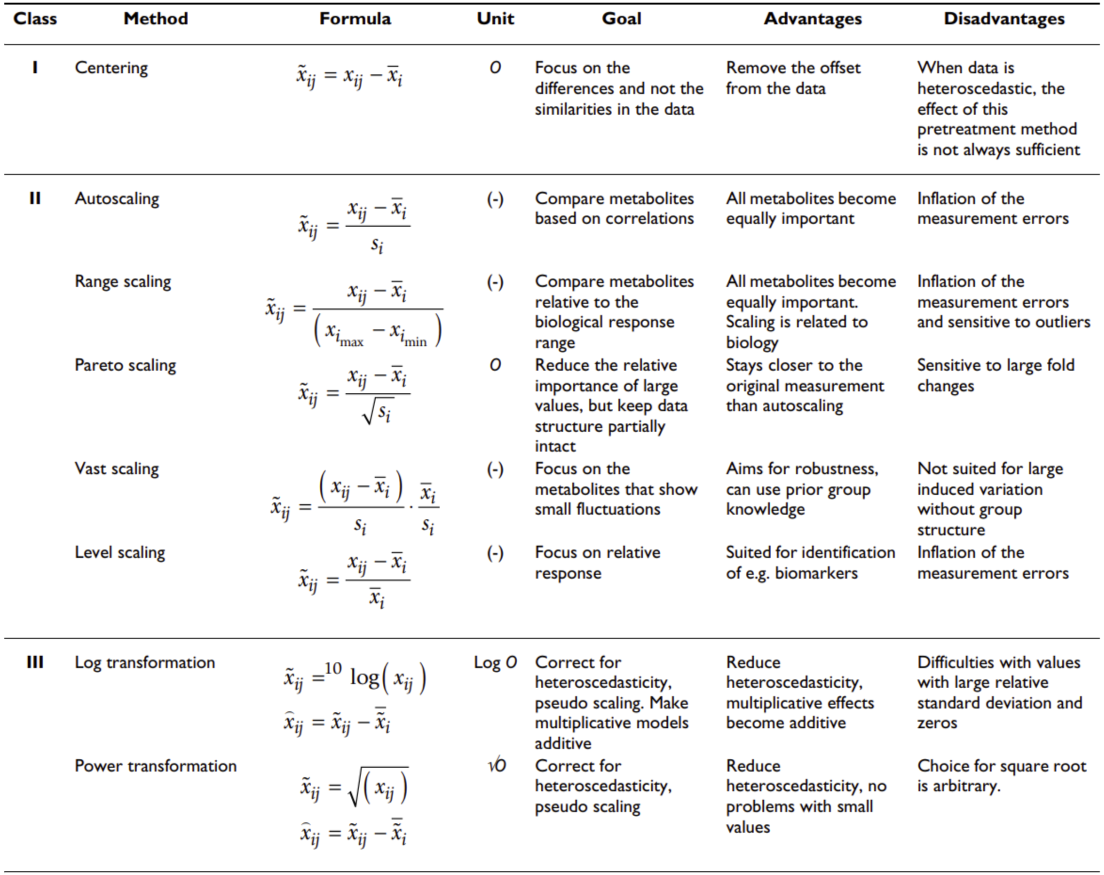
Imputation

- Missing values
    - All of them, including out of limits of detection
    - Endogen/exogen?

- Method
    - K-NN
    - Batch-specific median
    - Below lower limit of detection --> LLOD/2
    - Above upper limit of detection --> ULOD
    - zero

Transformations

- Which transformations
- When should data be transformed
  - Before/after filter/imputation
  - Before/after batch adjustment
- Normalization is transformation?



Class	Method	Formula	Unit	Goal	Advantages	Disadvantages
I	Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	*O*	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
II	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
	Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{\max}} - x_{i_{\min}} \right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
	Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	*O*	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
	Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i \right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
	Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
III	Log transformation	$\tilde{x}_{ij} = {}^{10}\log\left(x_{ij} \right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log *O*	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
	Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij} \right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√*O*	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

Identification of batch sources

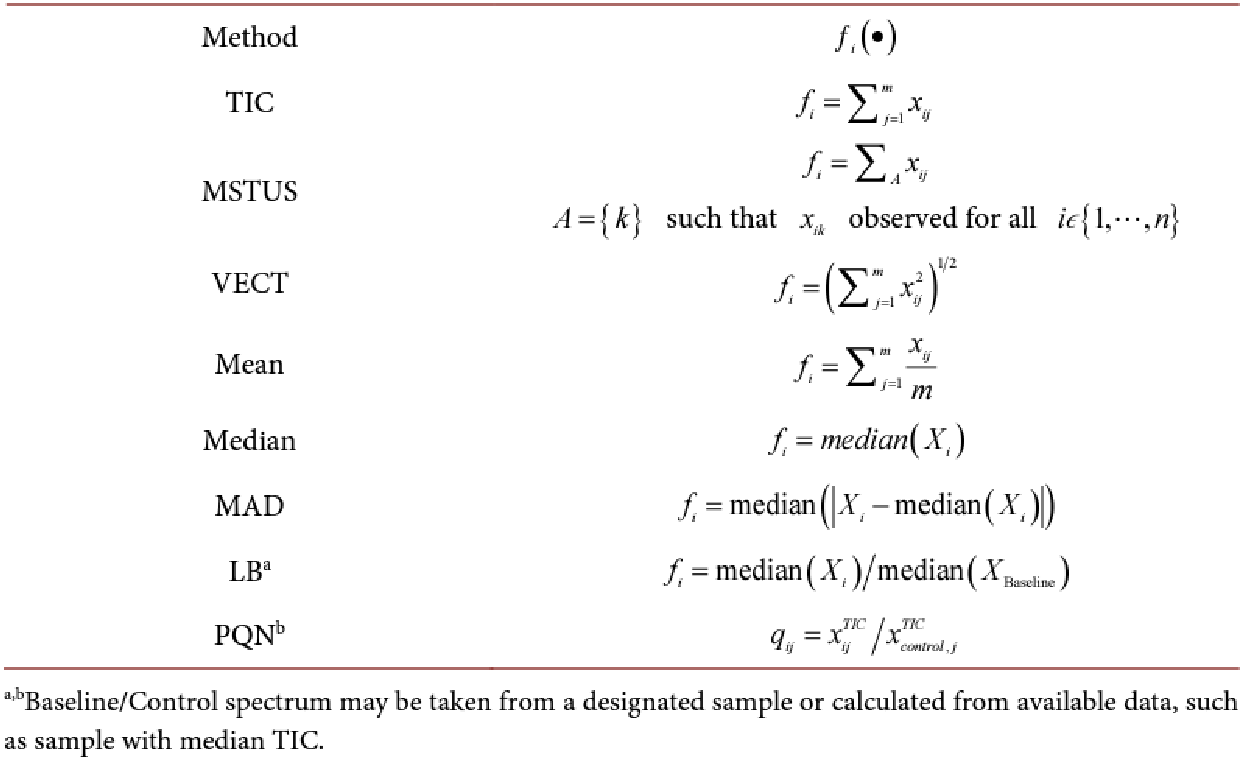
- Sources
  - Plate
  - Study
  - Drift (order of injection)?
  - To be determined
- Method
  - Principal Variance Component Analysis

Batch adjustment

- Methods
  - ComBat (sva)
  - Identification of sources of variation + modeling:
    - Principal Component Partial R-square
    - Linear Mixed Models
  - …?

Normalization

- What is normalization
- When should data be Normalized



Method	$f_i(\bullet)$
TIC	$f_i = \sum_{j=1}^{m} x_{ij}$
MSTUS	$f_i = \sum_{A} x_{ij}$ $A = \{k\}$ such that x_{ik} observed for all $i \epsilon \{1, \cdots, n\}$
VECT	$f_i = \left(\sum_{j=1}^{m} x_{ij}^2\right)^{1/2}$
Mean	$f_i = \sum_{j=1}^{m} \frac{x_{ij}}{m}$
Median	$f_i = median(X_i)$
MAD	$f_i = \text{median}\left(\lvert X_i - \text{median}(X_i)\rvert\right)$
LB[a]	$f_i = \text{median}(X_i)/\text{median}(X_{\text{Baseline}})$
PQN[b]	$q_{ij} = x_{ij}^{TIC} / x_{control,j}^{TIC}$

[a,b] Baseline/Control spectrum may be taken from a designated sample or calculated from available data, such as sample with median TIC.

A Comparison of Various Normalization Methods for LC/MS Metabolomics Data

Meeting minutes

En quin moment avaluar outliers?

En quin moment ajustar per efecte batch, si és que n'hi ha?

Raul: A l'article esmentat s'ajusta l'efecte batch al final de tot perquè en combinar diversos estudis, havia de ser així per força.

Tomás: Si amb els QCs (rèpliques repetides en una mateixa placa i entre plaques) no s'observa efecte batch, no cal ajustar.

Toni: Ha d'haver una coherència entre la transformació i el mètode d'ajust de l'efecte batch si es fa després de la transformació. Transformar pot ser requisit per aplicar un mètode d'ajust de l'efecte batch, per exemple, si requereix condicions de normalitat.

Raul: És important diferenciar endògens/exògens per tal de gestionar els missings. L'absència de metabolits exògens no hauria de ser criteri de "missingness".

- Cal tenir clar què representa un "NA". Pot significar absència del metabòlit o no detectable. Ara bé, quan els pics es troben per sota del llindar de detecció, tant pot ser que el resultat sigui un NA, com un valor numèric no vàlid. Per això necessitem que s'indiqui, per a cada valor, si es trobava dins del rang de detecció o no.
- Cristina: El rang de detecció és específic per a cada metabòlit i pot variar entre experiments.
- Cristina: Hi ha molts mètodes de càlcul dels límits de detecció
- Toni: Els missings són random? Si ho són, te sentit imputar per k-NN
- Alex. Cal tipificar els missings i establir solucions per a cada tipus de missing
- Enrique: Estaria bé buscar un mètode per tal d'avaluar quina és la millor combinació de passos
- Alba: Estaria bé fer un arbre de decisió
- Toni: Es poden fer estudis de simulació