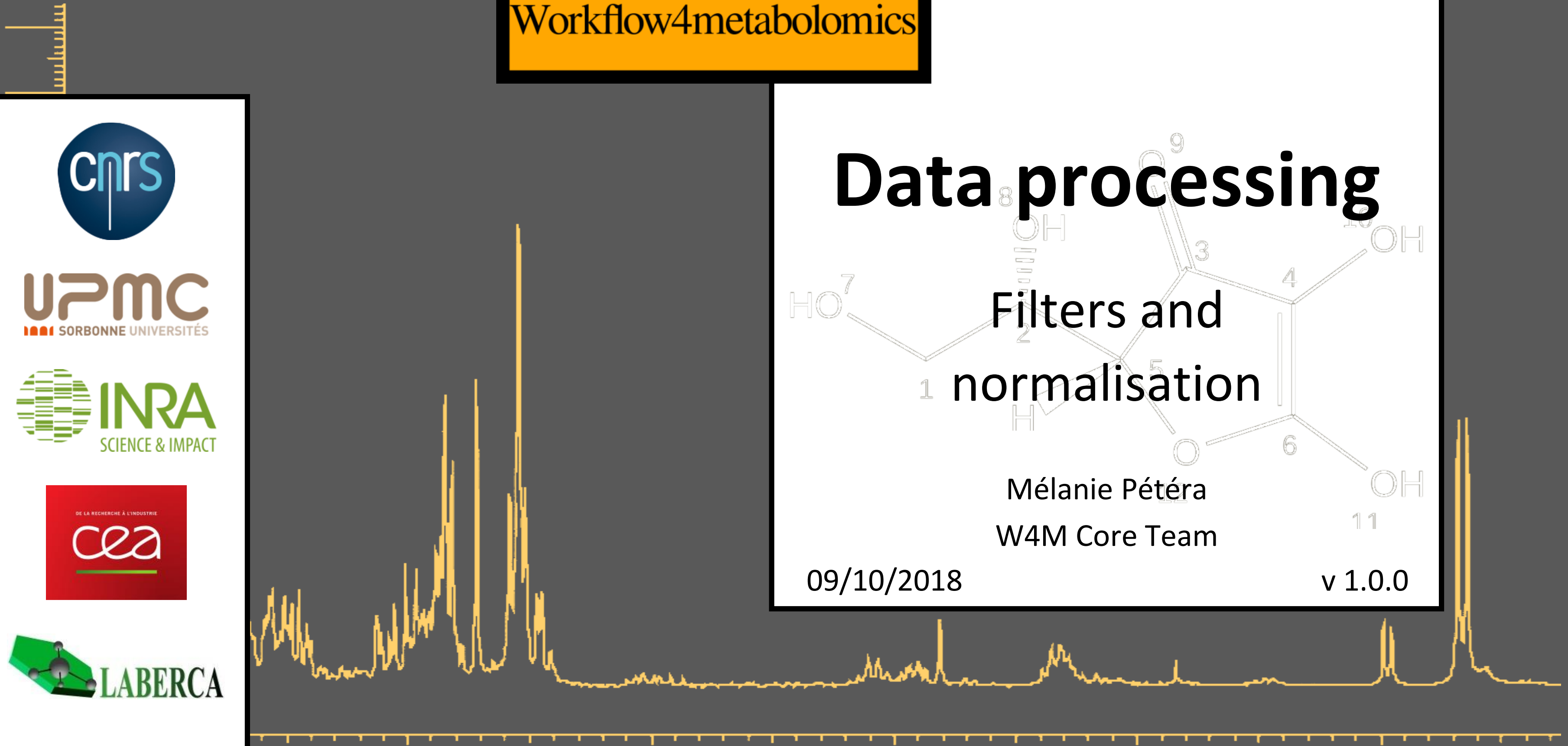
4 Wm

Workflow4metabolomics



Data processing

Filters and normalisation

Mélanie Pétéra

W4M Core Team

09/10/2018

v 1.0.0

Presentation map

1) Processing the data
   - W4M table format for Galaxy

2) A generic tool to filter in Galaxy

3) Signal drift and batch effect correction for MS data
   a) How does that work?
   b) One Galaxy tool, various possibilities

4) Checking for quality
   - Using your pools to check your data

5) Normalization: a tool to normalise

PROCESSING THE DATA

W4M Galaxy tools: a standard format

- A **variety of tools** to process extracted data
  - filters
  - normalisation
  - statistics…

- A **common way to handle data**
  - Easier to follow from a tool to another
  - **Less format switches** in the analysis pipeline
  - A standardised input files format to **easily find the information** needed or obtained

W4M table format for Galaxy

- 3 tables gathering all the information

the data matrix: *intensities of ions or buckets*

the sample metadata file: *information concerning your samples*

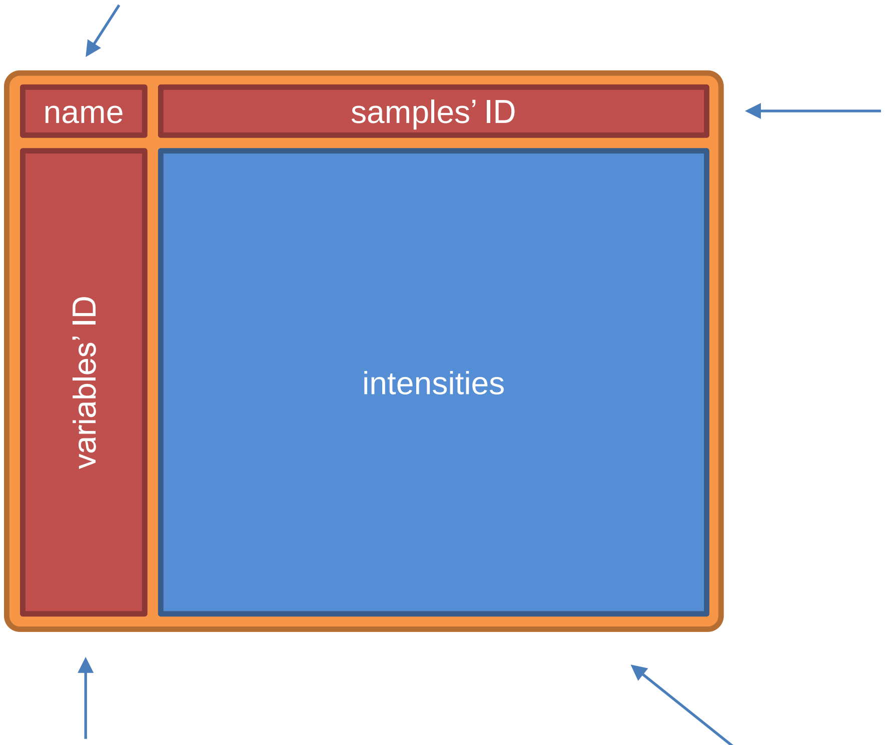
the variable metadata file: *information concerning your ions or buckets*

- Note that this 3 tables structure is already generated from the XCMS or bucketing modules

/!\ You must **complete the sample metadata file** with your samples' information (technical information about your samples, or factors of interest for example)

W4M table format for Galaxy

the name you want (just avoid it to begin with "ID" if you plan to open it with Excel later)



name	samples’ ID
variables’ ID	intensities

first row

Data matrix

first column

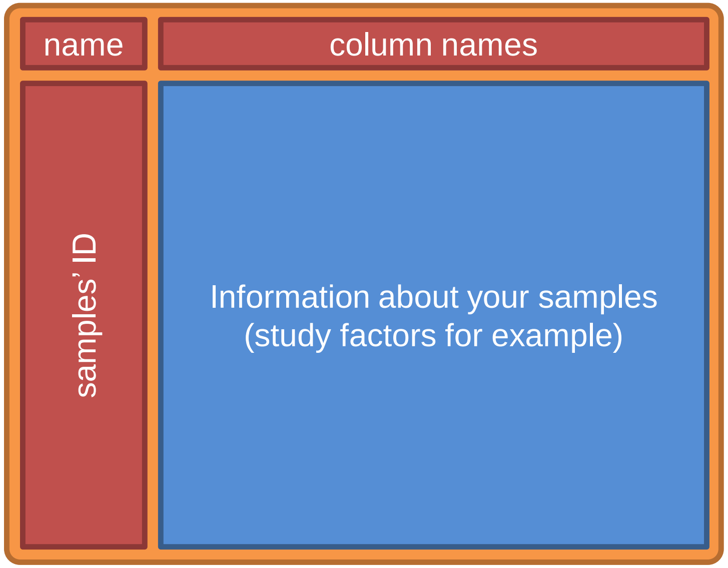
ONLY intensities (no other information)

Note: missing values should be coded NA

W4M table format for Galaxy

Sample metadata

the name you want (just avoid it to begin with "ID" if you plan to open it with Excel later)



first row

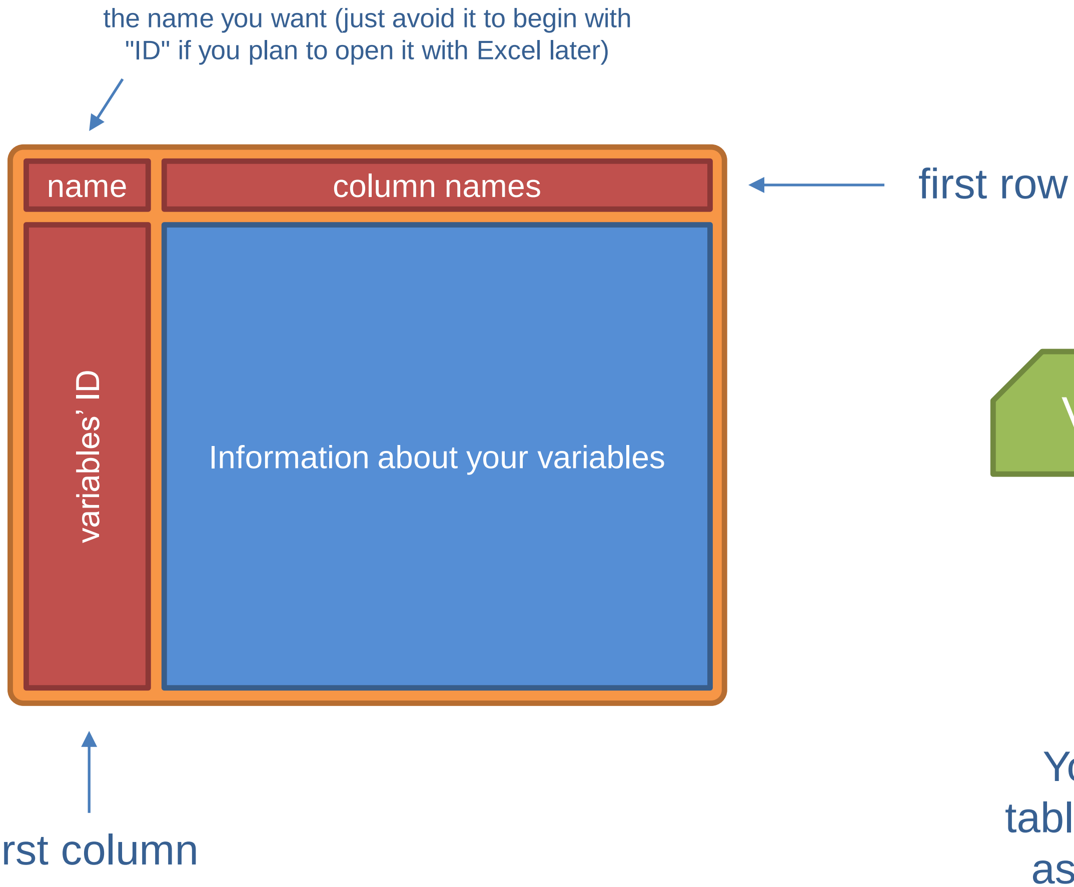
first column

You can add to this table as many columns as you want or need

Samples' ID must absolutely match those in the data matrix file

Note: some modules may need some specific columns with particular names (*e.g.* 'sampleType', 'injectionOrder' or 'batch' for the Batch Correction module)
Refer to the module's help section for more information

W4M table format for Galaxy


the name you want (just avoid it to begin with "ID" if you plan to open it with Excel later)
name
column names
first row
variables' ID
Information about your variables
first column

Variable metadata

You can add to this table as many columns as you want or need

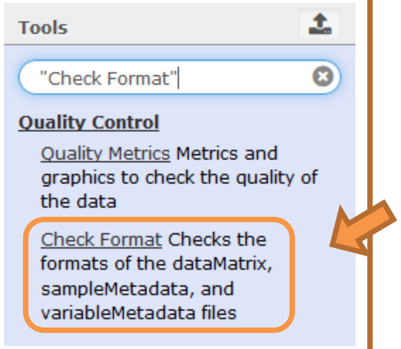
Variables' ID must absolutely match those in the data matrix file

W4M table format for Galaxy

- The files must be tabulated
  - TSV files
  - TXT files with tabulation as separator

- Convention for identifiers and column names
  - It should **not** contain any duplicate
  - Rather use only alphanumeric characters, and points (.) and underscores (_)

Some tools include preliminary tests for your table format, but if you want to make sure everything is alright you can use the Check Format module. It can also help sometimes when you encounter errors you do not understand.



"Check Format"

Quality Control

Quality Metrics Metrics and graphics to check the quality of the data

Check Format Checks the formats of the dataMatrix, sampleMetadata, and variableMetadata files

Generic Filter

A GENERIC TOOL TO FILTER IN GALAXY

A generic tool to filter in Galaxy

- Extracted data often contain more than what you want to use

Depending on your protocol and objectives

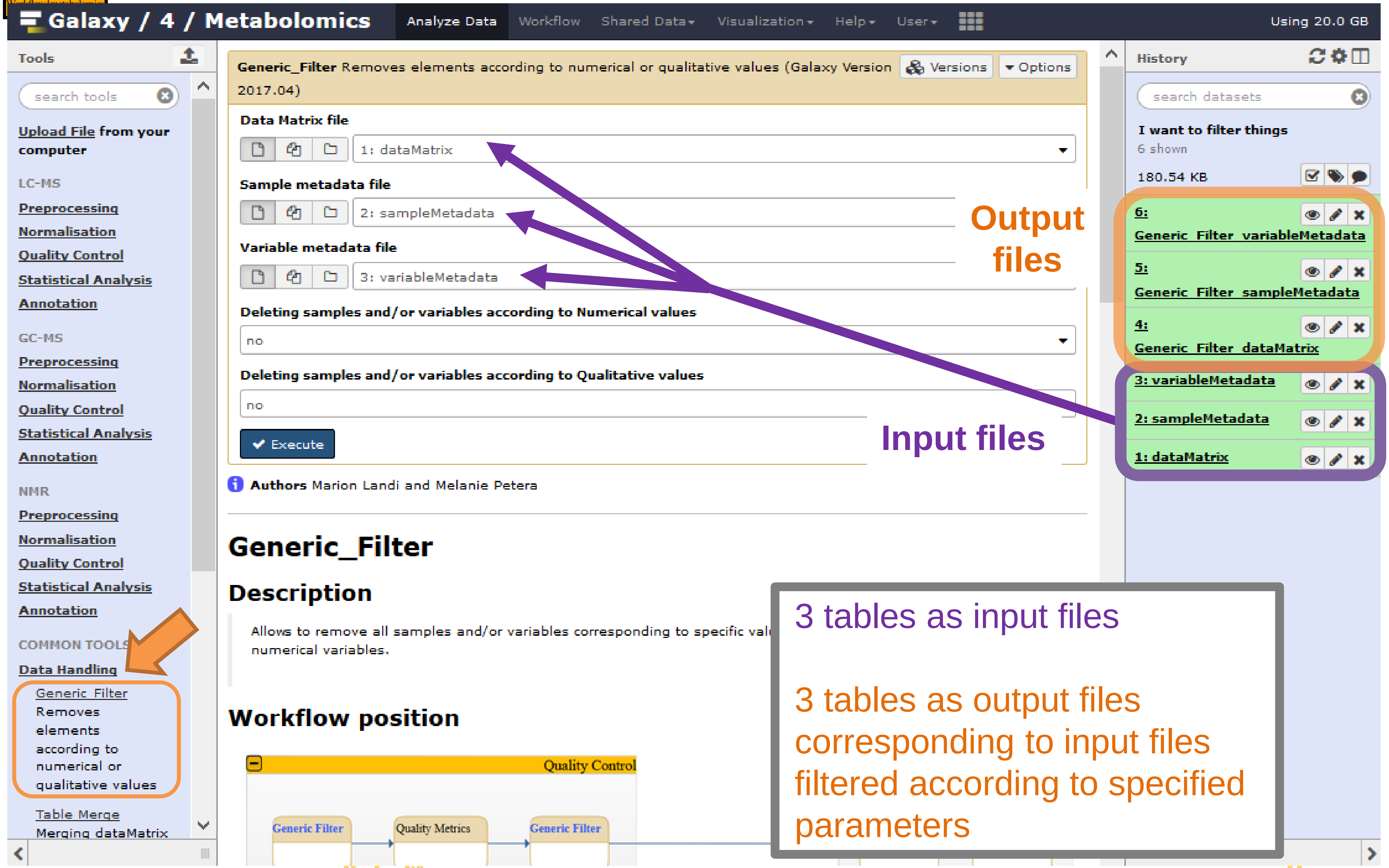
- You need to know what you want to filter

A generic tool invites you to specify exactly what you want to filter => this is **your choice**

- Where is the information to filter?

It must be contained in the sample or variable metadata file (depending on the filter)

Galaxy filtering module: "Generic Filter"



Upload File from your computer

LC-MS
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation

GC-MS
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation

NMR
Preprocessing
Normalisation
Quality Control
Statistical Analysis
Annotation

COMMON TOOLS
Data Handling

Generic_Filter Removes elements according to numerical or qualitative values

Table Merge Merging dataMatrix

Generic_Filter Removes elements according to numerical or qualitative values (Galaxy Version 2017.04)

Data Matrix file: 1: dataMatrix

Sample metadata file: 2: sampleMetadata

Variable metadata file: 3: variableMetadata

Deleting samples and/or variables according to Numerical values: no

Deleting samples and/or variables according to Qualitative values: no

Authors Marion Landi and Melanie Petera

Generic_Filter

Description

Allows to remove all samples and/or variables corresponding to specific val
numerical variables.

Workflow position

I want to filter things
6 shown
180.54 KB

6: Generic_Filter_variableMetadata
5: Generic_Filter_sampleMetadata
4: Generic_Filter_dataMatrix
3: variableMetadata
2: sampleMetadata
1: dataMatrix

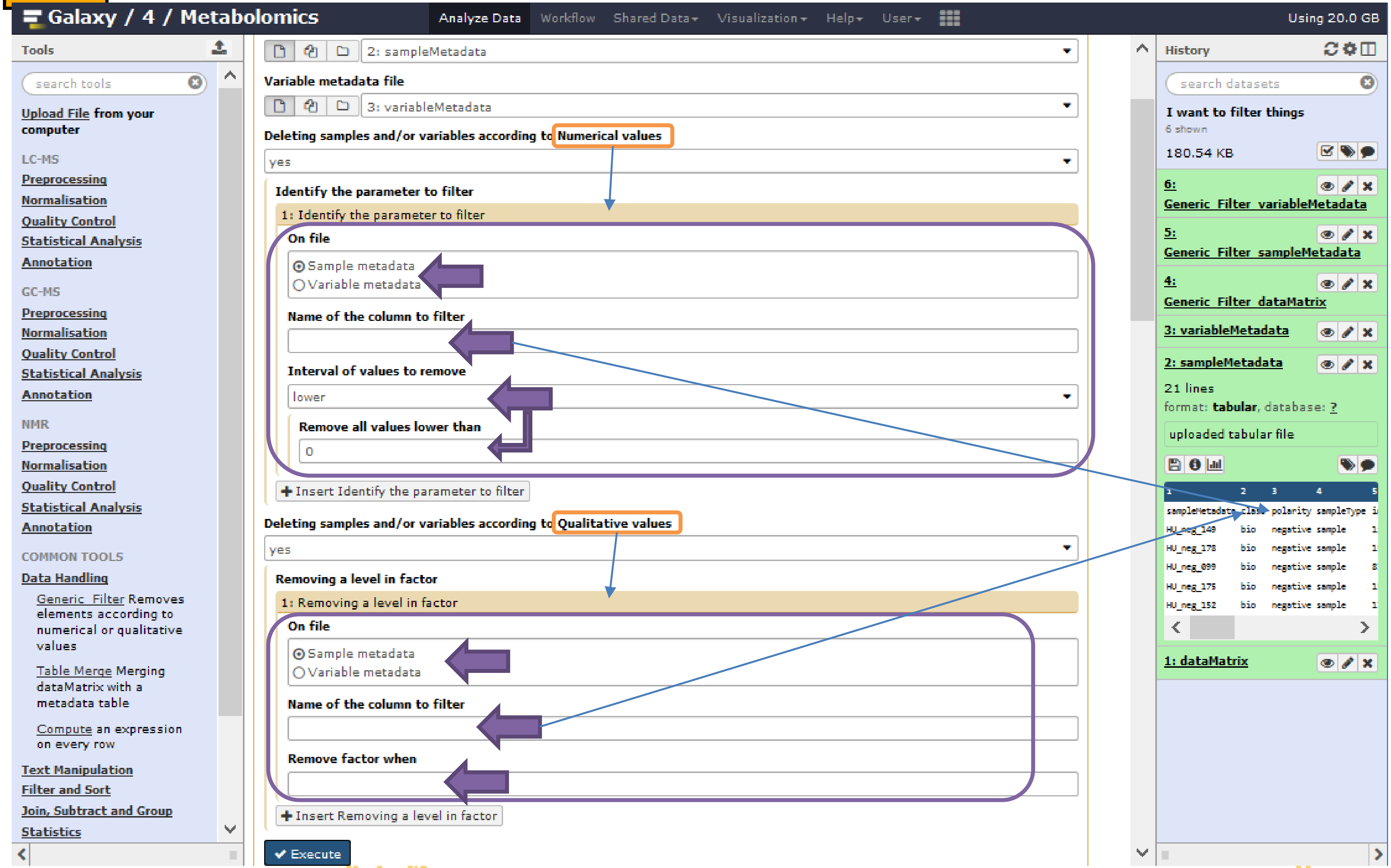
Output files

Input files

3 tables as input files

3 tables as output files corresponding to input files filtered according to specified parameters

Galaxy filtering module: "Generic Filter"



2: sampleMetadata

Variable metadata file

3: variableMetadata

Deleting samples and/or variables according to Numerical values

yes

Identify the parameter to filter

1: Identify the parameter to filter

On file

● Sample metadata
○ Variable metadata

Name of the column to filter

Interval of values to remove

lower

Remove all values lower than

0

+ Insert Identify the parameter to filter

Deleting samples and/or variables according to Qualitative values

yes

Removing a level in factor

1: Removing a level in factor

On file

● Sample metadata
○ Variable metadata

Name of the column to filter

Remove factor when

+ Insert Removing a level in factor

✔ Execute

I want to filter things

6 shown

180.54 KB

6: Generic_Filter_variableMetadata

5: Generic_Filter_sampleMetadata

4: Generic_Filter_dataMatrix

3: variableMetadata

2: sampleMetadata

21 lines

format: **tabular**, database: ?

uploaded tabular file

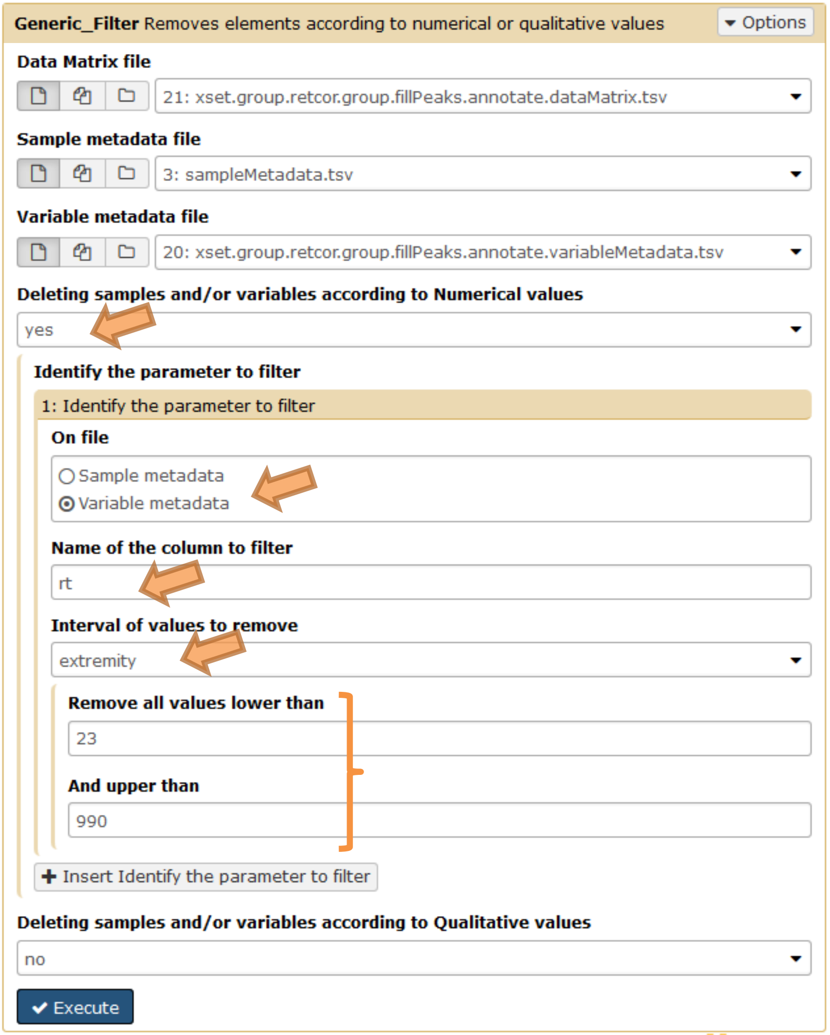
1	2	3	4	5
sampleMetadata	class	polarity	sampleType	i
HU_neg_148	bio	negative	sample	1
HU_neg_178	bio	negative	sample	1
HU_neg_099	bio	negative	sample	8
HU_neg_175	bio	negative	sample	1
HU_neg_152	bio	negative	sample	1

1: dataMatrix

Example1: filtering according to retention time

Example for LC-QTOF with dead volume between 0 and 0.4 min and column flush from 16.5 min

- When using a chromatography column for MS analysis, you may want to exclude some time range, for example to:
  - Exclude the dead volume
  - Exclude a calibration zone at the begining or the end
  - Exclude a column flush
  - …



Generic_Filter Removes elements according to numerical or qualitative values

Data Matrix file

21: xset.group.retcor.group.fillPeaks.annotate.dataMatrix.tsv

Sample metadata file

3: sampleMetadata.tsv

Variable metadata file

20: xset.group.retcor.group.fillPeaks.annotate.variableMetadata.tsv

Deleting samples and/or variables according to Numerical values

yes

Identify the parameter to filter

1: Identify the parameter to filter

On file

○ Sample metadata
● Variable metadata

Name of the column to filter

rt

Interval of values to remove

extremity

Remove all values lower than

23

And upper than

990

Deleting samples and/or variables according to Qualitative values

no

Example 2: using blanks to filter MS data

Step 1

Why using blanks?

Step 2

Getting the data

Step 3

Choosing the filter

- One unavoidable thing in mass spectrometry data is ***noise*** in the signal
- There are ways to reduce the impact on gathered data that may ***sometimes be too radical*** (for example filtering all intensities below a given threshold)
- One possible alternative is the use of ***blanks to estimate the noise***, as a reference

Example 2: using blanks to filter MS data

Step 1

Why using blanks?

Step 2

Getting the data

Step 3

Choosing the filter

- The idea is to ***compare*** blanks' intensities with other samples' intensities (biological samples and/or pools)
- Ideally blanks are your **injection solvent**
- Injected blanks should be **extracted along with** the biological samples.

Example 2: using blanks to filter MS data

Step 1

Why using blanks?

Step 2

Getting the data

Step 3

Choosing the filter

– One common way to compare may be **to set a minimum difference** (by ratio) between means or medians, or to test for significant difference with a statistical test (if you have enough blanks)

Example 2: using blanks to filter MS data

- Example with Galaxy

Available information when **specifying two groups** (blanks and other samples) for extraction steps (2nd column in sampleMetadata for xcms findChromPeaks Merger step):

Used for minimum ratio of means

a ***fold*** column: mean fold change (always greater than 1, see tstat for which set of sample classes is higher)

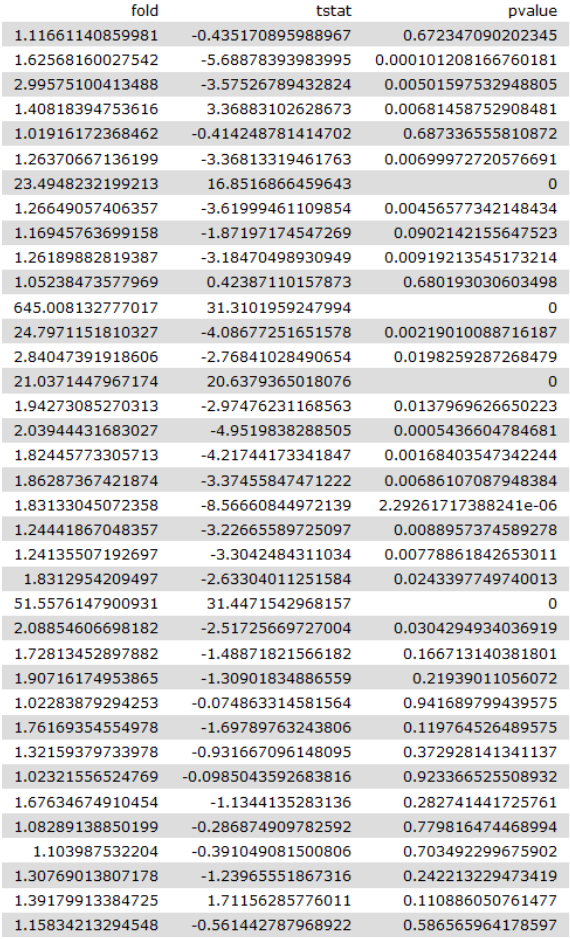
Used to know which group has higher intensities

a ***tstat*** column: Welch’s two sample t-statistic, positive for analytes having greater intensity in class2, negative for analytes having greater intensity in class1

Used for statistical difference

a ***pvalue*** column: p-value of t-statistic

Columns at the end of the **variable metadata** table



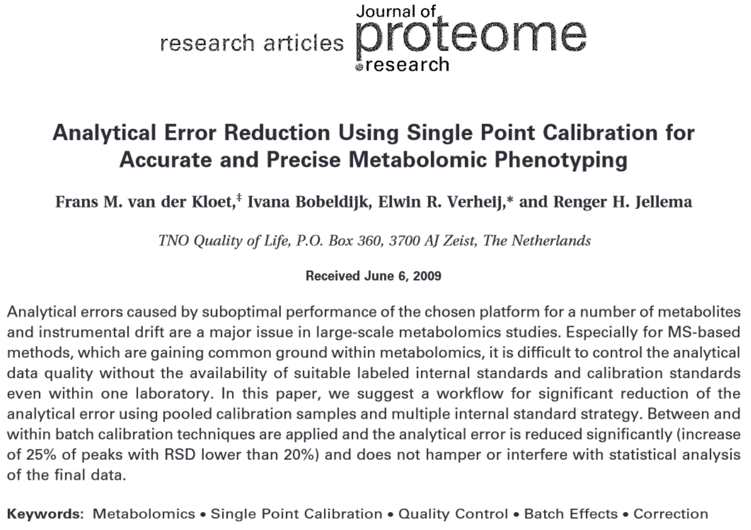
fold	tstat	pvalue
1.11661140859981	-0.435170895988967	0.672347090202345
1.62568160027542	-5.68878393983995	0.000101208166760181
2.99575100413488	-3.57526789432824	0.00501597532948805
1.40818394753616	3.36883102628673	0.00681458752908481
1.01916172368462	-0.414248781414702	0.687336555810872
1.26370667136199	-3.36813319461763	0.00699972720576691
23.4948232199213	16.8516866459643	0
1.26649057406357	-3.61999461109854	0.00456577342148434
1.16945763699158	-1.87197174547269	0.0902142155647523
1.26189882819387	-3.18470498930949	0.00919213545173214
1.05238473577969	0.42387110157873	0.680193030603498
645.008132777017	31.3101959247994	0
24.7971151810327	-4.08677251651578	0.00219010088716187
2.84047391918606	-2.76841028490654	0.0198259287268479
21.0371447967174	20.6379365018076	0
1.94273085270313	-2.97476231168563	0.0137969626650223
2.03944431683027	-4.9519838288505	0.0005436604784681
1.82445773305713	-4.21744173341847	0.00168403547342244
1.86287367421874	-3.37455847471222	0.00686107087948384
1.83133045072358	-8.56660844972139	2.29261717388241e-06
1.24441867048357	-3.22665589725097	0.0088957374589278
1.24135507192697	-3.3042484311034	0.00778861842653011
1.8312954209497	-2.63304011251584	0.0243397749740013
51.5576147900931	31.4471542968157	0
2.08854606698182	-2.51725669727004	0.0304294934036919
1.72813452897882	-1.48871821566182	0.166713140381801
1.90716174953865	-1.30901834886559	0.21939011056072
1.02283879294253	-0.074863314581564	0.941689799439575
1.76169354554978	-1.69789763243806	0.119764526489575
1.32159379733978	-0.931667096148095	0.372928141341137
1.02321556524769	-0.0985043592683816	0.923366525508932
1.67634674910454	-1.1344135283136	0.282741441725761
1.08289138850199	-0.286874909782592	0.779816474468994
1.103987532204	-0.391049081500806	0.703492299675902
1.30769013807178	-1.23965551867316	0.242213229473419
1.39179913384725	1.71156285776011	0.110886050761477
1.15834213294548	-0.561442787968922	0.586565964178597

Use Generic Filter tool to filter!

SIGNAL DRIFT AND BATCH EFFECT CORRECTION FOR MS DATA

How does that work?

- A normalisation process first established by Van Der Kloet *et al.*
  - *F.M. Van Der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema. (2009). "Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping." Journal of Proteome Research p5132-5141*



research articles Journal of proteome research

Analytical Error Reduction Using Single Point Calibration for Accurate and Precise Metabolomic Phenotyping

Frans M. van der Kloet,‡ Ivana Bobeldijk, Elwin R. Verheij,* and Renger H. Jellema

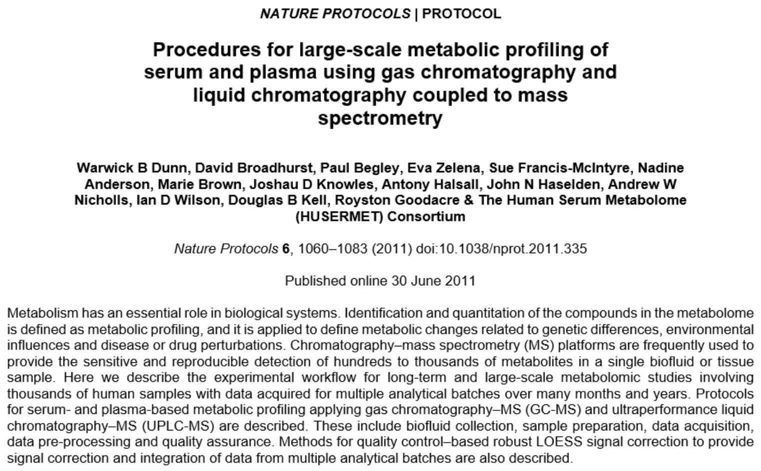
TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands

Received June 6, 2009

Analytical errors caused by suboptimal performance of the chosen platform for a number of metabolites and instrumental drift are a major issue in large-scale metabolomics studies. Especially for MS-based methods, which are gaining common ground within metabolomics, it is difficult to control the analytical data quality without the availability of suitable labeled internal standards and calibration standards even within one laboratory. In this paper, we suggest a workflow for significant reduction of the analytical error using pooled calibration samples and multiple internal standard strategy. Between and within batch calibration techniques are applied and the analytical error is reduced significantly (increase of 25% of peaks with RSD lower than 20%) and does not hamper or interfere with statistical analysis of the final data.

Keywords: Metabolomics • Single Point Calibration • Quality Control • Batch Effects • Correction

- which have made its way to nowadays procedures
  - *Dunn et al (2011). "Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry." Nature Protocols, 6:1060-1083*



NATURE PROTOCOLS **| PROTOCOL**

Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry

Warwick B Dunn, David Broadhurst, Paul Begley, Eva Zelena, Sue Francis-McIntyre, Nadine Anderson, Marie Brown, Joshau D Knowles, Antony Halsall, John N Haselden, Andrew W Nicholls, Ian D Wilson, Douglas B Kell, Royston Goodacre & The Human Serum Metabolome (HUSERMET) Consortium

Nature Protocols **6**, 1060–1083 (2011) doi:10.1038/nprot.2011.335

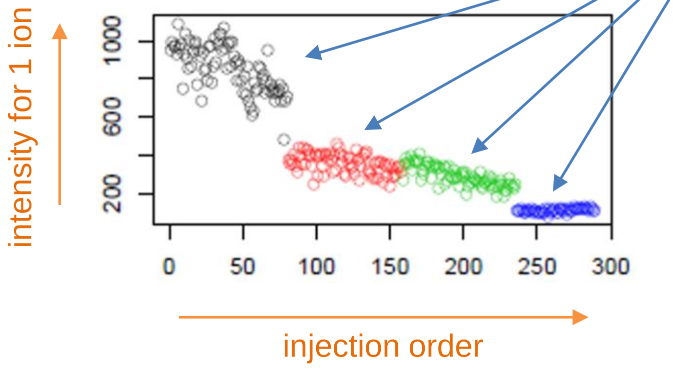
Published online 30 June 2011

Metabolism has an essential role in biological systems. Identification and quantitation of the compounds in the metabolome is defined as metabolic profiling, and it is applied to define metabolic changes related to genetic differences, environmental influences and disease or drug perturbations. Chromatography–mass spectrometry (MS) platforms are frequently used to provide the sensitive and reproducible detection of hundreds to thousands of metabolites in a single biofluid or tissue sample. Here we describe the experimental workflow for long-term and large-scale metabolomic studies involving thousands of human samples with data acquired for multiple analytical batches over many months and years. Protocols for serum- and plasma-based metabolic profiling applying gas chromatography–MS (GC-MS) and ultraperformance liquid chromatography–MS (UPLC-MS) are described. These include biofluid collection, sample preparation, data acquisition, data pre-processing and quality assurance. Methods for quality control–based robust LOESS signal correction to provide signal correction and integration of data from multiple analytical batches are also described.

How does that work?

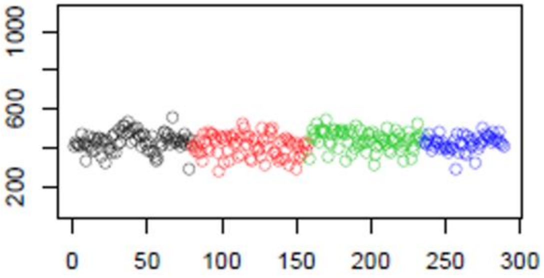
- Principle
  - What we have

distinct batches of analyse



particular intra-batch analytical effects

  - What we want

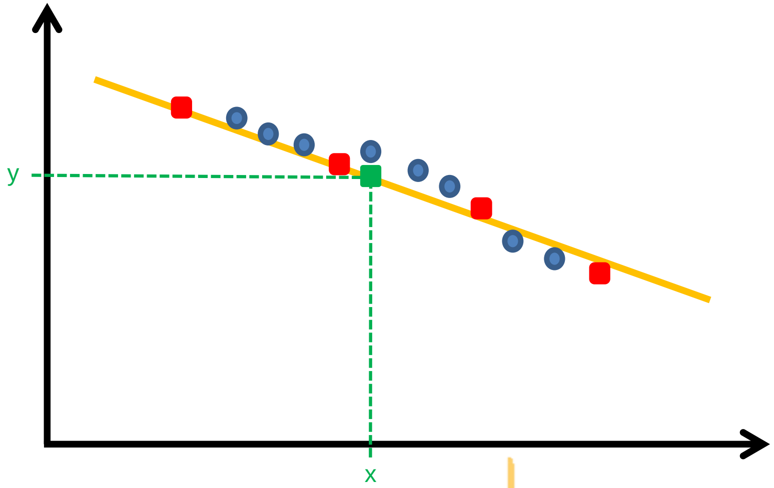


Comparable intensities

How does that work?

- Technically speaking
  - Correction is made for each ion independantly
  - For each ion:
    - An intra-batch correction is made for each batch independantly
      - Analytical effect is modelled using pools' intensities according to the injection order
      - Each sample intensity is devided by the estimation of analytical effect of corresponding injection number
      - Sample values are then multiplied by a reference value (to keep original ion scale)
    - Inter-batch effect is thus automatically corrected

Pools = Quality-control pooled samples, all identical, injected regularly all through an analytical sequence



🟥 Observed pool value

● Observed sample value

— Regression curve of analytical effect model

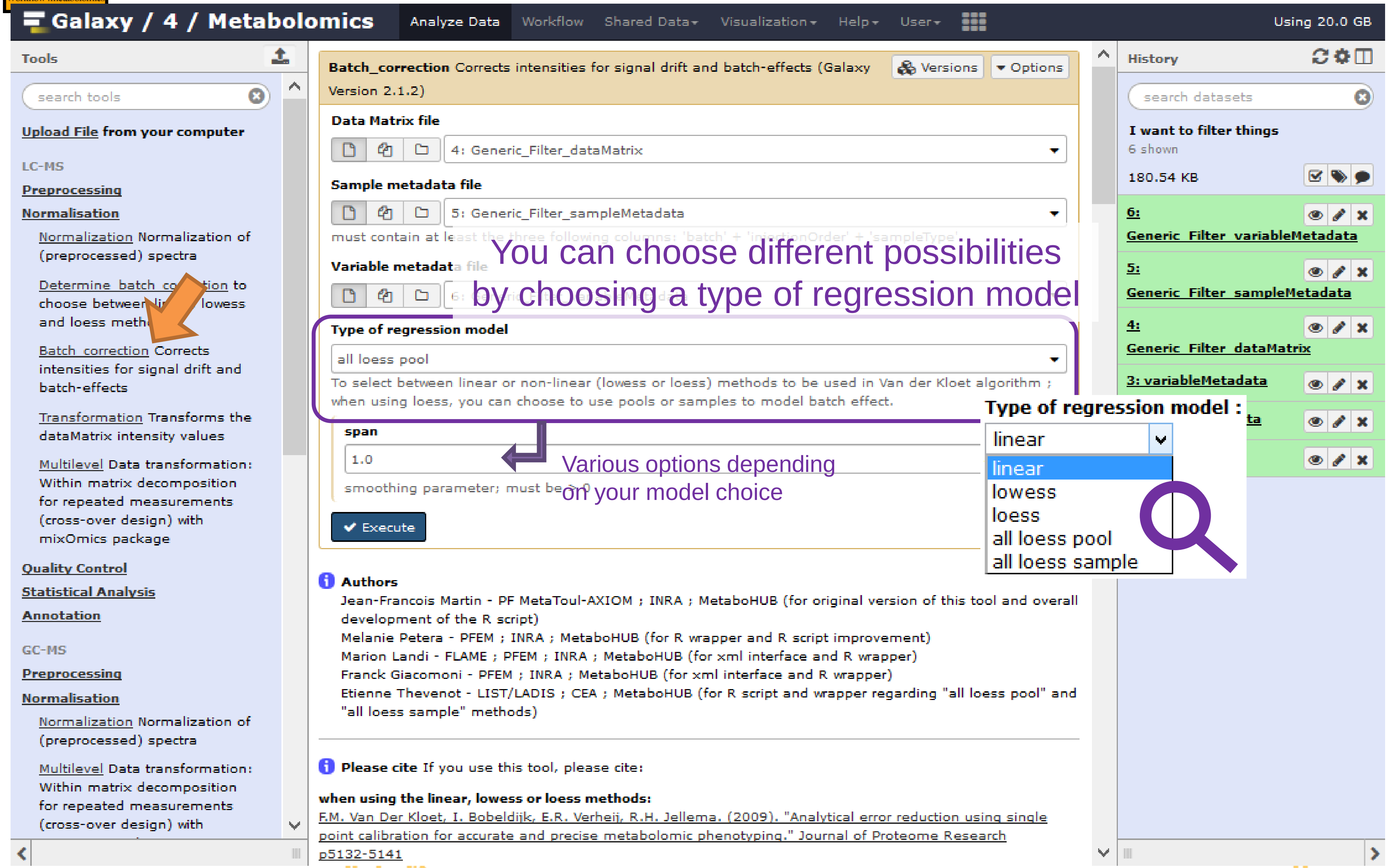
🟩 Estimated value for injection number x

$$\text{normalised value for sample obtained at injection number x} = \frac{\text{observed sample value at injection number x } ●}{\text{estimated value for injection number x } 🟩} \times \text{reference value}$$

How does that work?

- What you need to make it go smoothly
  - Pools should be ***injected regularly*** through your sequences
  - Pools should be ***identical***, preferably a mix of all your biological samples to be representative of molecule diversity
  - Pools should be ***numerous enough*** in each batch, for the regression to be reliable (must be, at the very least, of 3 per batch for linear methods and 8 for non-linear ones)
  - It's recommended that your biological samples may be ***randomised*** for injection order
  - Your data ***must contain*** specific information in sample metadata file:
    - the injection order
    - the batches of analyse
    - the sample type (pool or sample)

One Galaxy tool, various possibilities



Upload File from your computer

LC-MS

Preprocessing

Normalisation

Normalization Normalization of (preprocessed) spectra

Determine batch co[...]tion to choose between li[...] lowess and loess meth[...]

Batch correction Corrects intensities for signal drift and batch-effects

Transformation Transforms the dataMatrix intensity values

Multilevel Data transformation: Within matrix decomposition for repeated measurements (cross-over design) with mixOmics package

Quality Control

Statistical Analysis

Annotation

GC-MS

Preprocessing

Normalisation

Normalization Normalization of (preprocessed) spectra

Multilevel Data transformation: Within matrix decomposition for repeated measurements (cross-over design) with

Batch_correction Corrects intensities for signal drift and batch-effects (Galaxy Version 2.1.2)

Data Matrix file

4: Generic_Filter_dataMatrix

Sample metadata file

5: Generic_Filter_sampleMetadata

must contain at least the three following columns: 'batch' + 'injectionOrder' + 'sampleType'

Variable metadata file

6: Generic_Filter_variableMetadata

Type of regression model

all loess pool

To select between linear or non-linear (lowess or loess) methods to be used in Van der Kloet algorithm ; when using loess, you can choose to use pools or samples to model batch effect.

span

1.0

smoothing parameter; must be > 0

Authors

Jean-Francois Martin - PF MetaToul-AXIOM ; INRA ; MetaboHUB (for original version of this tool and overall development of the R script)
Melanie Petera - PFEM ; INRA ; MetaboHUB (for R wrapper and R script improvement)
Marion Landi - FLAME ; PFEM ; INRA ; MetaboHUB (for xml interface and R wrapper)
Franck Giacomoni - PFEM ; INRA ; MetaboHUB (for xml interface and R wrapper)
Etienne Thevenot - LIST/LADIS ; CEA ; MetaboHUB (for R script and wrapper regarding "all loess pool" and "all loess sample" methods)

Please cite If you use this tool, please cite:

when using the linear, lowess or loess methods:

F.M. Van Der Kloet, I. Bobeldijk, E.R. Verheij, R.H. Jellema. (2009). "Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping." Journal of Proteome Research p5132-5141

I want to filter things

6 shown

180.54 KB

6: Generic_Filter_variableMetadata

5: Generic_Filter_sampleMetadata

4: Generic_Filter_dataMatrix

3: variableMetadata

You can choose different possibilities by choosing a type of regression model

Various options depending on your model choice

Type of regression model :

linear

- linear
- lowess
- loess
- all loess pool
- all loess sample

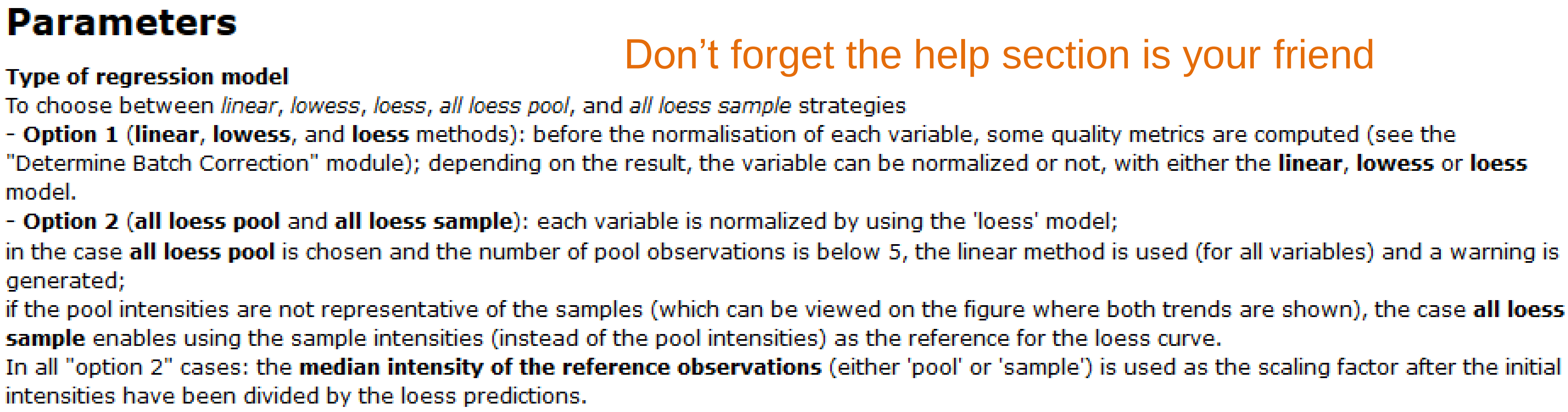
What’s different?

- Two strategies implemented
  - linear / lowess / loess
  - all loess pool / all loess sample

● choice in regression model type
● intra-batch correction is conditioned to internal quality metrics

● possibility to apply correction based on sample intensities only

- Distinct graphical output for each strategie
  - Different variations of before/after overview



Parameters

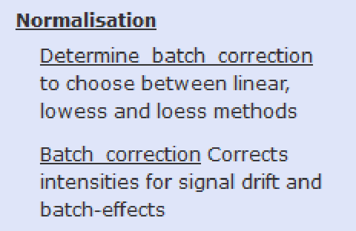
Don’t forget the help section is your friend

Type of regression model
To choose between *linear*, *lowess*, *loess*, *all loess pool*, and *all loess sample* strategies
- **Option 1** (**linear**, **lowess**, and **loess** methods): before the normalisation of each variable, some quality metrics are computed (see the "Determine Batch Correction" module); depending on the result, the variable can be normalized or not, with either the **linear**, **lowess** or **loess** model.
- **Option 2** (**all loess pool** and **all loess sample**): each variable is normalized by using the 'loess' model;
in the case **all loess pool** is chosen and the number of pool observations is below 5, the linear method is used (for all variables) and a warning is generated;
if the pool intensities are not representative of the samples (which can be viewed on the figure where both trends are shown), the case **all loess sample** enables using the sample intensities (instead of the pool intensities) as the reference for the loess curve.
In all "option 2" cases: the **median intensity of the reference observations** (either 'pool' or 'sample') is used as the scaling factor after the initial intensities have been divided by the loess predictions.

How to use this tool

- ***Mandatory columns*** in sample metadata table
  - *injectionOrder*: numerical column of injection order
  - *sampleType*: specifies if a pool or a sample (coded “pool” or “sample”)
  - *batch*: categorical column indicating the batches of analyse (if only one, must be a constant)
- In the data matrix (containing intensities), ***missing values*** are allowed ***only for all loess*** methods
- In case you want to use the linear / lowess / loess strategy, you can use the “Determine batch correction” tool to help you in the choice of a regression type

This module computes graphics and indicators, but the user remains the only judge regarding which model is the more appropriate for his data.



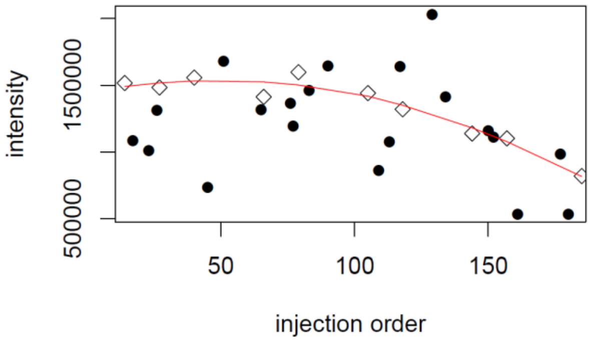
Normalisation

Determine_batch_correction to choose between linear, lowess and loess methods

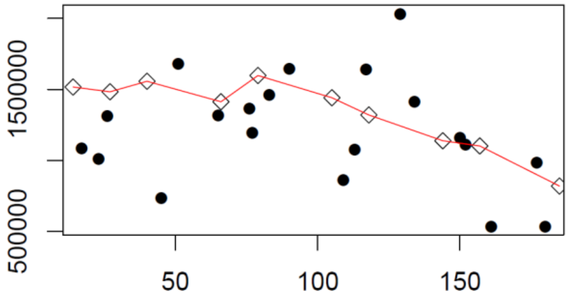
Batch_correction Corrects intensities for signal drift and batch-effects

How to use this tool

- Parameters
  - ***Span*** (*not available for ‘linear’ method*): smoothing parameter for lo(w)ess regression



quite a smooth curve (span=1)

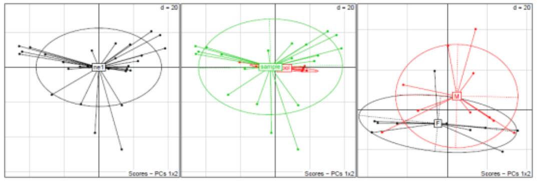


not smooth at all (span=0.3)

How to use this tool

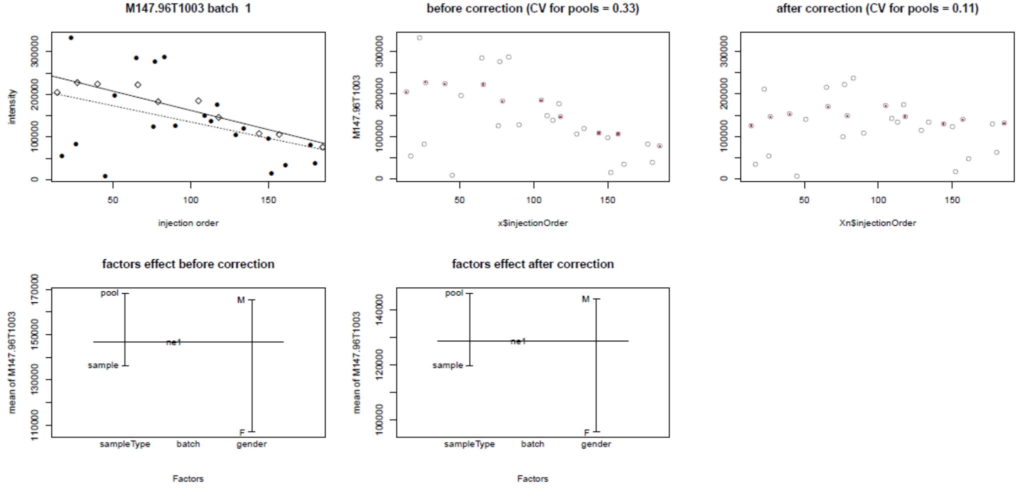
- Parameters *(not available for 'all loess' strategy)*
  - ***Null values***:
    what to do when negative or infinite intensity values are generated during calculations
  - ***Factor of interest***:
    a categorical column in sample metadata table, used to have a quick graphical overview of the effect of normalisation on this variable in the data; this does not affect correction calculation
  - ***Level of details for plots***:
    to choose the amount of graphical output to produce in the pdf file

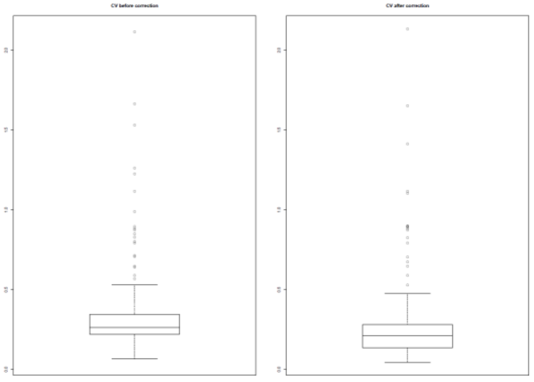
Coloration depending on factors

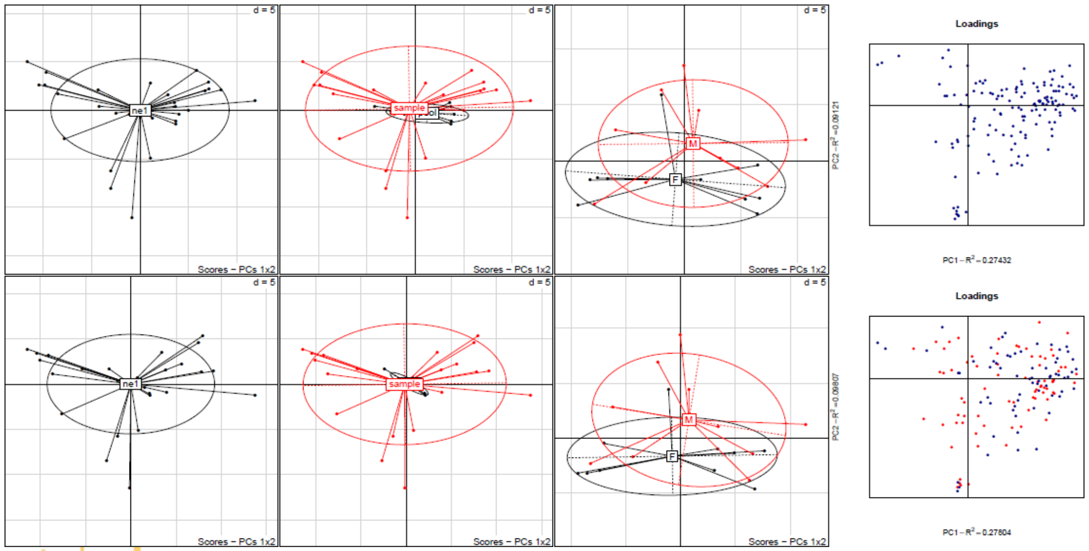


batch sample type factor of interest

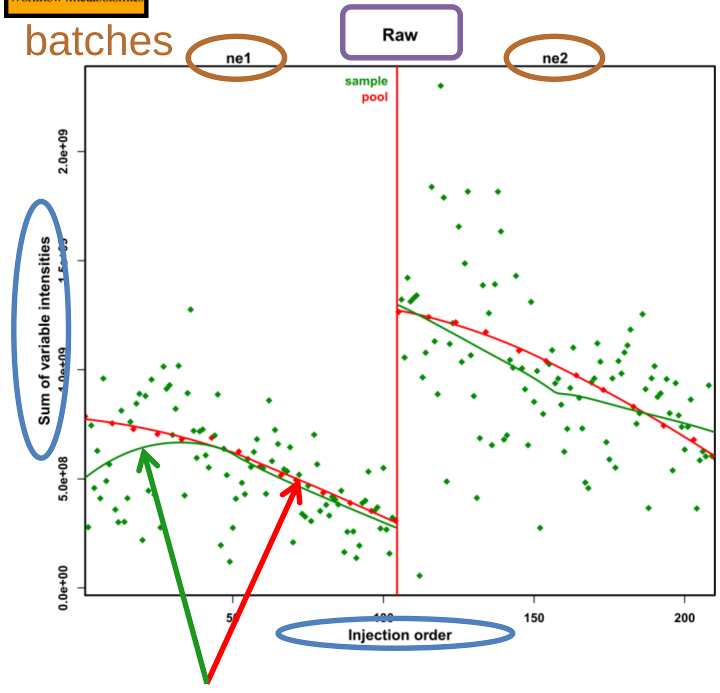
Graphical output: linear/lowess/loess





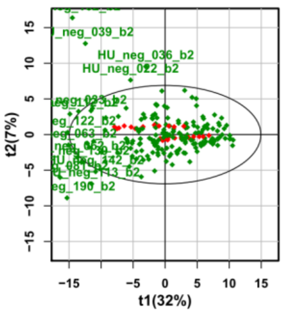


Graphical output: all_loess

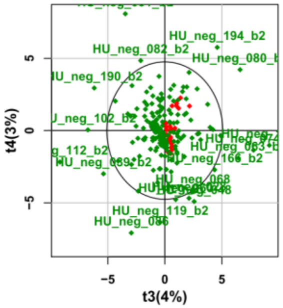


batches

visualization of loess curves

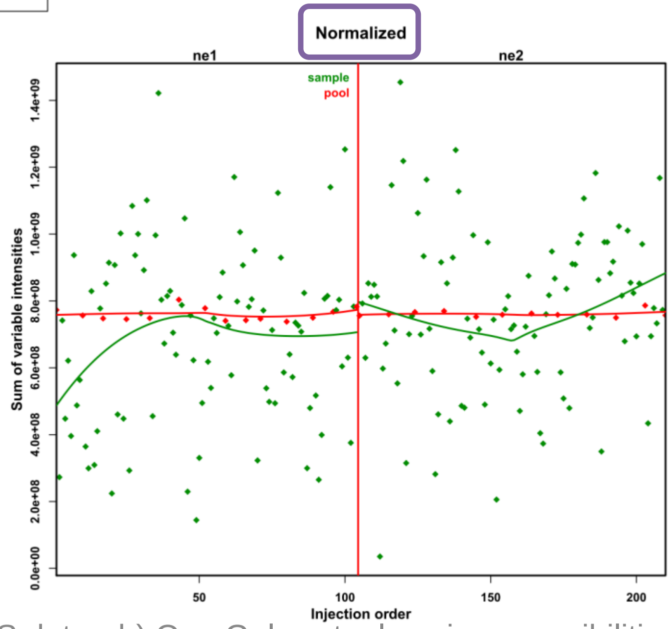


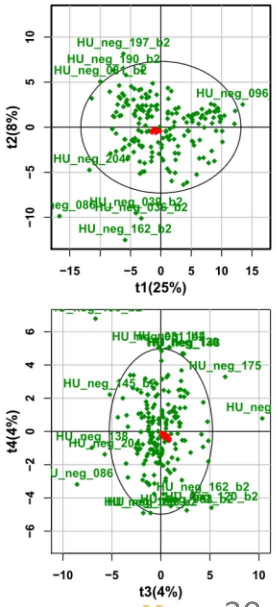
PCA (t1,t2)



PCA (t3,t4)

all_loess_pool method





CHECKING FOR QUALITY

Using your pools to check your data

- What to check

calculation per ion

$$CV = \frac{\sigma}{\mu}$$

where:

$\sigma = standard\ deviation$

$\mu = mean$

  - ***Coefficient of variation***:

used individually

or

used with ratio

e.g. pools' CV is often considered to be too high if upper than 0.3

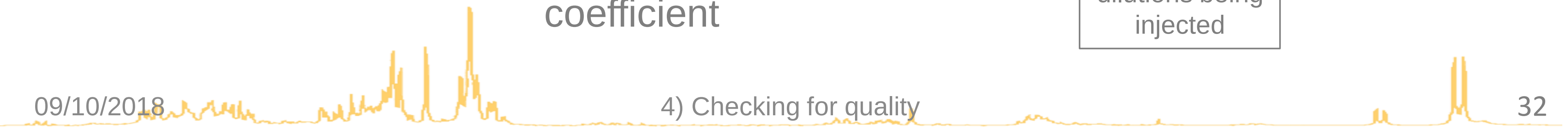
e.g. ration between pools and samples may be considered too high if upper than 1 (⇔ pools are more variable than samples)

global boxplot available in Batch Correction output with linear/loess/lowess methods

  - ***Correlation with pool dilutions***:
    "Does intensity evolve according to dilution?"

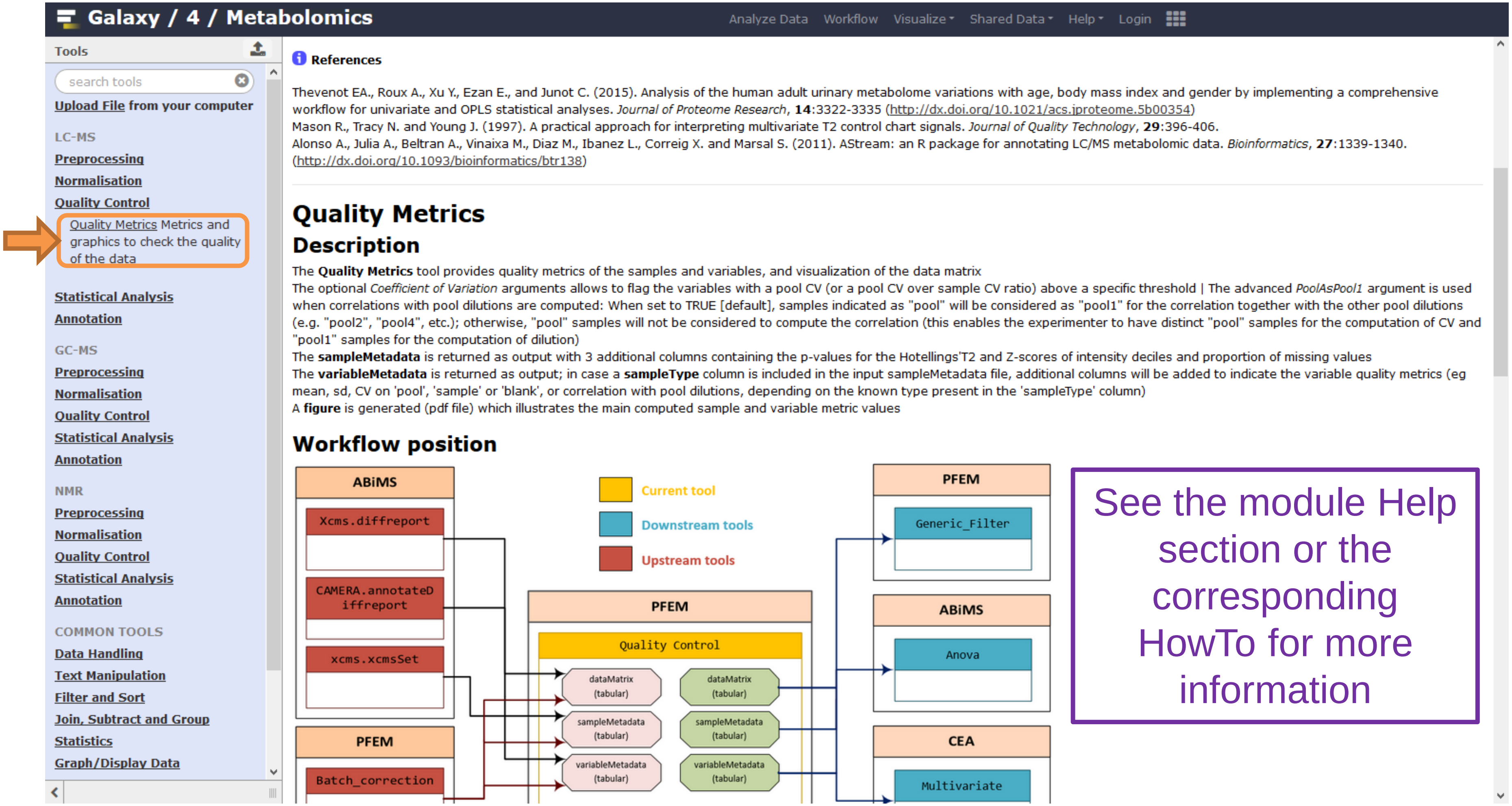
Pearson's correlation coefficient

Needs pool dilutions being injected



Using your pools to check your data

Use the Quality Metrics module to compute your indicators



References

Thevenot EA., Roux A., Xu Y., Ezan E., and Junot C. (2015). Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *Journal of Proteome Research*, **14**:3322-3335 (http://dx.doi.org/10.1021/acs.jproteome.5b00354)
Mason R., Tracy N. and Young J. (1997). A practical approach for interpreting multivariate T2 control chart signals. *Journal of Quality Technology*, **29**:396-406.
Alonso A., Julia A., Beltran A., Vinaixa M., Diaz M., Ibanez L., Correig X. and Marsal S. (2011). AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics*, **27**:1339-1340. (http://dx.doi.org/10.1093/bioinformatics/btr138)

Quality Metrics

Description

The **Quality Metrics** tool provides quality metrics of the samples and variables, and visualization of the data matrix
The optional *Coefficient of Variation* arguments allows to flag the variables with a pool CV (or a pool CV over sample CV ratio) above a specific threshold | The advanced *PoolAsPool1* argument is used when correlations with pool dilutions are computed: When set to TRUE [default], samples indicated as "pool" will be considered as "pool1" for the correlation together with the other pool dilutions (e.g. "pool2", "pool4", etc.); otherwise, "pool" samples will not be considered to compute the correlation (this enables the experimenter to have distinct "pool" samples for the computation of CV and "pool1" samples for the computation of dilution)
The **sampleMetadata** is returned as output with 3 additional columns containing the p-values for the Hotellings'T2 and Z-scores of intensity deciles and proportion of missing values
The **variableMetadata** is returned as output; in case a **sampleType** column is included in the input sampleMetadata file, additional columns will be added to indicate the variable quality metrics (eg mean, sd, CV on 'pool', 'sample' or 'blank', or correlation with pool dilutions, depending on the known type present in the 'sampleType' column)
A **figure** is generated (pdf file) which illustrates the main computed sample and variable metric values

Workflow position

See the module Help section or the corresponding HowTo for more information

Note: this module can be used even without pools since it computes other interesting quality information and graphics

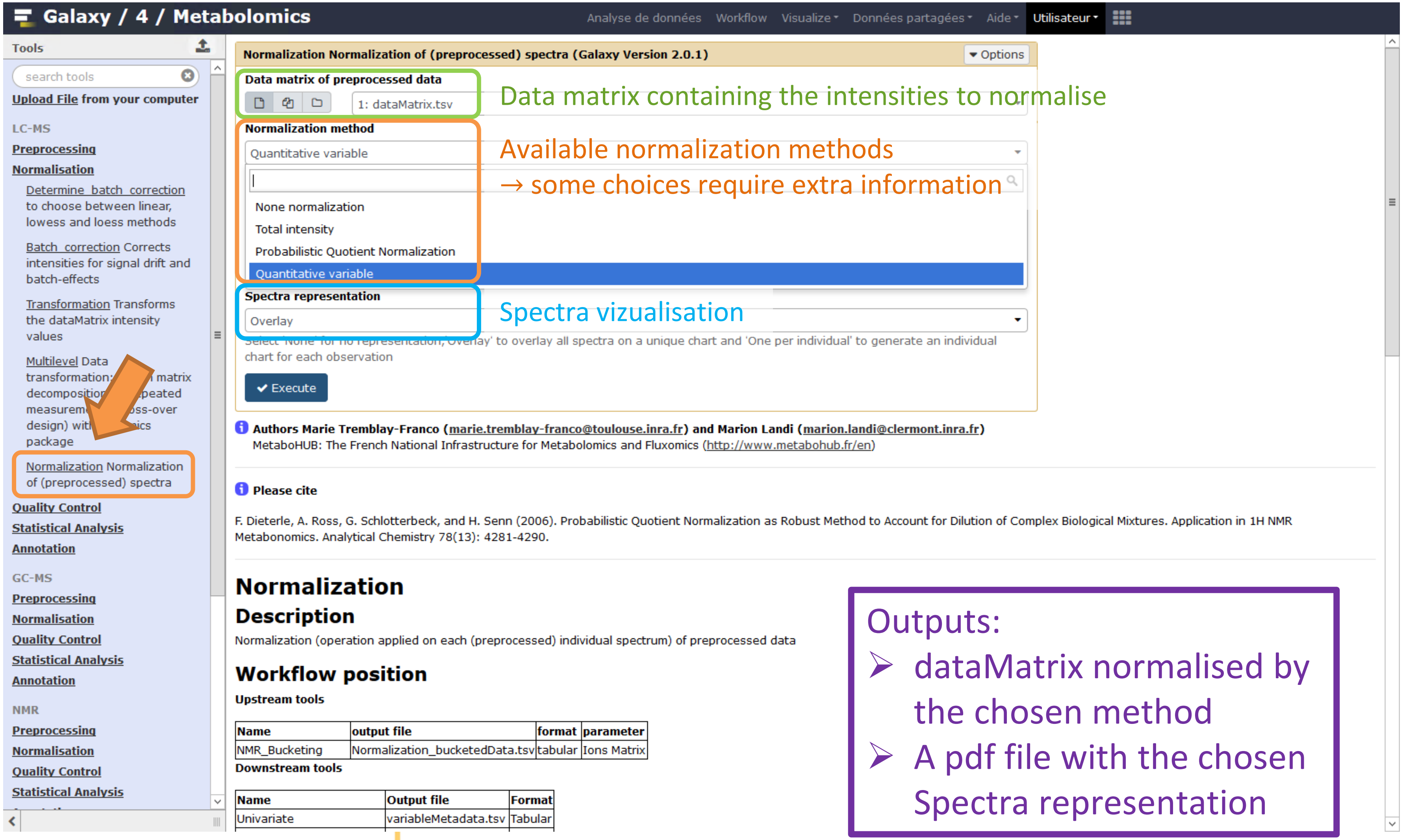
NORMALIZATION: A TOOL TO NORMALISE

About normalisation

- Operation applied to each sample to make the data from all samples directly comparable with each other (to take into account variations of the overall concentrations of samples due to biological and technical reasons)

⇒ To ensure that a measured concentration observed for a metabolite at the lower end of the dynamic range is as reliable as it is for a metabolite at the upper end

About the Normalization tool



Upload File from your computer

LC-MS

Preprocessing

Normalisation

Determine_batch_correction to choose between linear, lowess and loess methods

Batch_correction Corrects intensities for signal drift and batch-effects

Transformation Transforms the dataMatrix intensity values

Multilevel Data transformation: matrix decomposition repeated measurem oss-over design) wit ics package

Normalization Normalization of (preprocessed) spectra

Quality Control

Statistical Analysis

Annotation

GC-MS

Preprocessing

Normalization

Quality Control

Statistical Analysis

Annotation

NMR

Preprocessing

Normalization

Quality Control

Statistical Analysis

Normalization Normalization of (preprocessed) spectra (Galaxy Version 2.0.1)

Data matrix of preprocessed data

1: dataMatrix.tsv

Data matrix containing the intensities to normalise

Normalization method

Quantitative variable

- None normalization
- Total intensity
- Probabilistic Quotient Normalization
- Quantitative variable

Available normalization methods
→ some choices require extra information

Spectra representation

Overlay

Spectra vizualisation

Select 'None' for no representation, 'Overlay' to overlay all spectra on a unique chart and 'One per individual' to generate an individual chart for each observation

✓ Execute

Authors Marie Tremblay-Franco (marie.tremblay-franco@toulouse.inra.fr) and Marion Landi (marion.landi@clermont.inra.fr)
MetaboHUB: The French National Infrastructure for Metabolomics and Fluxomics (http://www.metabohub.fr/en)

Please cite

F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn (2006). Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. Analytical Chemistry 78(13): 4281-4290.

Normalization

Description

Normalization (operation applied on each (preprocessed) individual spectrum) of preprocessed data

Workflow position

Upstream tools

Name	output file	format	parameter
NMR_Bucketing	Normalization_bucketedData.tsv	tabular	Ions Matrix

Downstream tools

Name	Output file	Format
Univariate	variableMetadata.tsv	Tabular

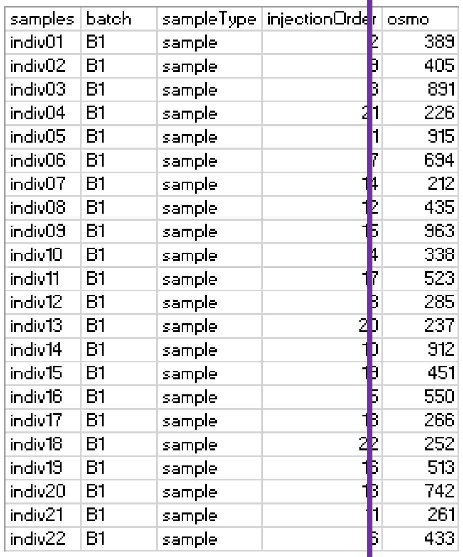
Outputs:

- dataMatrix normalised by the chosen method
- A pdf file with the chosen Spectra representation

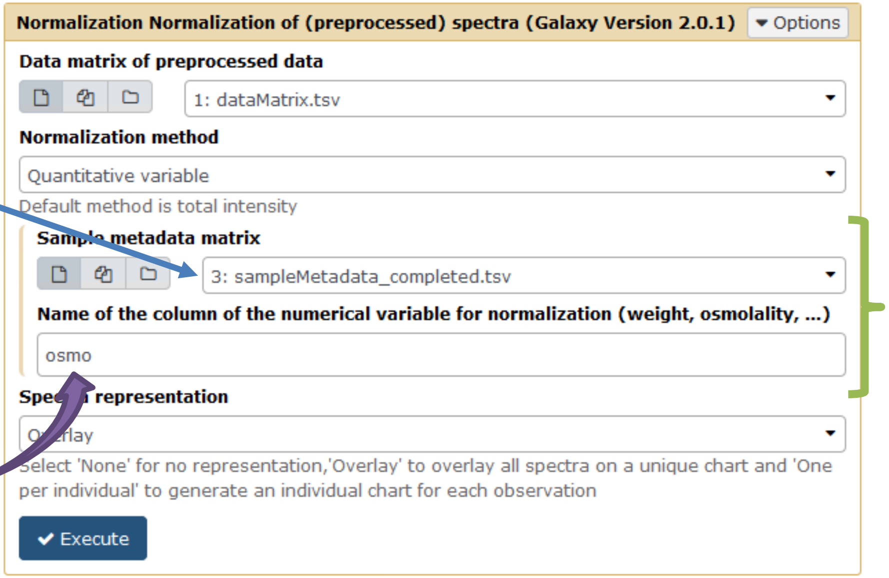
Example: quantitative variable

- Intensity of each feature is divided by the value of a known quantitative variable: weight for tissue, osmolality, …

Sample metadata file



samples	batch	sampleType	injectionOrder	osmo
indiv01	B1	sample	2	389
indiv02	B1	sample	9	405
indiv03	B1	sample	3	891
indiv04	B1	sample	21	226
indiv05	B1	sample	1	915
indiv06	B1	sample	7	694
indiv07	B1	sample	14	212
indiv08	B1	sample	12	435
indiv09	B1	sample	15	963
indiv10	B1	sample	4	338
indiv11	B1	sample	17	523
indiv12	B1	sample	8	285
indiv13	B1	sample	20	237
indiv14	B1	sample	10	912
indiv15	B1	sample	19	451
indiv16	B1	sample	5	550
indiv17	B1	sample	18	266
indiv18	B1	sample	22	252
indiv19	B1	sample	16	513
indiv20	B1	sample	13	742
indiv21	B1	sample	11	261
indiv22	B1	sample	6	433



Normalization Normalization of (preprocessed) spectra (Galaxy Version 2.0.1)

Data matrix of preprocessed data

1: dataMatrix.tsv

Normalization method

Quantitative variable

Default method is total intensity

Sample metadata matrix

3: sampleMetadata_completed.tsv

Name of the column of the numerical variable for normalization (weight, osmolality, ...)

osmo

Spectra representation

Overlay

Select 'None' for no representation,'Overlay' to overlay all spectra on a unique chart and 'One per individual' to generate an individual chart for each observation

mandatory for "Quantitative Variable" and "PQN normalization"