CINDERELLA A semi-automatic user-friendly tool for curation of mass spectrometry metabolomic data

Nathalie Lacrampe, Mary-Lorène Goddard

CINDERELLA
A semi-automatic user-friendly tool for curation of mass spectrometry metabolomic data

Nathalie Lacrampe[1] and Mary-Lorène Goddard[2,3]

1 Laboratoire innovations en spectrométrie de masse pour la santé, CEA, Gif-sur-Yvette, France (curent position)
2 Laboratoire d'innovation moléculaire et applications, UMR 7042 CNRS Université de Haute-Alsace, Mulhouse, France
3 Laboratoire Vigne Biotechnologies et Environnement, UPR 3991 Université de Haute-Alsace, Colmar, France
e-mail: nathalie.lacrampe@gmail.com and mary-lorene.goddard@uha.fr

Comprehensive integrated data editing and refinement for metabolomics by eliminating low-quality signals and artifacts

Introduction

CINDERELLA is an **R** tool with a **user-friendly interface** for semi-automatic and supervised **artifact elimination** of non-targeted **GC-MS and LC-MS metabolomic data** prior to statistical analysis and biological interpretation. Based on pre-processed data from MS-Dial[1], Mzmine[2] or MetaboScape® (Bruker) along with sample metadata, CINDERELLA can, upon user request, (i) handle missing values, (ii) perform sample corrections using internal standard values and/or the amount of biological samples, (iii) correct batch effects, (iv) filter out artifacts and biologically irrelevant signals and (v) assess the quality of the remaining features. Each step is supported by interactive decision-support tables and graphs, which are compiled into an automatic output report and Excel file to enable the monitoring of the data treatment process.

1. Input

- Sample metadata (*.xlsx*) containing:
  - Biological groups
  - Sample type: sample, blank, QC, diluted QC
  - Batches: extraction, analytic
- Additionnal biological information (*.xlsx*):
  - Sample mass: gDW, mgDW, gFW or mgFW
  - Water content: ratio DW/FW, %DW or $\%H_2O$
- Unedited feature file from MS-Dial, Mzmine or Metaboscape®

⇨ **Raw data table**

2. Curation

- Deletion of sample and feature containing too many missing values (NA)
- Replacement of remaining missing values with a very small value or the mean / median of the biological group
  - NAs in **all** biological groups (except blanks and QCs)
  - NAs in **several** biological groups (except blanks and QCs)
  - NAs in a **single** biological groups (except blanks and QCs)
  - NAs only in blanks and QCs

⇨ **NA-curated data table**

⇨ **Flagged data table**

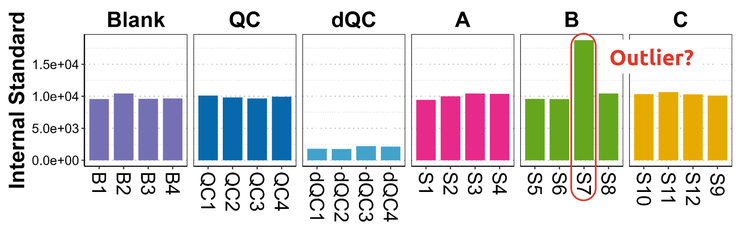
3. Sample correction

Perform internal standard normalisation ? ○ YES ◉ NO

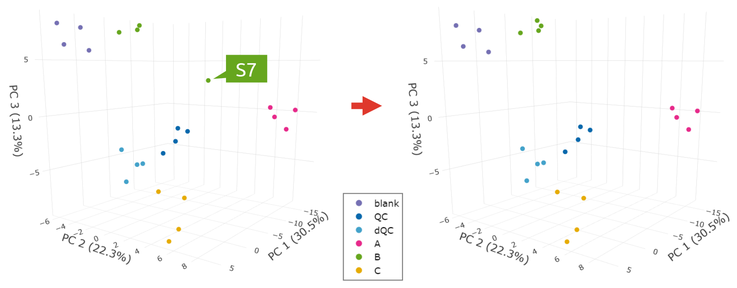
Perform mass normalisation ? ○ YES ◉ NO

Summarized distribution of internal standard by biological group

	Blank	QC	dQC	A	B	C
Mean	9826.08	9886.09	1984.47	10064.77	12091.29	10358.27
SD	410.3	192.15	235.86	460.73	4468.86	227.41
RSD (%)	4.2	1.9	11.9	4.6	37	2.2



Sample distribution before and after correction



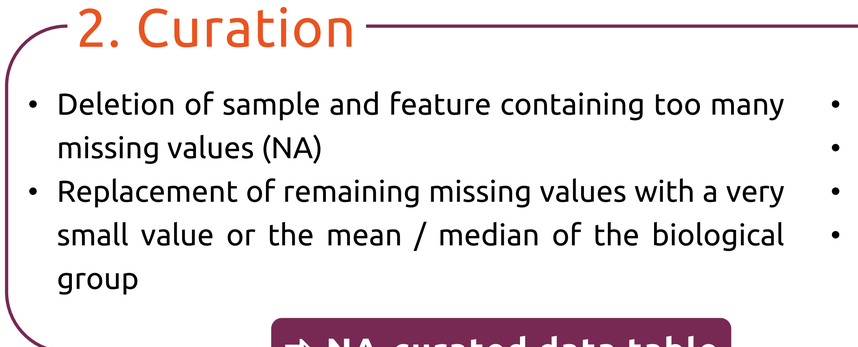
⇨ **IS-corrected data table**

4. Batch effect correction[3, 4]

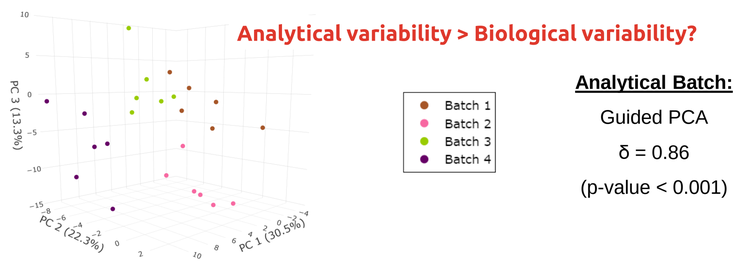
Perform a batch effect correction ? ◉ YES ○ NO

Which batch should be corrected? ○ Extraction batches ◉ Analytical batches

Intensity of internal standard *versus* injection order and colored according to analytical batch

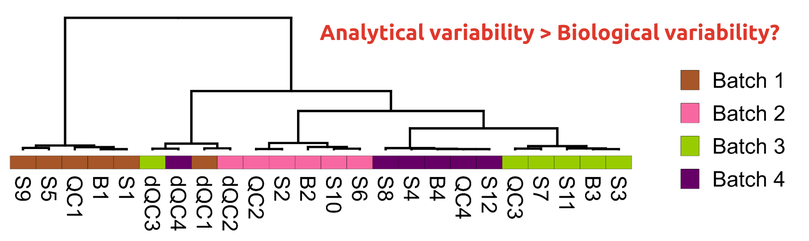


Principal component analysis (PCA) colored according to analytical batch and guided PCA results[5]

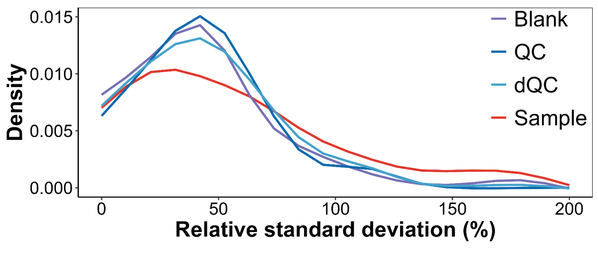


Analytical Batch:
Guided PCA
δ = 0.86
(p-value < 0.001)

Hierarchical clustering analysis of samples



QC-based analysis



	RSD (%)
Blank	41.6
QC	44.6
dQC	46.7
Sample	58.5

⇨ **Batch-corrected data table**

5. Filtering

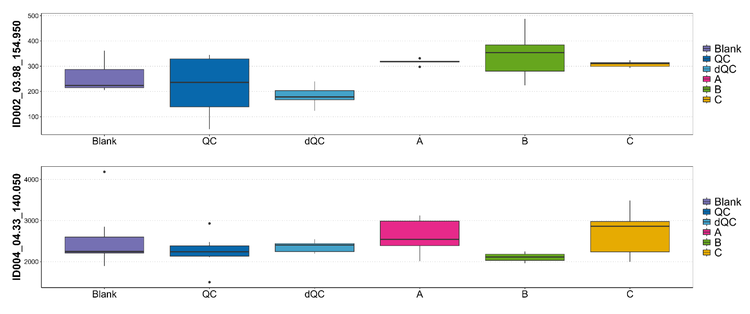
- Features with very small value
- Features with near-constant value
- Features with low repetability
- Features with a value close to blanks

```
Number of features after data curation: 224
Number of features removed by small values filtering: 0 (0%)
Number of features removed by near-constant values filtering: 11 (4.9%)
Number of features removed by low repetability filtering: 2 (0.9%)
Number of features removed by blank presence filtering: 119 (53.1%)
Number of features in the filtered dataset: 103
```

Overview of mean intensities by biological group, showing the effect of filtering features with a value close to blanks

	Filtered	Blank	QC	dQC	A	B	C
ID001_03.55_352.050	FALSE	1761.9	2069.0	1768.3	6405.3	1369.6	6224.1
ID002_03.98_154.950	TRUE	261.2	196.8	170.3	313.1	373.8	317.1
ID003_04.10_102.050	FALSE	54.8	63.2	51.1	24.5	501.8	55.7
ID004_04.33_140.050	TRUE	2779.4	2445.5	2234.2	2600.1	2075.9	2583.7

Boxplot of filtered features



⇨ **Filtered data table**

6. Linearity assessment

Check if features are within the linear dynamic range of the instrument

$$\frac{\text{Intensity (QC)}}{\text{Intensity (diluted QC)}} = \text{Theoretical dilution factor ?}$$

⇨ **Linearity-flagged data table**

7. Export

- All data tables in one file (*.xlsx*)
- Report containing processing steps, figures and additionnal comments (*.html*)

STATISTICAL ANALYSIS & BIOLOGICAL INTERPRETATION

References

1. Tsugawa *et al.*, 2020 (DOI: 10.1038/s41587-020-0531-2); **2.** Schmid *et al.*, 2023 (DOI: 10.1038/s41587-023-01690-2); **3.** Han and LI, 2022 (DOI: 10.1002/mas.21672); **4.** Johnson *et al.*, 2007 (DOI: 10.1093/biostatistics/kxj037); **5.** Reese *et al.*, 2013 (DOI: 10.1093/bioinformatics/btt480)

This project was supported by the Rectorat de l'Académie de Strasbourg