Data preprocessing in non-targeted metabolomics – best practices and pitfalls

Workshop at the Metabolomics 2023, Niagara Falls, Canada, June 19, 2023

Kati Hanhineva, Ville Koistinen, Topi Meuronen, Retu Haikonen

Food Sciences unit, Department of Life Technologies, University of Turku

Institute of Public Health and Clinical Nutrition, University of Eastern Finland

UNIVERSITY OF EASTERN FINLAND

UNIVERSITY OF TURKU

Contents

- Introduction to data preprocessing and notame pipeline
- Demonstration of the pipeline
- Discussion and feedback


notame
RAW DATA
NOISE
SIGNALS
DRIFT
NH
OH

Introduction

- *Notame* (”non-targeted metabolomics”) is an R-based tool to preprocess and clean metabolomics data
- The first version of notame was written by Anton Klåvus for his master's thesis in Bioinformatics at Aalto university (published under former name Anton Mattsson), while working for University of Eastern Finland and Afekta Technologies
- The key elements in notame are to
  - assess the general quality of the data
  - prepare the raw data matrix for statistical analysis/feature selection
  - provide meaningful visualizations on the results

UNIVERSITY OF TURKU

MS-DIAL

Data extraction & spreadsheet formation — MS-DIAL Spreadsheet

Pre-processing

Drift correction	`correct_drift()`	**Quality control**	`flag_detection()` `flag_quality()`
Imputation	`impute_rf()` *missForest*	**Feature clustering**	`cluster_features()`

Feature-wise analysis

Parametric analysis

2 groups	Welch's *t*-test
> 2 groups	Welch's ANOVA
2 factors	Linear mode with interaction
Repeated measurements	Linear mixed effects model

Non-parametric analysis

2 groups	Mann–Whitney *U* test
> 2 groups	Kruskal–Wallis *H* test
2 factor	Friedman test

Multivariate analysis

Exploratory analysis

PCA — *pcaMethods*
t-SNE — *Rtsne*
Hierarchical clustering

Classification & regression

PLS(-DA)
Random forest
} *MUVR*

Transformation & normalisation

notame

Visualisation of the data and key results

Feature selection and identification — Spreadsheet MS-DIAL

http://prime.psc.riken.jp/compms/msdial/main.html

Metabolomics data preprocessing in *notame* R package

W10: Data preprocessing in non-targeted metabolomics – best practices and pitfalls

Ville Koistinen, Topi Meuronen, Retu Haikonen

UNIVERSITY OF EASTERN FINLAND

UNIVERSITY OF TURKU

Overview of the workshop

- Basic R skills are useful to follow this tutorial workshop and to use notame – but we try to make it simple
- These slides contain simplified screenshots of the script, which is available on our group's website and GitHub
- What you need: untargeted LC-MS or GC-MS data, MS-DIAL (or similar), RStudio, Excel

Workshop material
GitHub link

https://hanhinevalab.com/
→ Material & Tools

UNIVERSITY OF TURKU

Further information

https://www.mdpi.com/2218-1989/10/4/135

https://github.com/antonvsdata/notame

Publication

Notame GitHub link

metabolites

MDPI

Protocol

"Notame": Workflow for Non-Targeted LC–MS Metabolic Profiling

Anton Klåvus [1,*,†], Marietta Kokla [1,*,†], Stefania Noerman [1], Ville M. Koistinen [1], Marjo Tuomainen [1], Iman Zarei [1], Topi Meuronen [1], Merja R. Häkkinen [2], Soile Rummukainen [2], Ambrin Farizah Babu [1], Taisa Sallinen [1,2], Olli Kärkkäinen [2], Jussi Paananen [3], David Broadhurst [4], Carl Brunius [5,6] and Kati Hanhineva [1,5,7,*]

1 Department of Clinical Nutrition and Public Health, University of Eastern Finland, 70210 Kuopio, Finland; stefania.noerman@uef.fi (S.N.); ville.m.koistinen@uef.fi (V.M.K.); marjo.tuomainen@uef.fi (M.T.); iman.zarei@uef.fi (I.Z.); topi.meuronen@uef.fi (T.M.); ambbab@student.uef.fi (A.F.B.); taisa.sallinen@uef.fi (T.S.)
2 School of Pharmacy, University of Eastern Finland, 70210 Kuopio, Finland; merja.hakkinen@uef.fi (M.R.H.); soile.rummukainen@uef.fi (S.R.); olli.karkkainen@uef.fi (O.K.)
3 Institute of Biomedicine, University of Eastern Finland, 70210 Kuopio, Finland; jussi.paananen@uef.fi
4 Centre for Integrative Metabolomics & Computational Biology, School of Science, Edith Cowan University, Joondalup, WA 6027, Australia; d.broadhurst@ecu.edu.au
5 Department of Biology and Biological Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden; carl.brunius@chalmers.se
6 Chalmers Mass Spectrometry Infrastructure, Chalmers University of Technology, 41296 Gothenburg, Sweden
7 Department of Biochemistry, Food Chemistry and Food Development unit, University of Turku, 20014 Turun yliopisto, Finland
* Correspondence: anton.klavus@uef.fi (A.K.); marietta.kokla@uef.fi (M.K.); kati.hanhineva@uef.fi (K.H.)
† These authors have contributed equally to this work.

Received: 2 March 2020; Accepted: 28 March 2020; Published: 31 March 2020

check for updates

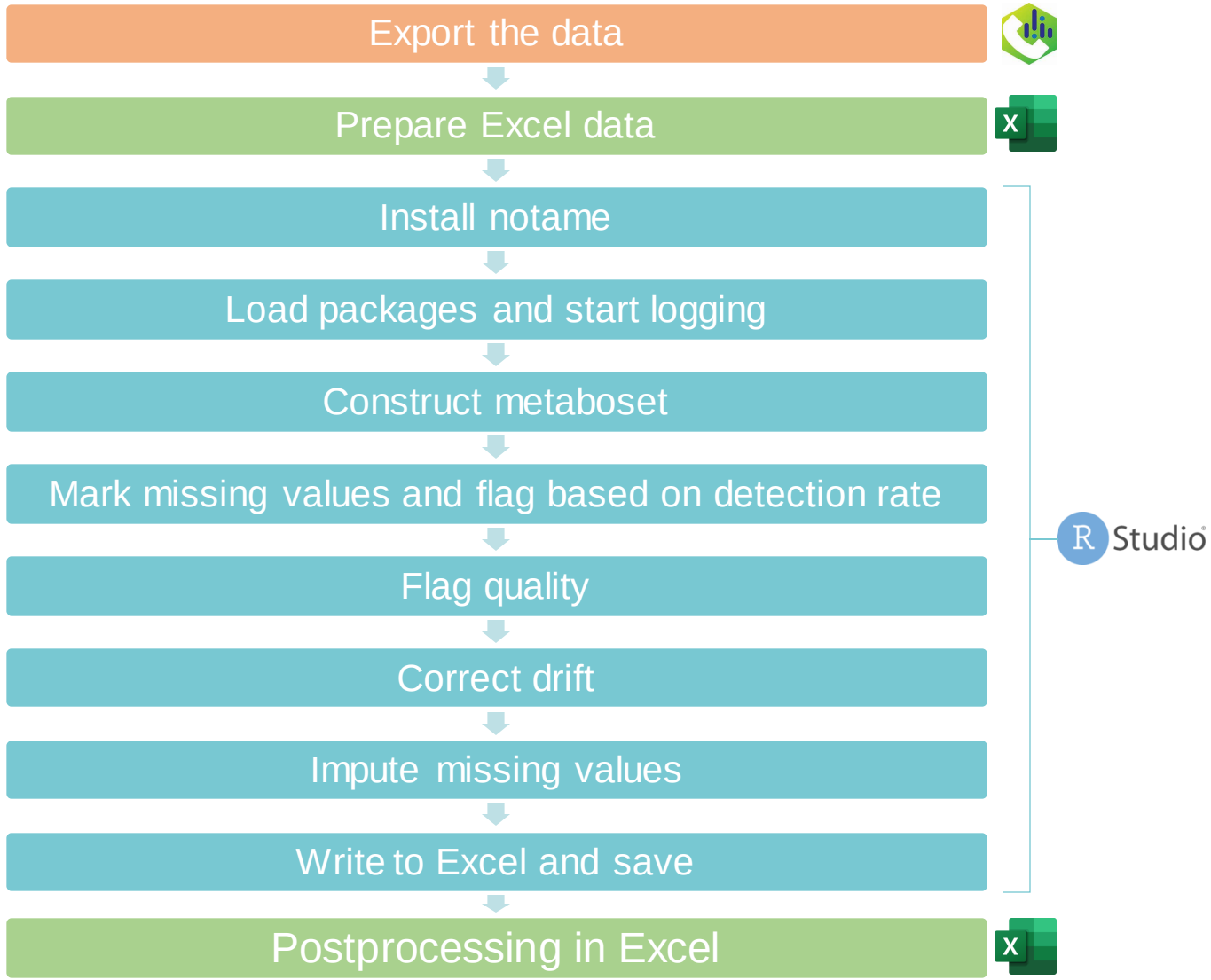
Abstract: Metabolomics analysis generates vast arrays of data, necessitating comprehensive workflows involving expertise in analytics, biochemistry and bioinformatics in order to provide coherent and high-quality data that enable discovery of robust and biologically significant metabolic findings. In this protocol article, we introduce notame, an analytical workflow for non-targeted metabolic profiling approaches, utilizing liquid chromatography–mass spectrometry analysis. We provide an overview of lab protocols and statistical methods that we commonly practice for the analysis of nutritional metabolomics data. The paper is divided into three main sections: the first and second sections introducing the background and the study designs available for metabolomics research and the third section describing in detail the steps of the main methods and protocols used to produce, preprocess and statistically analyze metabolomics data and, finally, to identify and interpret the compounds that have emerged as interesting.

Keywords: metabolomics; LC–MS; mass spectrometry; metabolic profiling; computational statistical; unsupervised learning; supervised learning; pathway analysis

1. Introduction

The rapid technical development of instrumentation for biomolecule analysis has led to a wide application of metabolomics in biological and biomedical research. Due to its very high sensitivity and the ability to concomitantly assess thousands of molecular features, liquid chromatography coupled with mass spectrometry (LC–MS) is making its way as the key analytical tool in the field of discovery-driven metabolic profiling [1–3]. The LC–MS platform generates large amounts of

General preprocessing workflow and software used

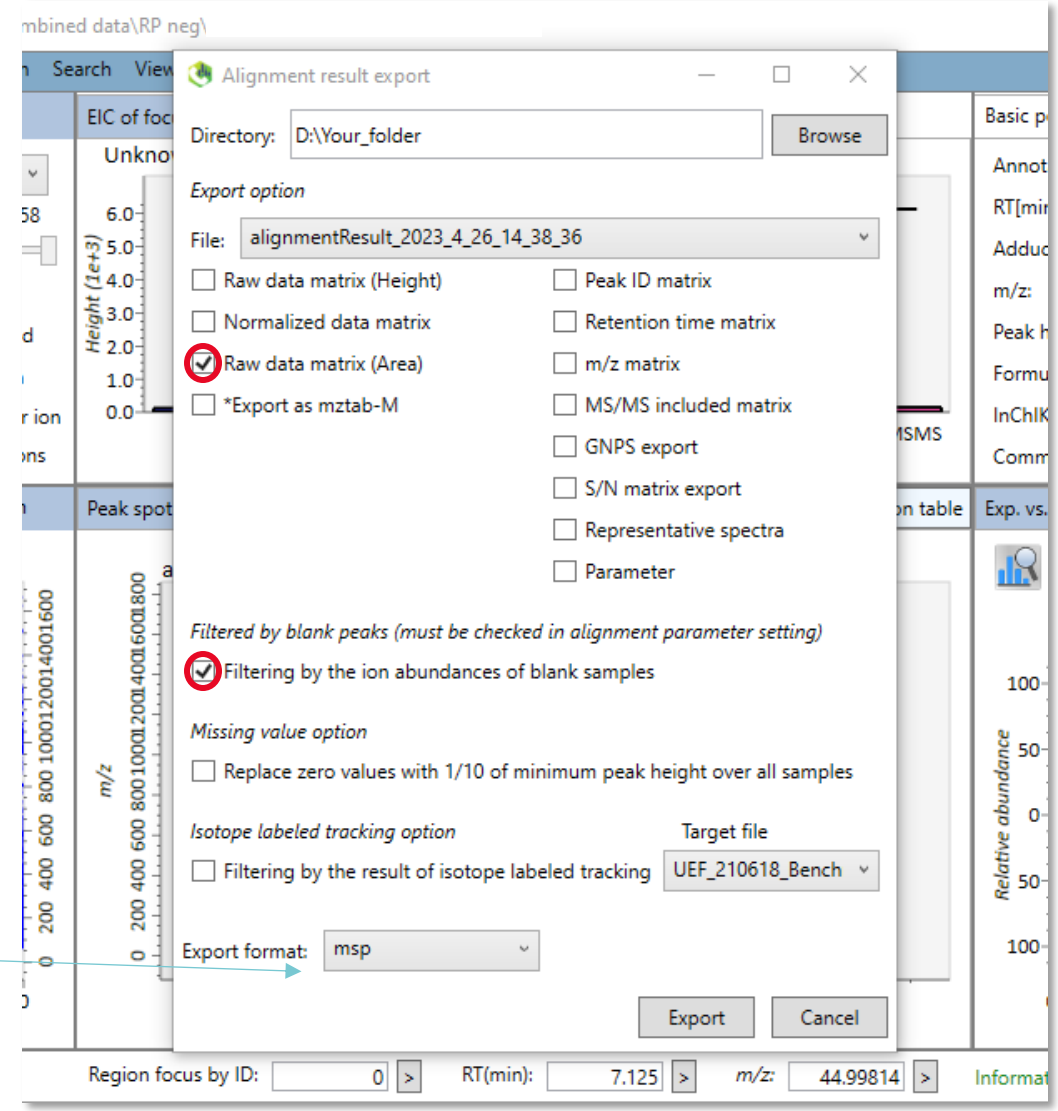


UNIVERSITY OF TURKU

MS-DIAL

- Notame accepts the alignment export directly from MS-DIAL
- Note that you need to keep track of the chromatographic /ion modes if you have several in your data

The ”msp” export format is actually a tab-separated CSV file, which can be opened in Excel



UNIVERSITY OF TURKU

Prepare Excel data

- The data will be used to construct a MetaboSet object with the following structure:
  - **exprs:** feature abundances across the samples
  - **pheno_data:** sample information
  - **feature_data:** feature information
- You can combine data from different modes in Excel or later in R
  - In Excel, you can make a new column describing the mode of each feature (e.g., RP pos, RP neg, HILIC pos, HILIC neg), for example named Split
  - Make sure to match the sequence across modes

notame will automatically try to find this "corner" where the expression values begin

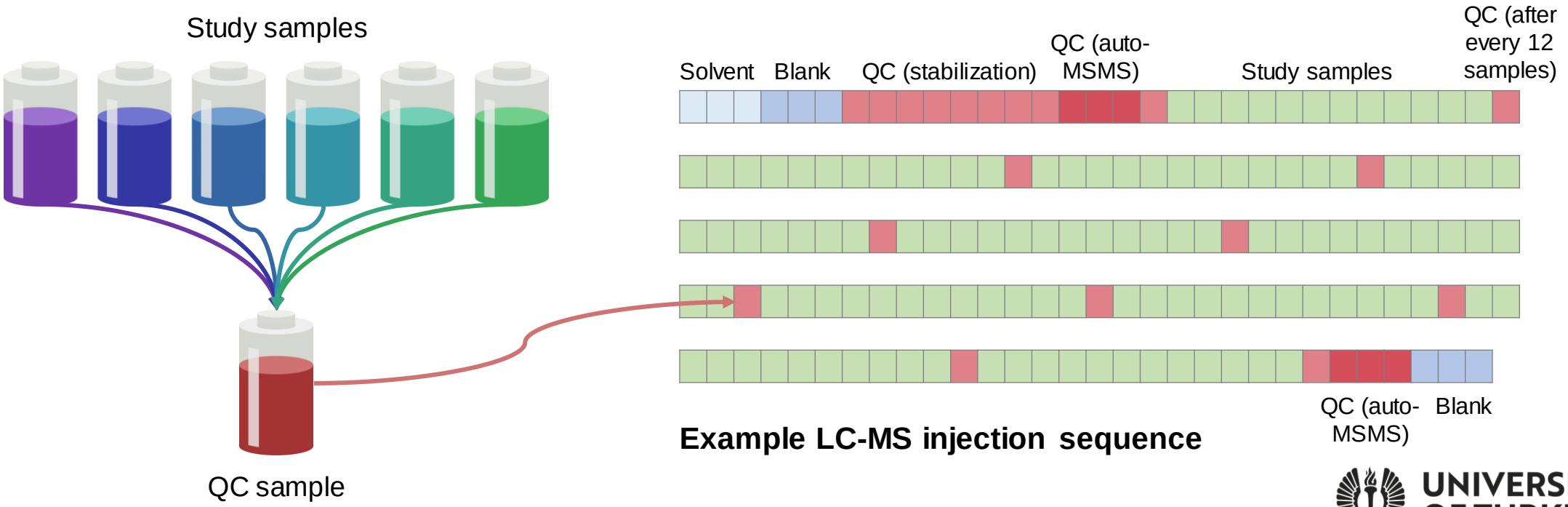
Example of a suitable data format

Keep this area empty

				Class		A	A	A	B	B	B
				File type	QC	Sample	Sample	Sample	Sample	Sample	Sample
				Injection order	1	2	3	4	5	6	7
				Batch ID	1	1	1	1	1	1	1
Alignment ID	Average Rt(min)	Average Mz	Metabolite name	Split	File_1	File_2	File_3	File_4	File_5	File_6	File_7
1	14.05	100.0793	Metabolite 1	RP pos	15389	16004	67816	37654	2976	2867	1085
2	8.78	112.1583	Metabolite 2	RP pos	1654516	6549684	4352415	1986545	6854646	4632184	7874546
3	2.16	132.0428	Metabolite 3	RP pos	4594146	8242334	3261199	1728190	2475302	3164750	9225945
4	3.53	140.0467	Metabolite 4	RP pos	4821951	113017	690884	2961649	6238804	3533544	1099746
5	8.35	153.9692	Metabolite 5	RP pos	8701229	9870760	91094	4073828	6291024	5607944	710761
6	10.37	165.0105	Metabolite 6	RP pos	8029117	6695877	910399	3880946		1644123	5253212
7	9.01	165.0351	Metabolite 7	RP pos	419517	120	7489125	792169	5787382	952883	8570666
8	4.77	177.0273	Metabolite 8	RP pos	3664172	9280985	8763212	2334167	6399593	7653679	9248892

The QC samples

- The *study-specific* QC sample is pooled from the actual samples by collecting 5–25 µL aliquots to make a sample that represents the whole sample set and can be used for quality control purposes, such as drift correction
- Ideally, the QC sample contains *in detectable levels* all the metabolites that the actual samples do—the less similar the samples are, the less this is true



UNIVERSITY OF TURKU

Now, let's go to R...

UNIVERSITY OF TURKU

Install notame in R

- https://github.com/antonvsdata/notame

```
#1. Package installations

#1a. notame

if (!requireNamespace("devtools", quietly = TRUE)) {
  install.packages("devtools")
}
devtools::install_github("antonvsdata/notame", dependencies = c("Depends", "Imports", "Suggests"))
```

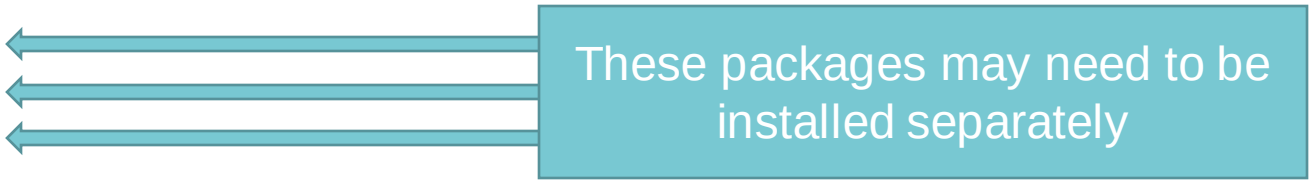
This script is available in our group webpage:
https://hanhinevalab.com

UNIVERSITY OF TURKU

Load packages, set project path, logging

- We recommend using logging to keep track of the package versions etc.

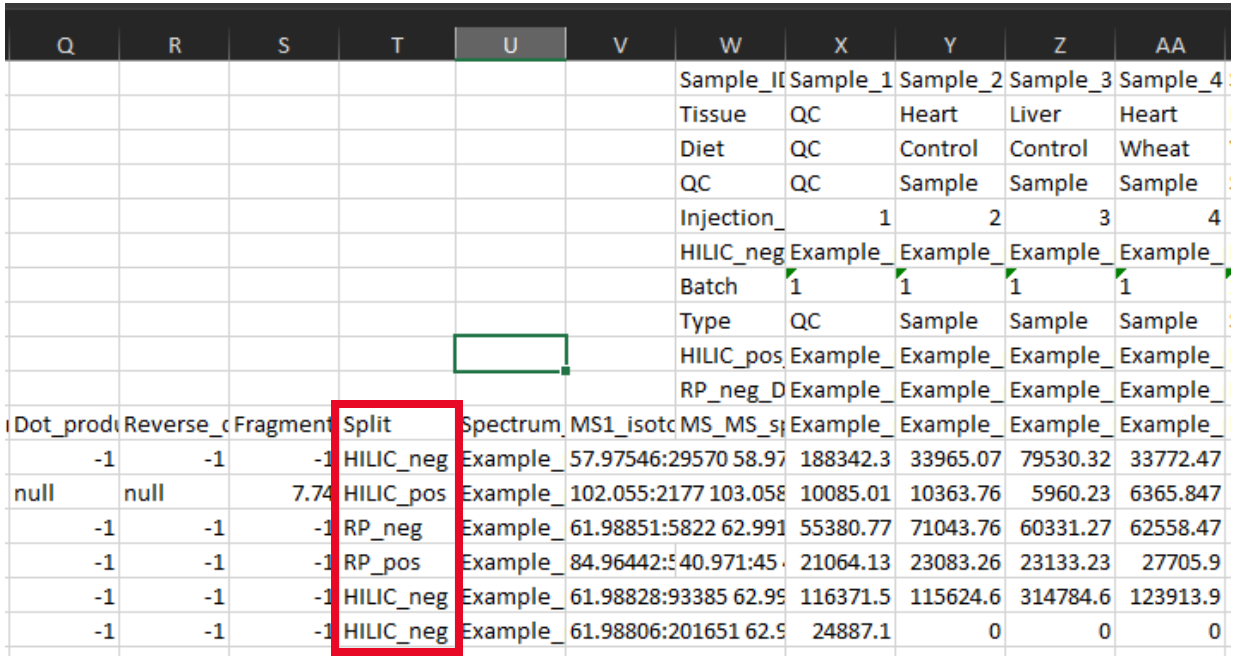
```
59  #2. Load libraries, set up path and logging
60  library(notame)
61  library(doParallel)
62  library(dplyr)
63  library(openxlsx)
64
65  path <- "G:/"
66  init_log(log_file = paste0(path, "log.txt"))
67  #> Logging started: ~/test_project/log.txt.
68  # Check logging state
69  log_state()
70  #> Current log file: ~/test_project/log.txt
71
```



Construct metaboset (option A: combined Excel file)

- Use this if your Excel file has several modes combined and you have a column containing the mode information

```
71
72  #3. Read data (note: assumption is that the data contains signals from two or more modes)
73  data <- read_from_excel(file = "Metaboset_to_metsoc.xlsx", split_by = "Split")
74
75  #Check how it looks
76  names(data)
77  sapply(data, class)
78  sapply(data, dim)
79
80  #4a. Construct MetaboSet objects
81  modes <- construct_metabosets(
82    exprs = data$exprs,
83    pheno_data = data$pheno_data,
84    feature_data = data$feature_data,
85    group_col = "Tissue", time_col = "Diet")
86
```



Q	R	S	T	U	V	W	X	Y	Z	AA
						Sample_ID	Sample_1	Sample_2	Sample_3	Sample_4
						Tissue	QC	Heart	Liver	Heart
						Diet	QC	Control	Control	Wheat
						QC	QC	Sample	Sample	Sample
						Injection_	1	2	3	4
						HILIC_neg	Example_	Example_	Example_	Example_
						Batch	1	1	1	1
						Type	QC	Sample	Sample	Sample
						HILIC_pos	Example_	Example_	Example_	Example_
						RP_neg_D	Example_	Example_	Example_	Example_
Dot_prod	Reverse_c	Fragment	Split	Spectrum_	MS1_isotc	MS_MS_sp	Example_	Example_	Example_	Example_
-1	-1	-1	HILIC_neg	Example_	57.97546:29570	58.97	188342.3	33965.07	79530.32	33772.47
null	null	7.74	HILIC_pos	Example_	102.055:2177	103.058	10085.01	10363.76	5960.23	6365.847
-1	-1	-1	RP_neg	Example_	61.98851:5822	62.991	55380.77	71043.76	60331.27	62558.47
-1	-1	-1	RP_pos	Example_	84.96442:5	40.971:45	21064.13	23083.26	23133.23	27705.9
-1	-1	-1	HILIC_neg	Example_	61.98828:93385	62.99	116371.5	115624.6	314784.6	123913.9
-1	-1	-1	HILIC_neg	Example_	61.98806:201651	62.9	24887.1	0	0	0

Construct metaboset (option B: one file per mode)

- Use this if you have several Excel outputs from different modes
- The example below combines the Excel files (inside a folder called ”data”) if you name them RP_pos_sample.xlsx etc.

```
#3b. Alternative: Read in the separate modes and construct MetaboSet objects
# (note: here we read in individual modefiles)

modes <- c("HILIC_neg", "HILIC_pos", "RP_neg", "RP_pos")
objects <- list()
for (mode in modes) {
  # Read single mode
  filename <- paste0(ppath, "Data/", mode, "_sample.xlsx")
  data <- read_from_excel(file = filename,
                          name = mode)
  objects[[mode]] <- construct_metabosets(exprs = data$exprs,
                                          pheno_data = data$pheno_data,
                                          feature_data = data$feature_data,
                                          group_col = "Tissue", time_col = "Diet")[[1]]
}
```

UNIVERSITY OF TURKU

Mark missing values and flag based on detection rate

- By default, a feature with a detection rate **less than 70% in the QC samples** or **less than 80% in every group** is flagged

```
#Take several cores into use

cl <- makeCluster(8)
registerDoParallel(cl)

# Initialize empty list for processed objects
processed <- list()
for (i in seq_along(modes)) {
  name <- names(modes)[i]
  mode <- modes[[i]]
  # Set all zero abundances to NA
  mode <- mark_nas(mode, value = 0)
  mode <- flag_detection(mode, qc_limit = 0.7, group_limit = 0.8)

  corrected <- correct_drift(mode)

  corrected <- corrected %>% assess_quality() %>% flag_quality()
  processed[[i]] <- corrected
}
#Stop using several cores (releases them for other use)
stopCluster(cl)

# Merge the modes
merged <- merge_metabosets(processed)

#visualizations(merged, prefix = paste0(path, "figures/_FULL"))
```

Mark missing values

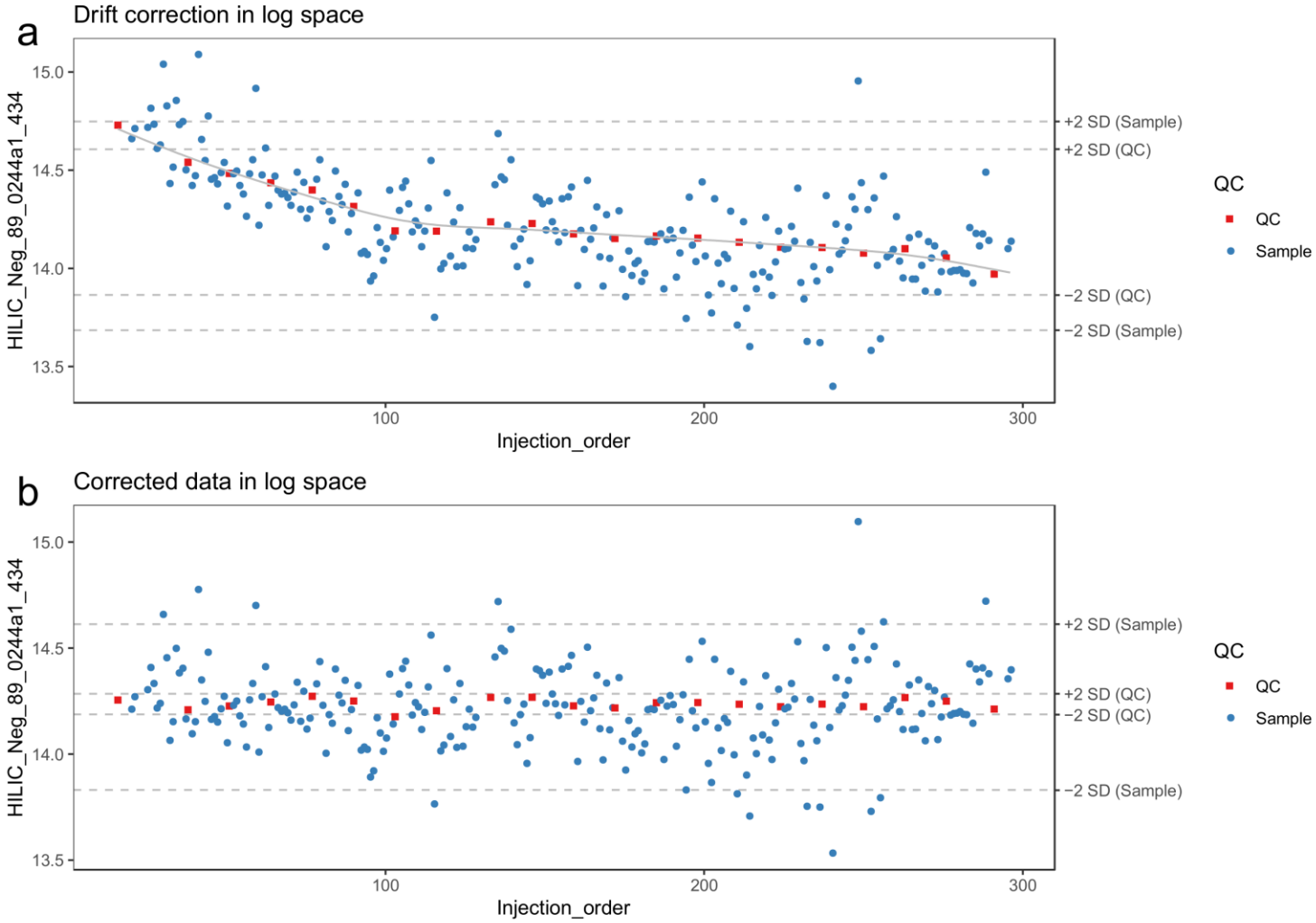
Set the detection rate thresholds

Note that group_limit refers to the group_col information in pdata

UNIVERSITY OF TURKU

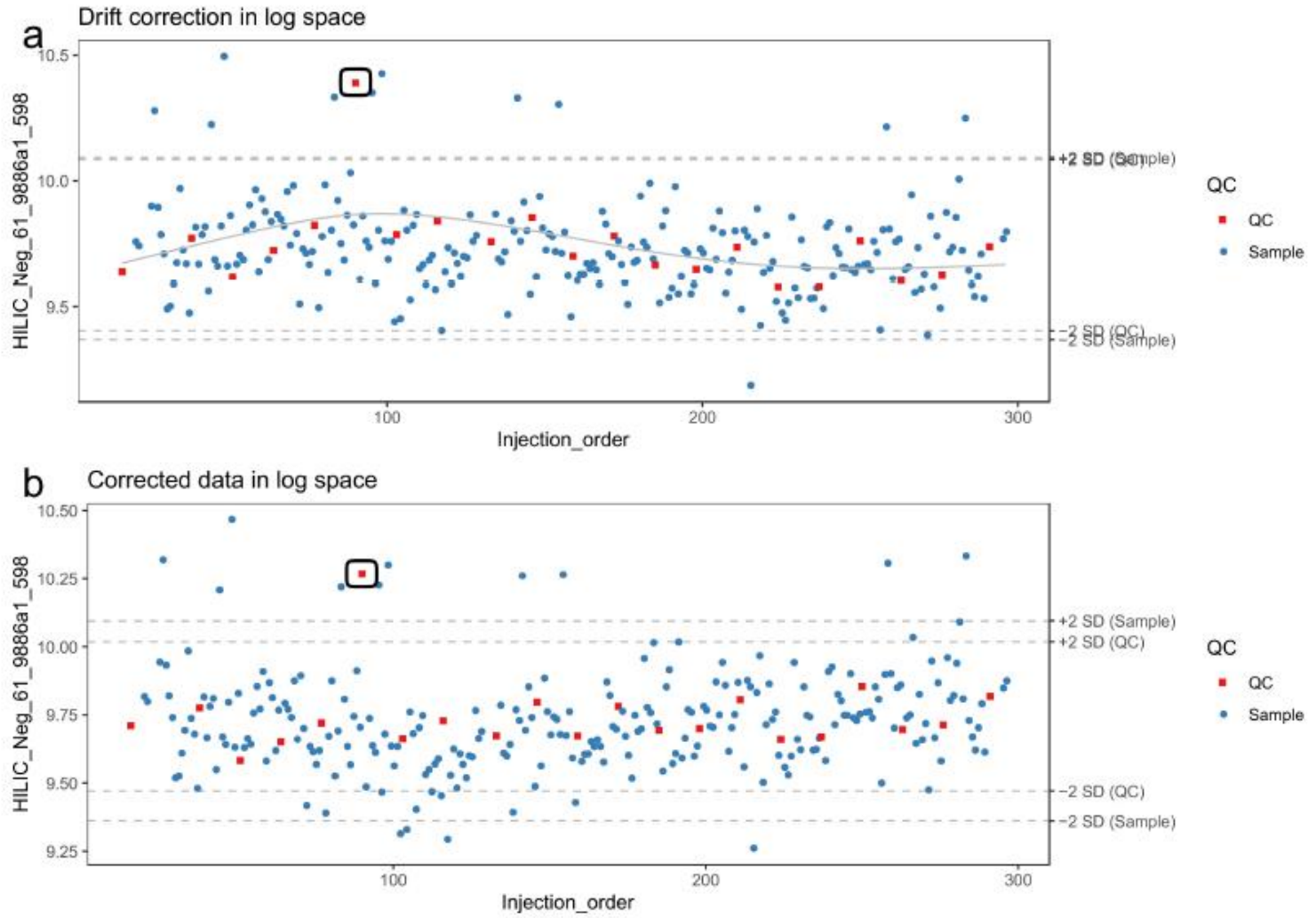
Correct drift

- Notame corrects signal intensity drift often present in long LC-MS sequences
- It fits a smoothed cubic spline curve across the QC samples to stratify the signals of the samples
- Done for every feature individually



Correct drift

- Due to the smoothing, the correction method is robust against the deviating QC sample and adjusts seemingly adequately for the global drift trend.
- Viant et al. 2013



Correct drift

```
108
109 #Take several cores into use
110
111 cl <- makeCluster(8)
112 registerDoParallel(cl)
113
114 # Initialize empty list for processed objects
115 processed <- list()
116 for (i in seq_along(modes)) {
117   name <- names(modes)[i]
118   mode <- modes[[i]]
119   # Set all zero abundances to NA
120   mode <- mark_nas(mode, value = 0)
121   mode <- flag_detection(mode, qc_limit = 0.7, group_limit = 0.8)
122
123   corrected <- correct_drift(mode)
124
125   corrected <- corrected %>% assess_quality() %>% flag_quality()
126   processed[[i]] <- corrected
127 }
128 #Stop using several cores (releases them for other use)
129 stopCluster(cl)
130
131 # Merge the modes
132 merged <- merge_metabosets(processed)
133
134 #visualizations(merged, prefix = paste0(path, "figures/_FULL"))
135
```

Drift correction

Flag quality

- The quality parameters are based on Broadhurst et al. 2018
- A feature is *not* flagged for low quality if the RSD and D-ratio conditions are met

```
#Take several cores into use

cl <- makeCluster(8)
registerDoParallel(cl)

# Initialize empty list for processed objects
processed <- list()
for (i in seq_along(modes)) {
  name <- names(modes)[i]
  mode <- modes[[i]]
  # Set all zero abundances to NA
  mode <- mark_nas(mode, value = 0)
  mode <- flag_detection(mode, qc_limit = 0.7, group_limit = 0.8)

  corrected <- correct_drift(mode)

  corrected <- corrected %>% assess_quality() %>% flag_quality()
  processed[[i]] <- corrected
}
#Stop using several cores (releases them for other use)
stopCluster(cl)

# Merge the modes
merged <- merge_metabosets(processed)

#visualizations(merged, prefix = paste0(path, "figures/_FULL"))
```

The exact conditions:
RSD_r < 0.2 & D_ratio_r < 0.4
or
RSD < 0.1 & RSD_r < 0.1 & D_ratio < 0.1

The default values are in-built into this function

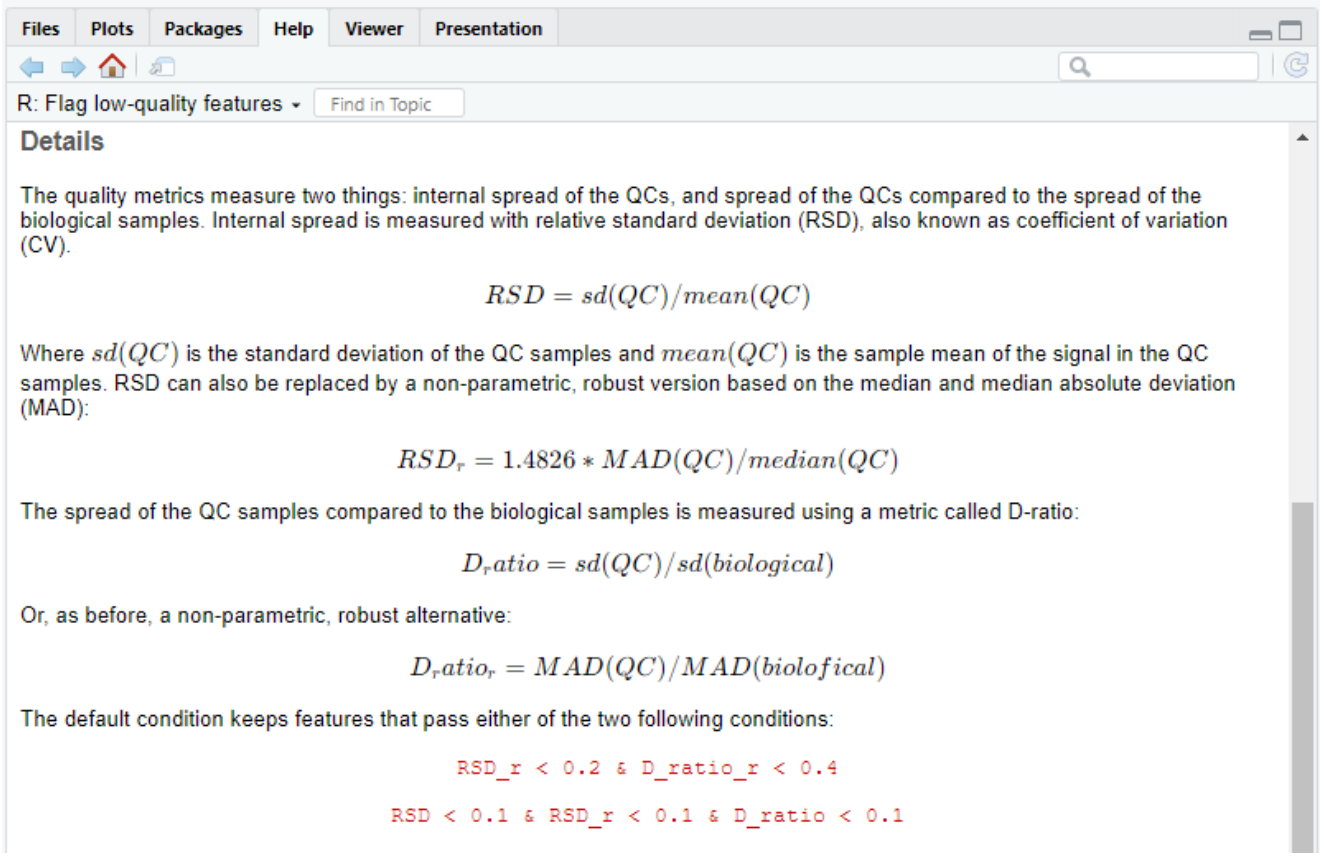
Merge all preprocessed modes before imputation

UNIVERSITY OF TURKU

Quality more information

Quality metrics are calculated based on B roadhurst *et al.*, *Metabolomics* 2018:

- **Detection rate** in the QC samples (*e.g.*, 70 %)
- Relative standard deviation (**RSD**)
- Non-parametric RSD (**RSD***)
- **D-ratio** the spread of the QC samples compared to the biological samples
- Non-parametric D-ratio (**D-ratio***)



Non-parametric quality metrics are more resilient to outliers in the dataset

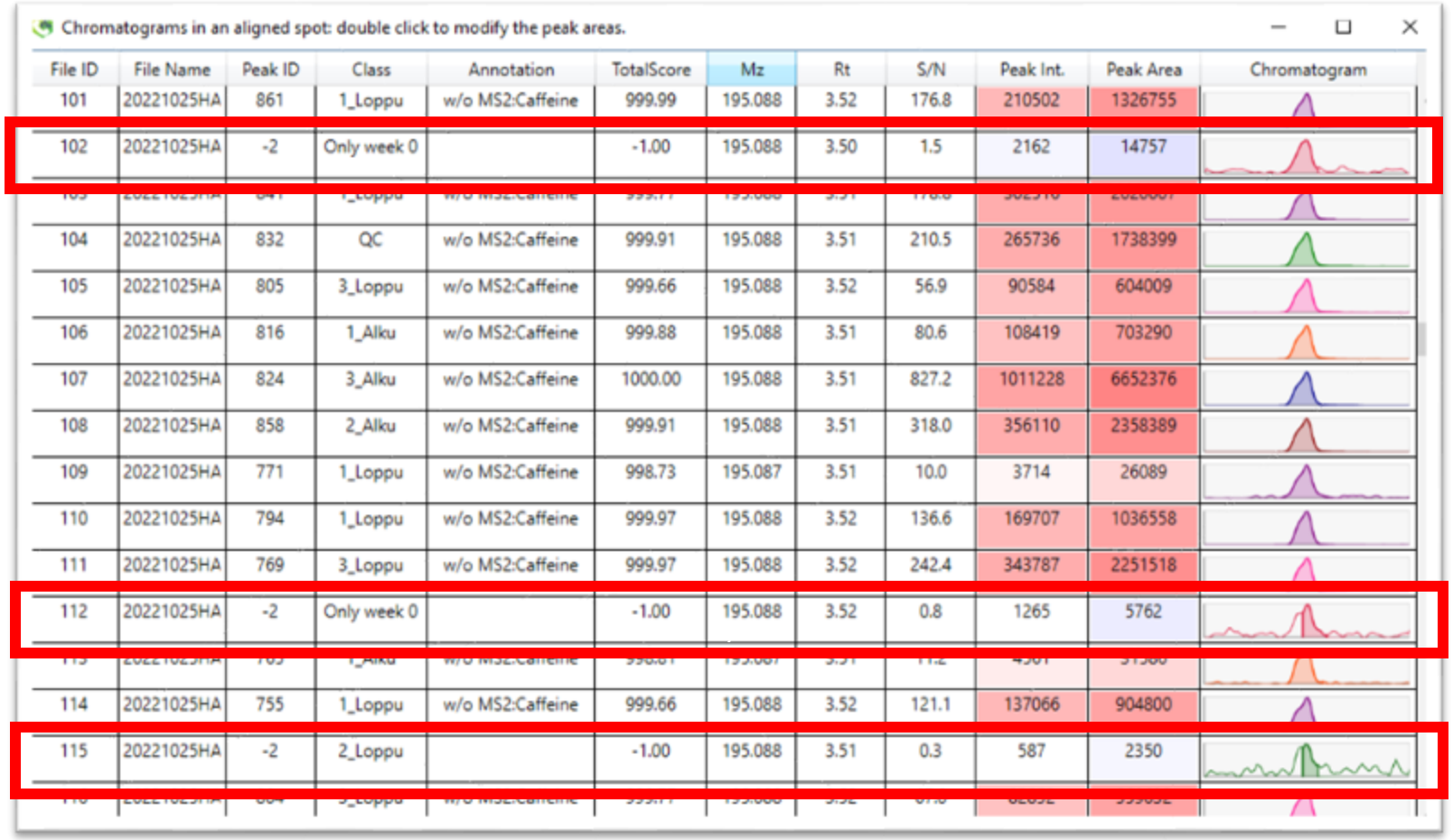
UNIVERSITY OF TURKU

Impute missing values

- Missing Completely at Random (MCAR)
  - "This mechanism is characterized by a randomness in the occurrence of missing data, which could also be accounted as a **zero correlation between the missing and the observed part in the data**"
- Missing at Random (MAR)
  - "systematic correlation between the missing molecular features and the observed data, but not with the missing data itself, meaning that the **probability of a molecular feature being missing is determined by other observed molecular features**"
- Missing Not at Random (MNAR)
  - "**value of the molecular feature is causal for missingness and that is why it is missing**, and this type of missing data is usually characterized in many metabolomics studies as left truncated data (molecular features occur below the detection limit)"

Reference: Kokla et al. 2019

Missing values: MS-DIAL and gap filling



File ID	File Name	Peak ID	Class	Annotation	TotalScore	Mz	Rt	S/N	Peak Int.	Peak Area	Chromatogram
101	20221025HA	861	1_Loppu	w/o MS2:Caffeine	999.99	195.088	3.52	176.8	210502	1326755	
102	20221025HA	-2	Only week 0		-1.00	195.088	3.50	1.5	2162	14757	
104	20221025HA	832	QC	w/o MS2:Caffeine	999.91	195.088	3.51	210.5	265736	1738399	
105	20221025HA	805	3_Loppu	w/o MS2:Caffeine	999.66	195.088	3.52	56.9	90584	604009	
106	20221025HA	816	1_Alku	w/o MS2:Caffeine	999.88	195.088	3.51	80.6	108419	703290	
107	20221025HA	824	3_Alku	w/o MS2:Caffeine	1000.00	195.088	3.51	827.2	1011228	6652376	
108	20221025HA	858	2_Alku	w/o MS2:Caffeine	999.91	195.088	3.51	318.0	356110	2358389	
109	20221025HA	771	1_Loppu	w/o MS2:Caffeine	998.73	195.087	3.51	10.0	3714	26089	
110	20221025HA	794	1_Loppu	w/o MS2:Caffeine	999.97	195.088	3.52	136.6	169707	1036558	
111	20221025HA	769	3_Loppu	w/o MS2:Caffeine	999.97	195.088	3.52	242.4	343787	2251518	
112	20221025HA	-2	Only week 0		-1.00	195.088	3.52	0.8	1265	5762	
114	20221025HA	755	1_Loppu	w/o MS2:Caffeine	999.66	195.088	3.52	121.1	137066	904800	
115	20221025HA	-2	2_Loppu		-1.00	195.088	3.51	0.3	587	2350	

- "Peak recognition is performed by the average peak width of samples having the metabolite feature even though no local maximum is observed in the chromatogram"
- https://mtbinfo-team.github.io/mtbinfo.github.io/MS-DIAL/tutorial.html

UNIVERSITY OF TURKU

Impute missing values

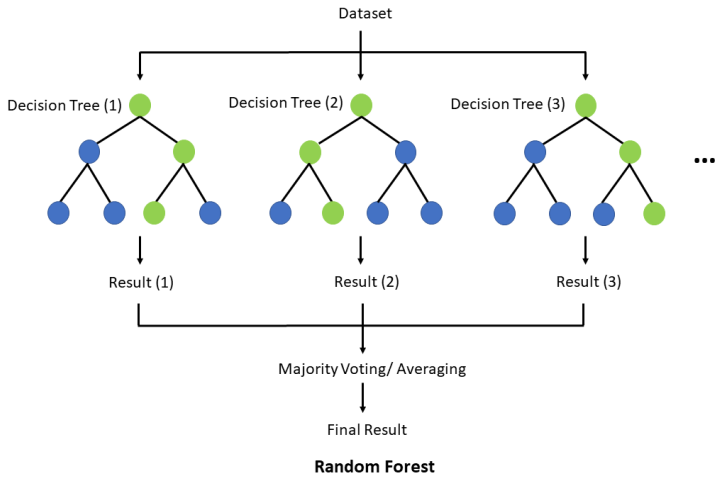
R: Impute missing values using random forest ▾ Find in Topic

impute_rf {notame} R Documentation

Impute missing values using random forest

Description

Impute the missing values in the exprs part of the object using a random forest. The estimated error in the imputation is logged. It is recommended to set the seed number for reproducibility (it is called random forest for a reason). This a wrapper around `missForest::missForest`. Use parallelize = "variables" to run in parallel for faster testing. NOTE: running in parallel prevents user from setting a seed number. **CITATION:** When using this function, cite the `missForest` package

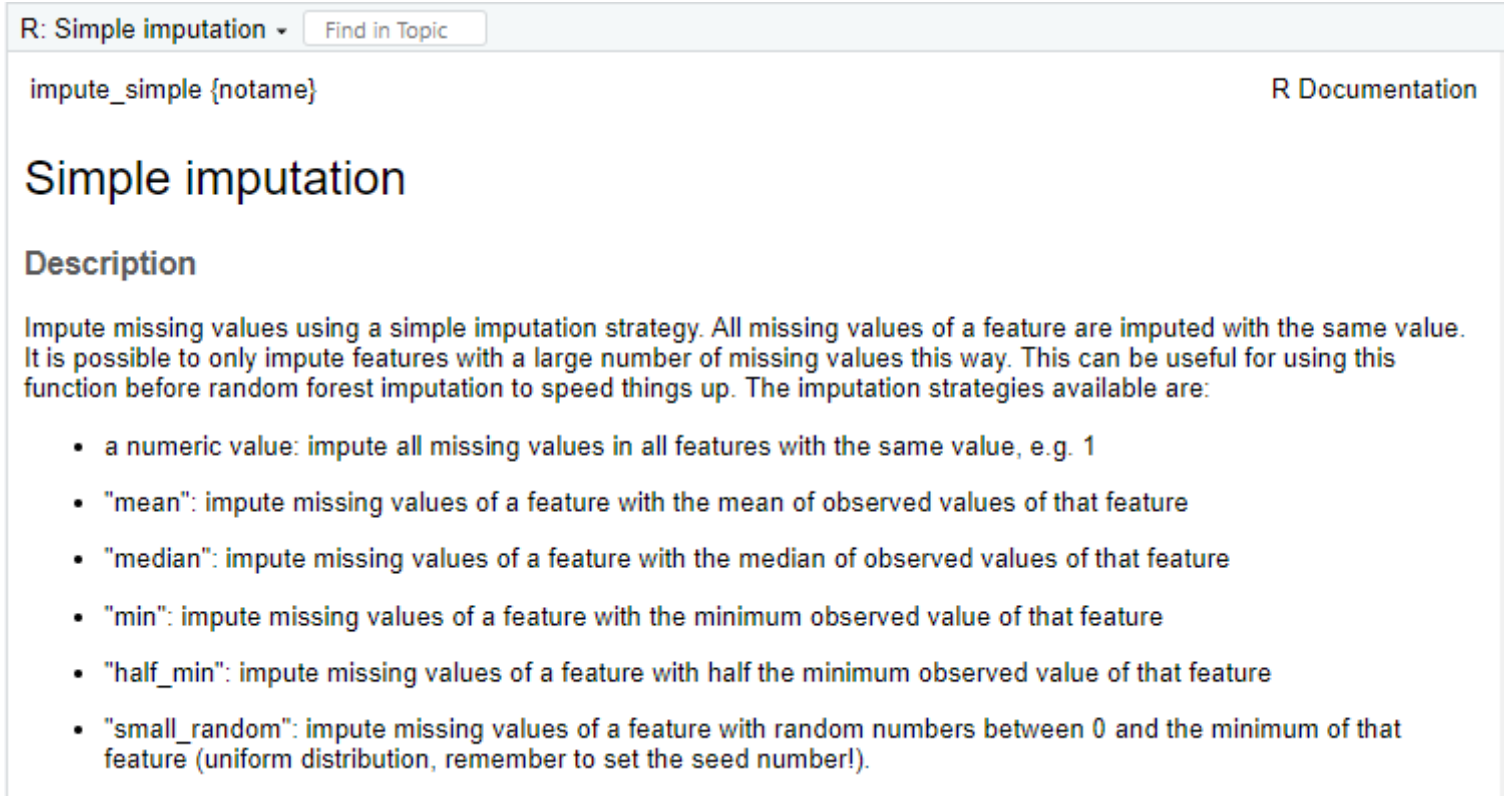


- "RF-based imputation method performs best in most of the tested scenarios, including combinations of different types and rates of missingness. Therefore, we recommend using RF-based imputation for imputing missing metabolomics data, since typically the origin of missingness is not known in advance."

Reference: Kokla et al. 2019

UNIVERSITY OF TURKU

Impute missing values



R: Simple imputation Find in Topic

impute_simple {notame} R Documentation

Simple imputation

Description

Impute missing values using a simple imputation strategy. All missing values of a feature are imputed with the same value. It is possible to only impute features with a large number of missing values this way. This can be useful for using this function before random forest imputation to speed things up. The imputation strategies available are:

- a numeric value: impute all missing values in all features with the same value, e.g. 1
- "mean": impute missing values of a feature with the mean of observed values of that feature
- "median": impute missing values of a feature with the median of observed values of that feature
- "min": impute missing values of a feature with the minimum observed value of that feature
- "half_min": impute missing values of a feature with half the minimum observed value of that feature
- "small_random": impute missing values of a feature with random numbers between 0 and the minimum of that feature (uniform distribution, remember to set the seed number!).

- Possibility to use also other common imputation methods includings:
  - Mean
  - Median
  - Min
  - Half of the minimum
  - Random number

UNIVERSITY OF TURKU

Impute missing values

- Imputation is necessary for many statistical models if the data contains missing values
- The seed number ensures that the imputation is reproducible (it's a process including random elements)

Remove QCs at this point (not needed anymore)

```
148
149  # Remove the QCs for imputation
150  merged_no_qc <- drop_qcs(merged)
151
152  #visualizations(merged_no_qc, prefix = paste0(path, "figures/FULL_NO_QC"))
153
154  #10. Imputation (note: may not be necessary especially if gap filling by compulsion was
155  #Set seed number for reproducibility
156
157  set.seed(38)
158  imputed <- impute_rf(merged_no_qc, all_features = FALSE)
159  imputed <- impute_rf(imputed, all_features = TRUE)
160
```

If your data contains samples from several species/tissues/etc, the missing values can be because of actual biological differences → be more careful with the imputation

Imputation of good-quality features based on good-quality features

Imputation for flagged features based on all features

UNIVERSITY OF TURKU

Write to excel and save object for later use

- Due to tehcnical limitations, the resulting Excel will have lost any style or format settings from the original file and contains numbers stored as text (can be fixed in Excel)

```
178
179 #11. Write to Excel (note: update the object and file name according to project)
180
181 write_to_excel(imputed, file = paste0(path, "preprosessed_excel.xlsx"))
182
183 #11 b. Save data in RDS format
184
185 saveRDS(imputed, file = paste0(path, "preprosessed_example.RDS"))
186
```

After the preprocessing

Visualizations

Excel tips

Statistics

UNIVERSITY OF TURKU

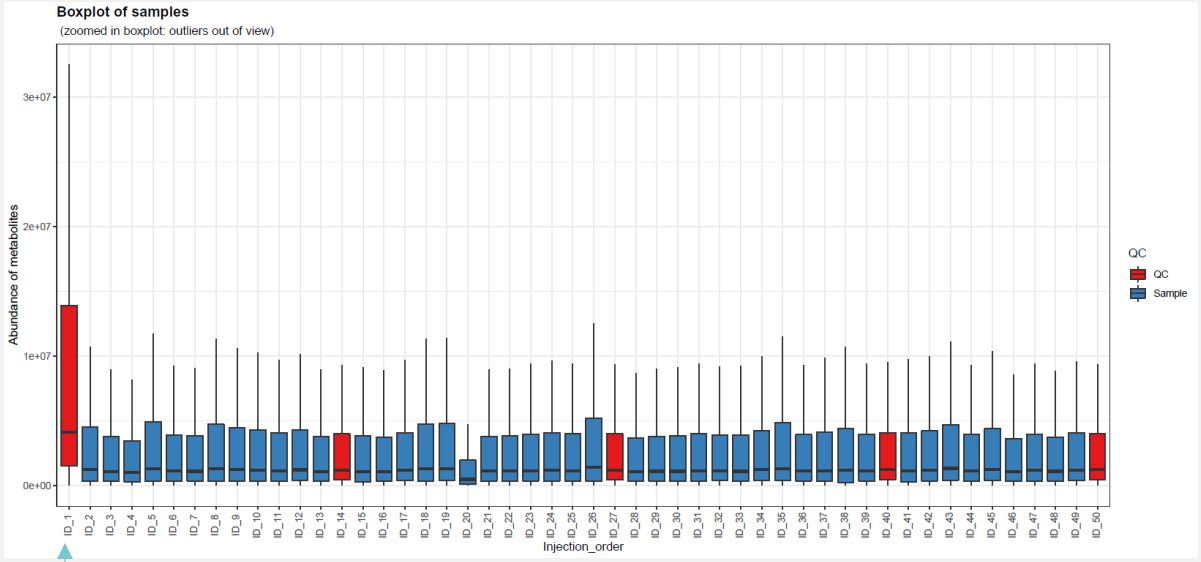
Visualization output from preprocessing

- The script creates optional visualizations that describe the data quality
  - Box plots of the samples arranged by group or injection order
  - Dendrogram of the hierarchically clustered samples
  - Density plot and heatmap of the differences between samples
  - Linear model *p*-value histograms
  - PCA by group or injection order
  - Quality metrics histograms
- The visualizations are made from the original data, drift-corrected data, and final cleaned-up data (good-quality signals) to allow a checkup on how the preprocessing has affected data quality

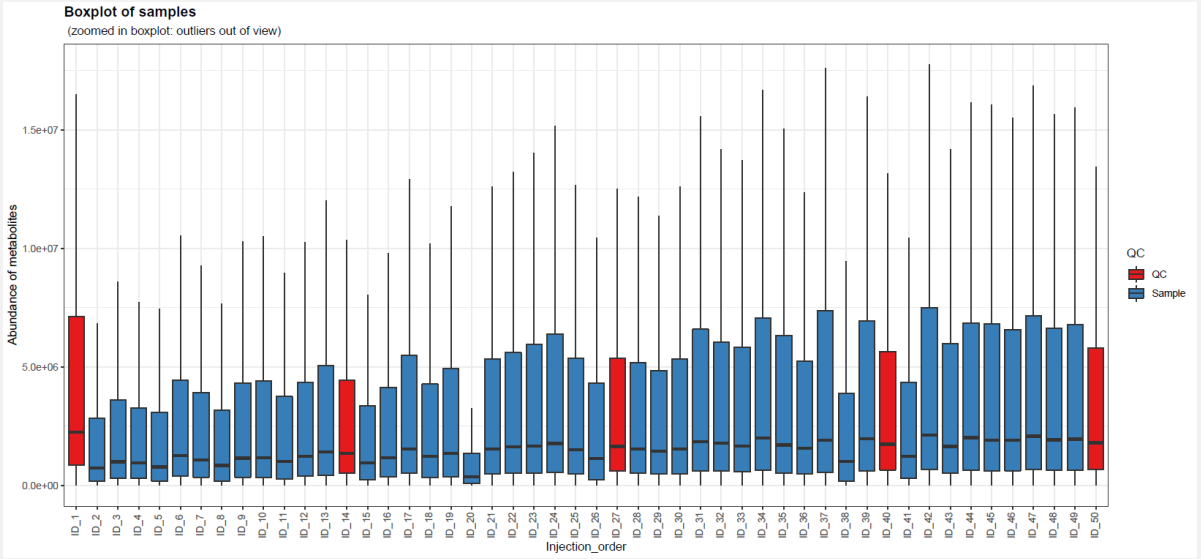
Box plots

- The abundance of molecular features in every sample

Before (notice the outlier QC)



After



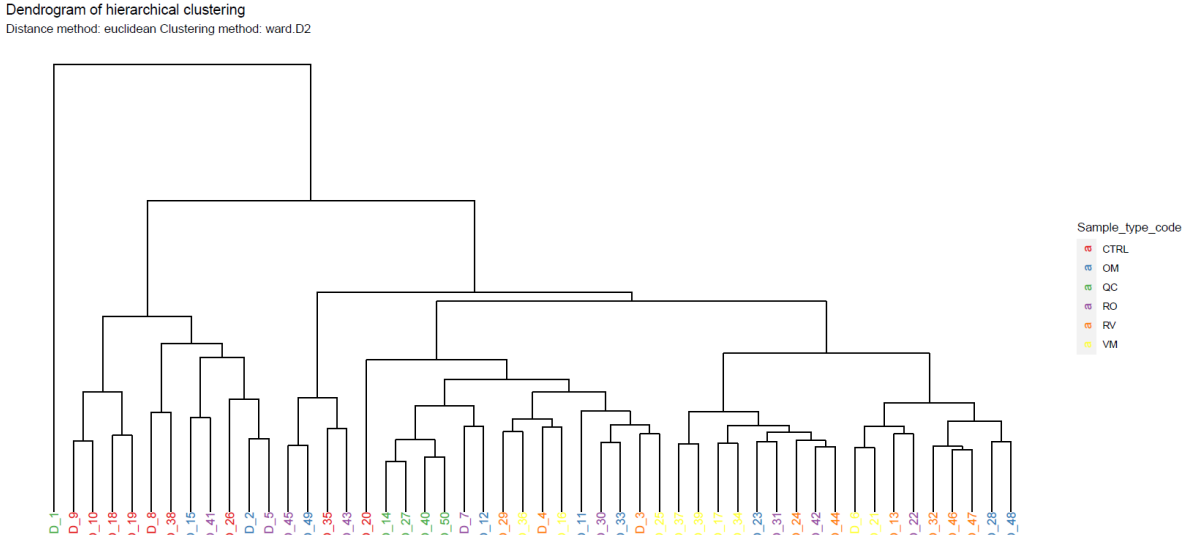
Notice the outlier QC in the data and how it will be treated

UNIVERSITY OF TURKU

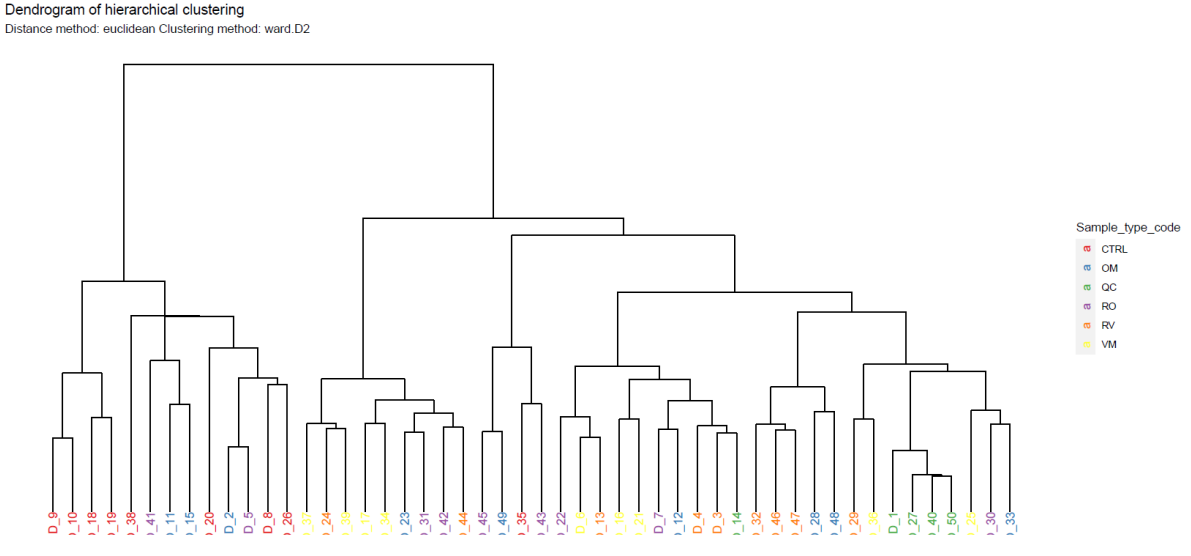
Dendrogram

- Hierarchical tree based on similarities between samples

Before



After

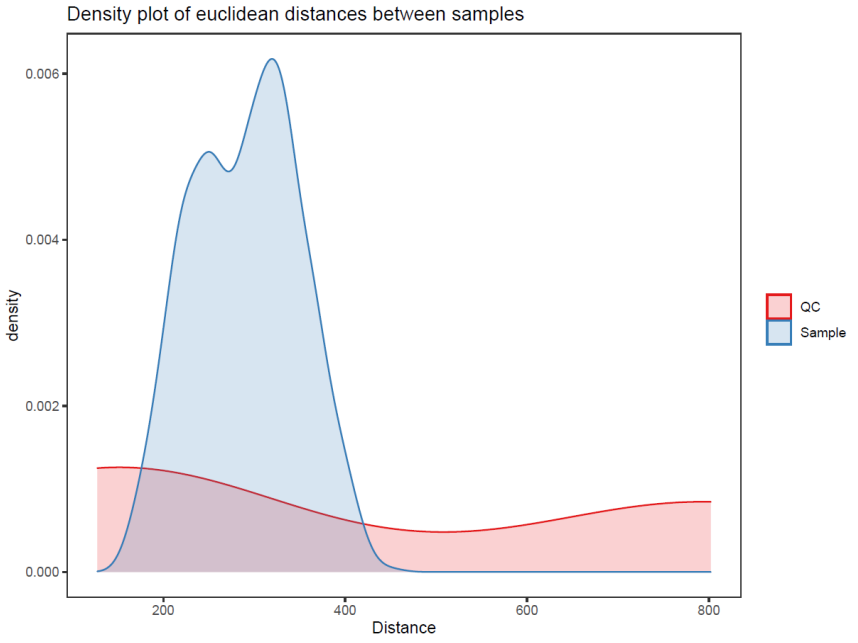


UNIVERSITY OF TURKU

Density plot

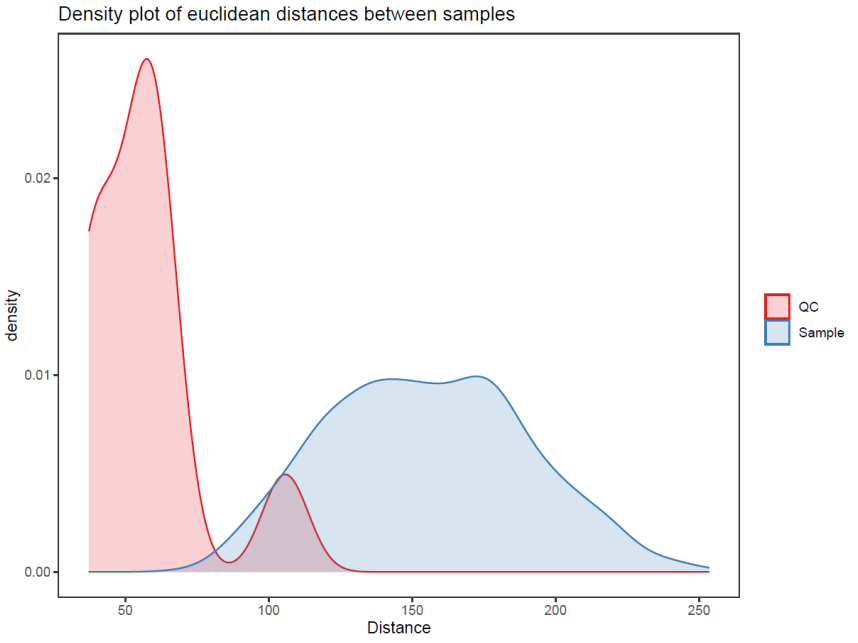
- Similarities between the samples (QC / actual sample) in a Euclidean distance matrix

Before



The QCs should be very similar

After

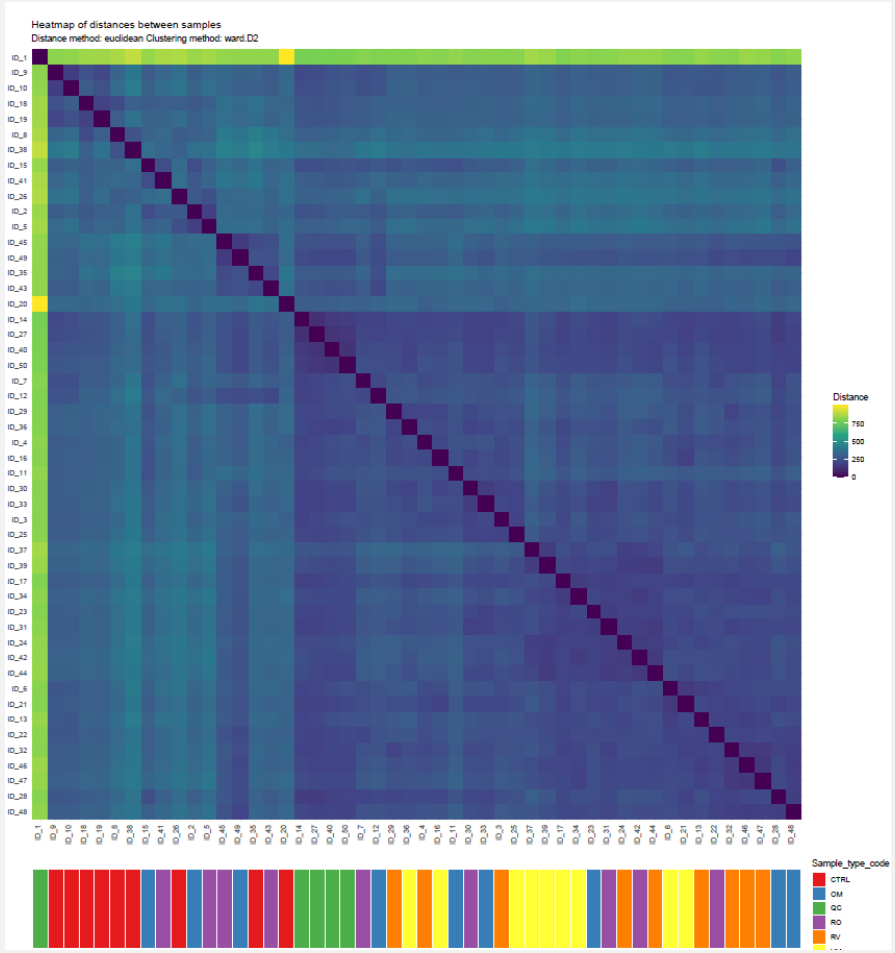


UNIVERSITY OF TURKU

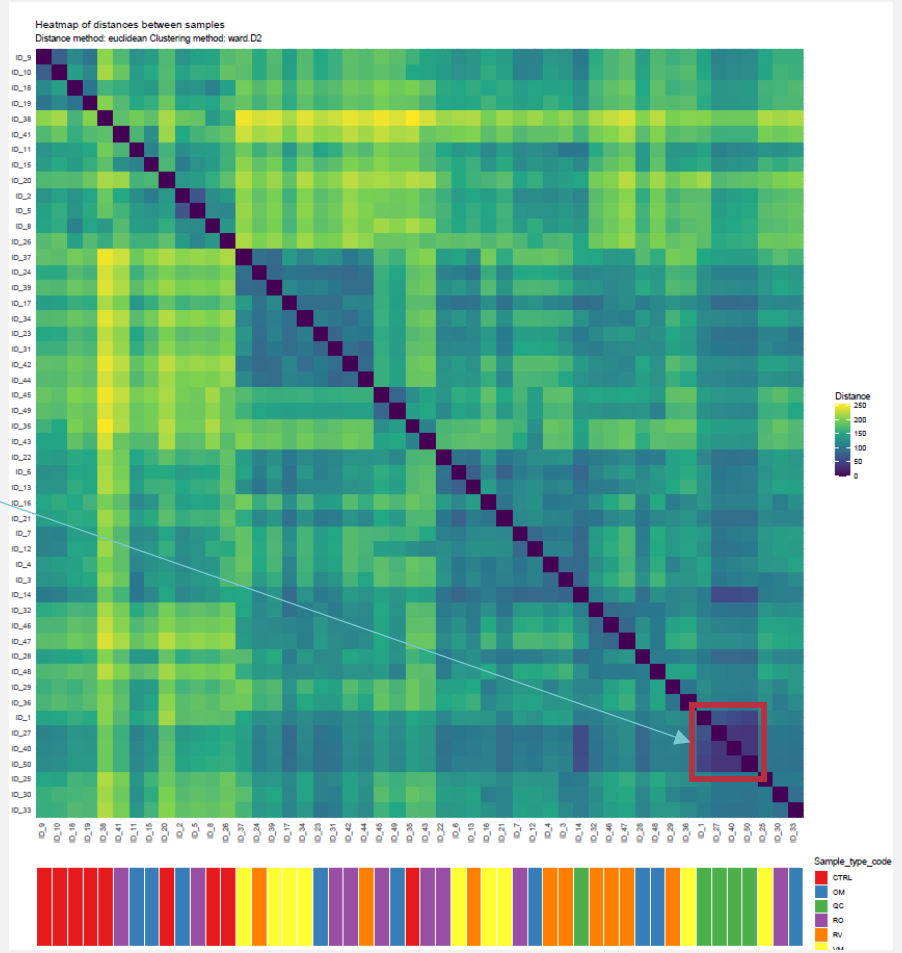
Heatmap

- Similarities between the samples

Before



After

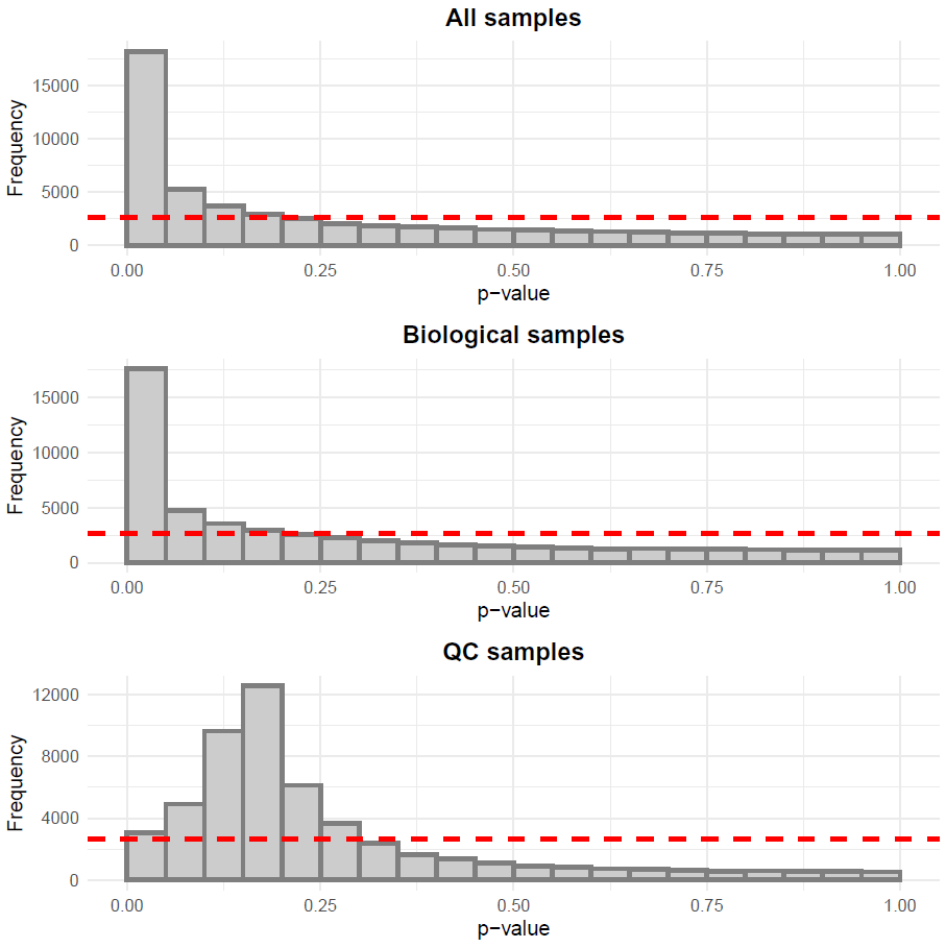


The QCs should form a dark rectangle

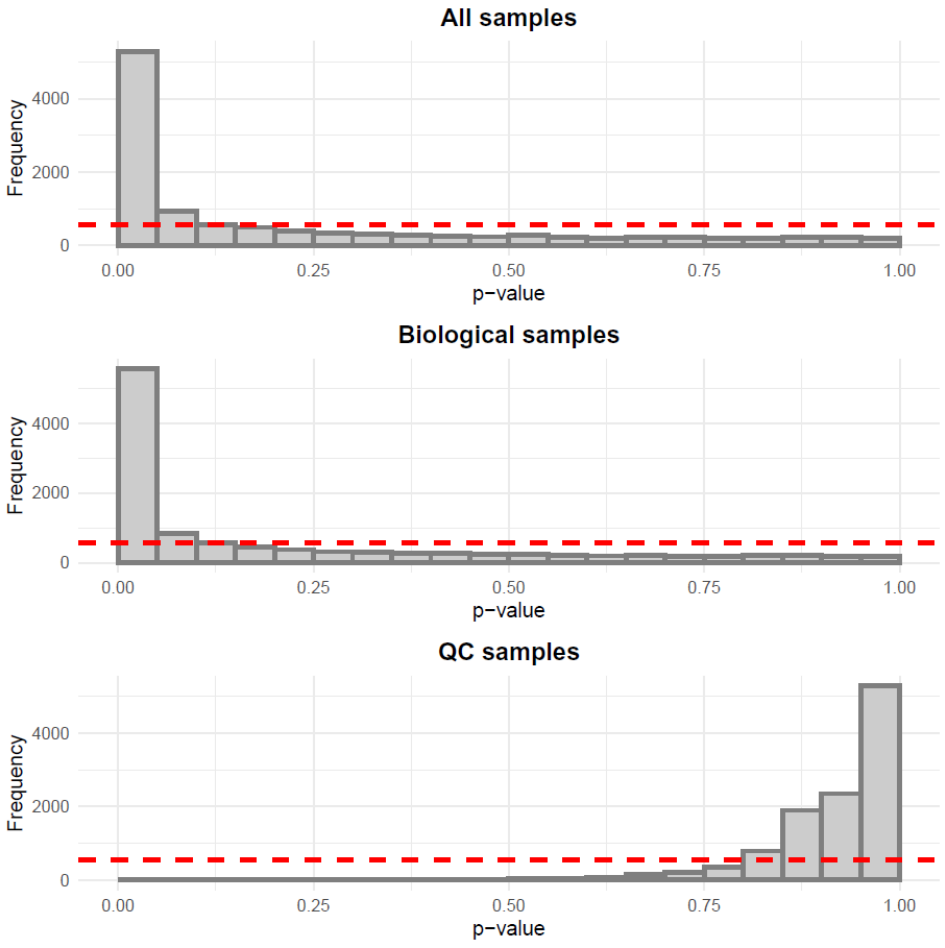
p-Value histograms

- Distribution of group-wise differences

Before



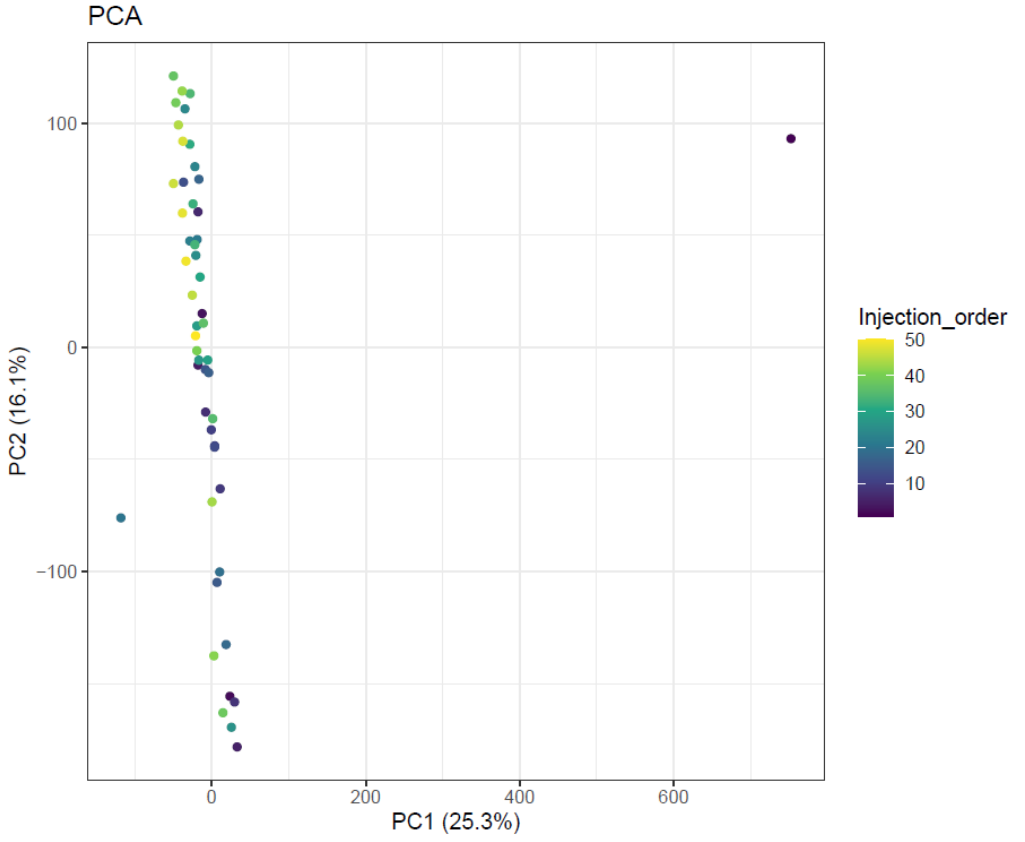
After



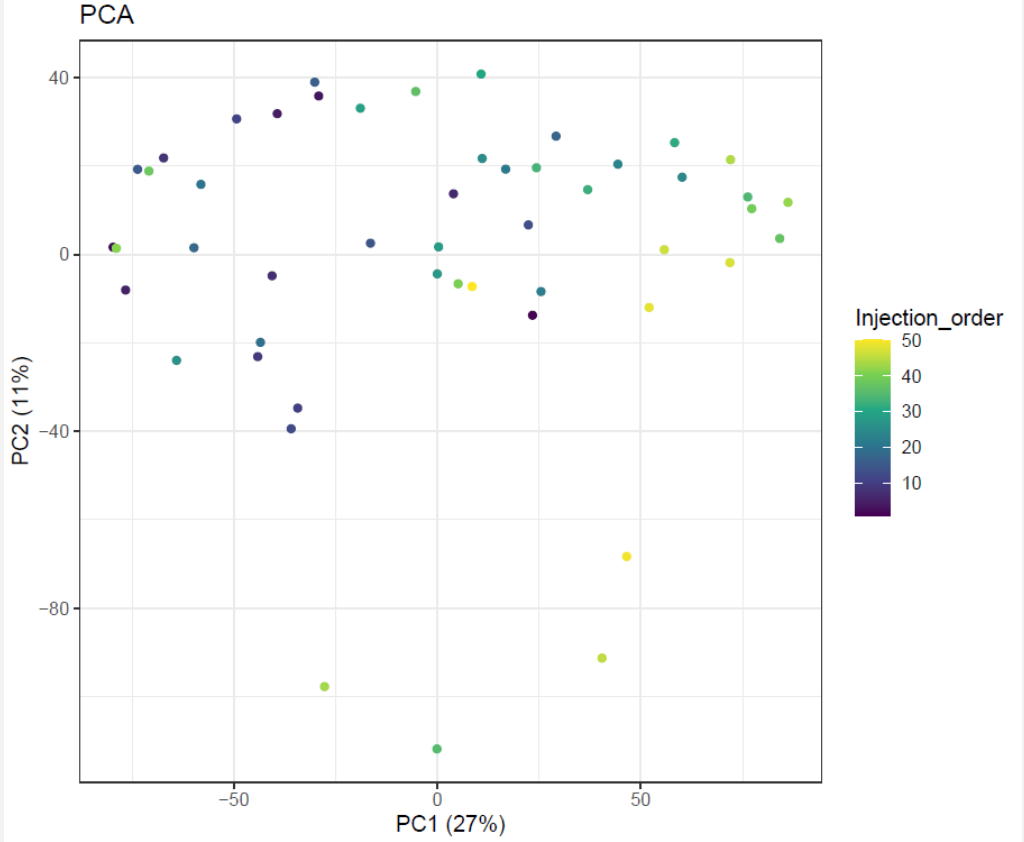
PCA by injection order

- Reveals issues from the signal drift and outliers

Before



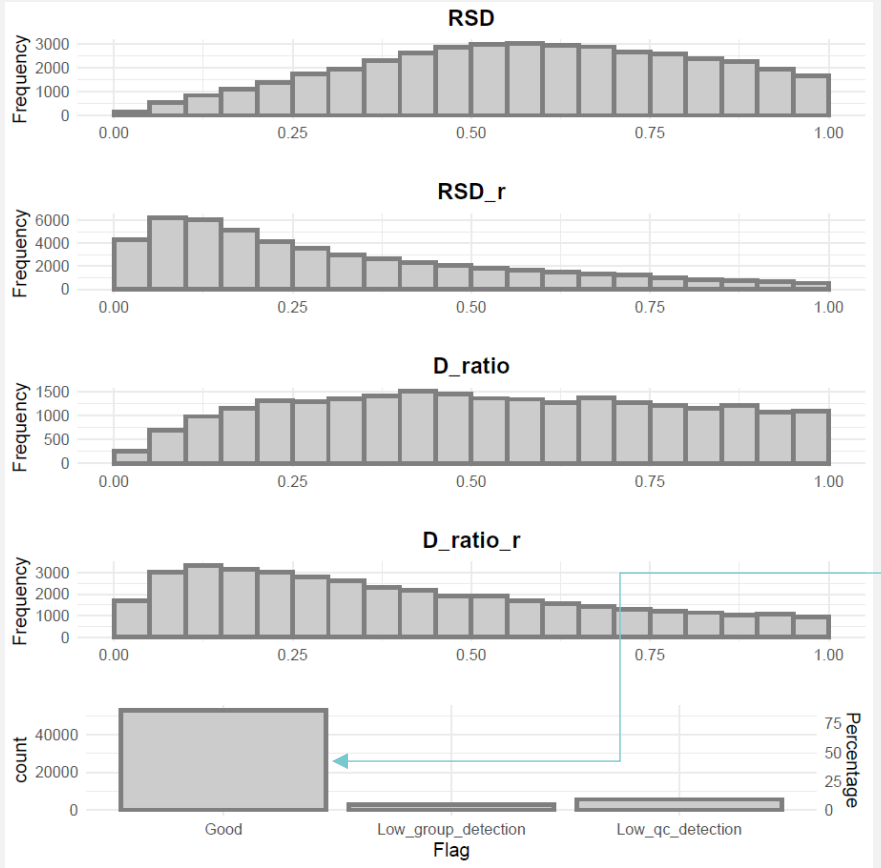
After



Quality metrics histograms

- Distribution of the quality metrics descriptives

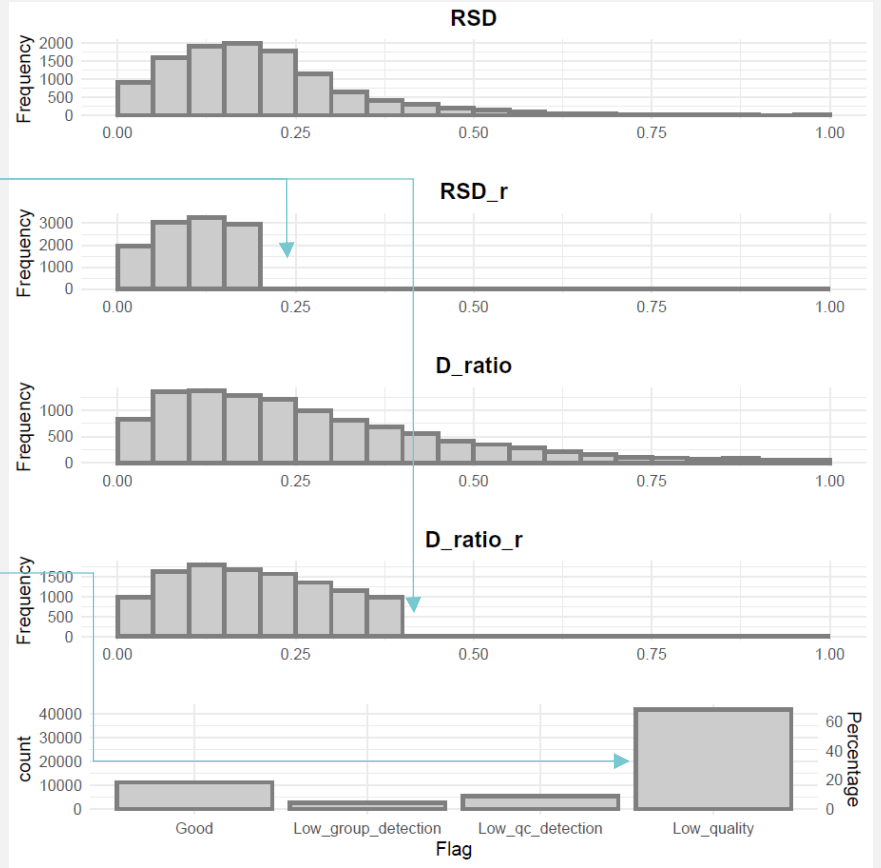
Before



This histogram only contains good-quality features

The good/low quality is determined during preprocessing

After



Tips for working with the preprocessed data in Excel

- The low detection rate & quality flags are meant for guidance and work best with plasma samples or within a single sample type
- We recommend keeping a ”master file” where all features are kept (also the flagged ones) – use filtering for hiding or sorting data
- Making additional columns in the Excel spreadsheet is useful for marking down manually curated identifications, ID level, chemical identifiers, comments, etc.

UNIVERSITY OF TURKU

Tips for working with the preprocessed data in Excel

Unique ID given by *notame*

This was given by the user earlier

Some column suggestions for the manual metabolite identification process

These are from MS-DIAL output

MS-DIAL output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
7																
8																
9																
10																
11																
12	Feature_ID	Split	Alignment_ID	Average_Rt_min	Average_Mz	Metabolite_name	Curated ID	HMDB subclass	HMDB ID	ID level	Kept for reporting	Tab	Comment	Adduct_type	Post_curatio	Fill__
13	RP_neg_120_0384a0_55	RP_neg	0	0.55	120.0384	w/o MS2:Anthranilate								[M-H]-	ion correlated	
14	RP_neg_120_05335a4_78	RP_neg	1	4.78	120.05335	w/o MS2:1H-Benzotriazole								[M-H]-	ion correlated	
15	RP_neg_121_02949a3_787	RP_neg	2	3.787	121.02949	Benzoic acid, 4-hydroxy-	4-Hydroxybenzaldehyde	Carbonyl compounds	HMDB0011718	1	x	1		[M-H]-		
16	RP_neg_122_97561a0_555	RP_neg	4	0.555	122.97561	Unknown								[M-H]-	similar chrom	
17	RP_neg_123_11755a2_506	RP_neg	5	2.506	123.11755	w/o MS2:trans,cis-3,6-Non								[M-H]-	ion correlated	
18	RP_neg_124_00718a0_517	RP_neg	6	0.517	124.00718	Taurine	Taurine	Organosulfonic acids ar	HMDB0000251	2			Use peak from HILIC neg	[M-H]-	ion correlated	
19	RP_neg_125_00101a0_636	RP_neg	7	0.636	125.00101	w/o MS2:Syringic acid								[M-H]-	similar chrom	
20	RP_neg_125_02428a0_55	RP_neg	8	0.55	125.02428	w/o MS2:4-Hydroxy-6-meth								[M-H2O-H]-	ion correlated	
21	RP_neg_125_62222a0_53	RP_neg	9	0.53	125.62222	Unknown								[M-H]-	ion correlated	
22	RP_neg_125_63573a2_339	RP_neg	10	2.339	125.63573	Unknown								[M-H]-		
23	RP_neg_125_87329a15_308	RP_neg	12	15.308	125.87329	Unknown								[M-H]-	ion correlated	
24	RP_neg_126_00314a0_512	RP_neg	13	0.512	126.00314	w/o MS2:2-Mercaptopyridi	Unknown C5H5NOS			4	x			[M-H]-		
25	RP_neg_126_94918a0_624	RP_neg	14	0.624	126.94918	Unknown								[M-H2O-H]-	ion correlated	
26	RP_neg_127_05117a0_514	RP_neg	15	0.514	127.05117	w/o MS2:6-Methyluracil								[M-H2O-H]-	ion correlated	
27	RP_neg_127_87012a15_316	RP_neg	16	15.316	127.87012	Unknown								[M-H]-	ion correlated	
28	RP_neg_127_94933a0_696	RP_neg	18	0.696	127.94933	Unknown								[M-H]-	ion correlated	
29	RP_neg_128_03519a1_062	RP_neg	19	1.062	128.03519	D-Pyroglutamic acid							In-source fragment	[M-H]-	found in highe	
30	RP_neg_128_03522a1_376	RP_neg	20	1.376	128.03522	D-Pyroglutamic acid								[M-H]-	found in highe	
31	RP_neg_128_03526a0_898	RP_neg	21	0.898	128.03526	D-Pyroglutamic acid								[M-H]-	adduct linked	
32	RP_neg_128_03539a0_523	RP_neg	22	0.523	128.03539	D-Pyroglutamic acid								[M-H]-	found in highe	

Data 1

Additional info of each identified metabolite (spectra, etc.) in its own tab

Kept for reporting = the molecular feature best representing each metabolite (≈ base peak)

UNIVERSITY OF TURKU

Tips for working with the preprocessed data in Excel

Quality flag (empty if good quality)

Detection rate (non-missing values) in each sample class; 1 means no missing values

Quality metrics

MS-DIAL output

Drift-corrected peak areas

	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE
7															Replicate	1	3	2	
8															Subject	OM_1	RV_3	RV_2	RO_2
9															QC	Sample	Sample	Sample	Sample
10															Injection_orde				
11															Batch_ID				
12	Spectrum_ref	MS1_isotopic_	MS_MS_spec	Flag	Detection_rate_S	Detection_rate_S	Detection_rate_S	Detection_rate_S	Detection_rate_S	Detection_rate_C	DC_note	RSD	RSD_r	D_ratio	D_ratio_r	220514_Koisti	220514_Koisti	220514_Koisti	220514_Ko
13	220514_Koisti	120.0384:951537	121.04175	Low_quality	1	1	0.777777778	0.444444444	1	0.8	Drift_corrected	0.236599095	0.269818877	0.070550246	0.15323188	126964.4442	29996.23923	47815.21157	78224.249
14	220514_Koisti	120.05335:1865236	121.056	Low_quality	1	1	1	1	1	1	Drift_corrected	0.306138069	0.191711387	1.045673533	0.759791676	1327109.468	946716.9261	1468789.846	1384543.1
15	220514_Koisti	121.02949:16	49.50377:0	49	Low_quality	1	1	1	1	1	1	Drift_corrected	0.425972673	0.554319455	1.202773276	1.378818329	230421.4879	108672.5998	477951.9949	223279.92
16	220514_Koisti	122.97561:613912	123.9789	Low_quality	1	1	1	1	1	1	Drift_corrected	0.35671491	0.374719735	0.884841892	0.753477737	92033.20503	187901.9137	125852.4069	118506.12
17	220514_Koisti	123.11755:610530	124.1209	Low_quality	0.888888889	0.888888889	0.777777778	0.666666667	0.777777778	1	Drift_corrected	0.35370655	0.283849553	0.138382184	1.663099604	96501.79221	23568.36172	1329.114331	272503.44
18	220514_Koisti	124.00718:16	49.50402:0	49	Low_quality	1	1	1	1	1	1	Drift_corrected	0.316477541	0.267442664	0.82567256	0.639439442	3299940.294	6551302.571	4443899.036	4109226.4
19	220514_Koisti	125.00101:30	49.50388:0	49	Low_quality	1	1	1	1	1	1	Drift_corrected	0.370274534	0.513345045	0.610981603	1.300421372	2561491.019	264625.5497	13346.40649	329096
20	220514_Koisti	125.02428:4734507	126.02763:0	127.0309	0.666666667	0.888888889	0.888888889	0.888888889	1	0.8	Drift_corrected	0.451382553	0.153732084	0.668899157	0.224797973	9644.875342		13762.35176	
21	220514_Koisti	125.62222:1484137	126.625	Low_qc_detec	0.111111111	0.222222222	0.222222222	0.111111111	0	0.2	Missing_QCS		0		0	835			
22	220514_Koisti	125.63573:500345	126.6390	Low_group_d	0.555555556	0.444444444	0.555555556	0.333333333	0.666666667	1	Drift_corrected	0.4228141	0.613308802	0.300370359	0.391415434	862.9759685	191638.6301		3361.053
23	220514_Koisti	125.87329:811000	126.8766	Low_quality	1	1	1	1	1	1	Drift_corrected	0.220271664	0.274639273	1.354757095	1.454202027	104067.4894	114581.3575	116924.2182	130504.97
24	220514_Koisti	126.00314:783340	127.0064	Low_quality	1	1	1	1	1	1	Drift_corrected	0.264130521	0.363845379	0.603688681	0.847845483	166662.3633	348670.152	239500.5911	152612.21
25	220514_Koisti	126.94918:57	49.50373:0	49	Low_quality	1	1	0.666666667	1	1	1	Drift_corrected	0.505275011	0.502849766	0.459363918	0.678568923	36390.19773	15338.45759	12398.06054	33371.131
26	220514_Koisti	127.05138:70	49.50385:0	49	Low_quality	1	1	1	0.888888889	1	1	Drift_corrected	0.331474042	0.40601163	0.211965666	1.572306369	107230.5431	12138.4626	3856.100064	71745.533
27	220514_Koisti	127.87012:518214	128.8734	Low_quality	1	1	1	1	1	1	Drift_corrected	0.274032291	0.4164902	1.080055962	1.340909489	34516.61225	66050.00177	57129.67265	72818.757
28	220514_Koisti	127.94933:677572	128.9526	Low_qc_detec	0	0.222222222	0.111111111	0.222222222	0.222222222	0	Missing_QCS					7412		2368	
29	220514_Koisti	128.03523:18	49.50373:0	49	Low_quality	1	1	1	1	1	1	Drift_corrected	0.296816802	0.251528801	0.829626795	0.644749417	116127414.2	158750707	85526007.64	13064096
30	220514_Koisti	128.03522:21	49.5038:0	49.5	Low_quality	1	1	1	1	1	1	Drift_corrected	0.733055274	1.035753784	1.571691694	1.558185673	632453.417	1021320.231	722108.2613	1462496.4
31	220514_Koisti	128.03546:19	49.50381:0	49	Low_quality	1	1	1	1	1	1	Drift_corrected	0.309120968	0.244197489	1.049487993	0.712782875	175653602.6	177365173.3	117414206	18906376
32	220514_Koisti	128.03539:37	49.50446:0	49	Low_quality	1	1	1	1	1	1	Drift_corrected	0.343534721	0.291111762	0.59705737	0.64970071	3858251.958	640614.6159	669185.4815	5427491.2

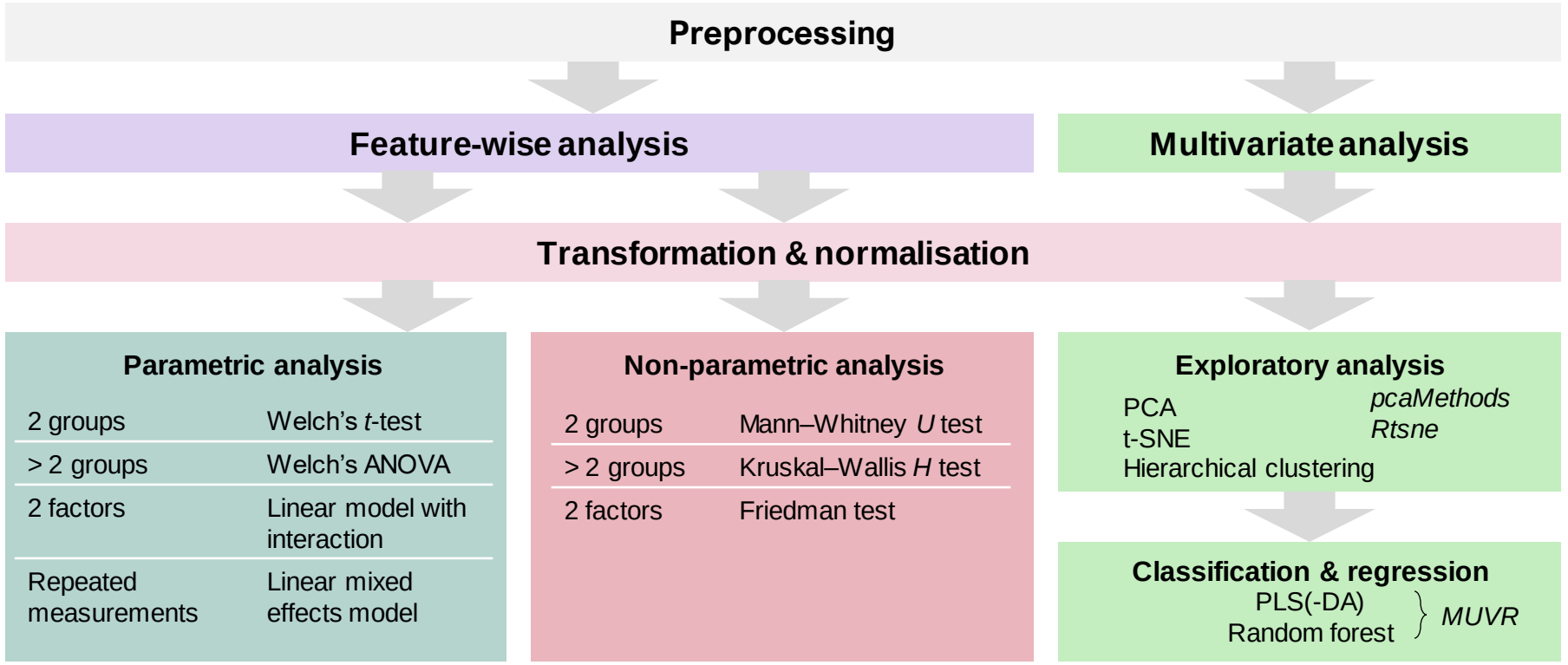
Data 1

UNIVERSITY OF TURKU

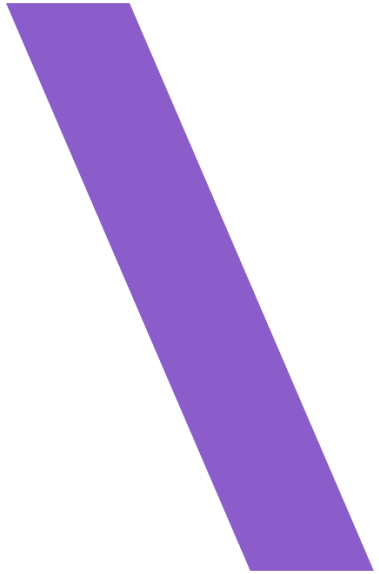
Detected or not detected?

- The definition of what is detected or not is not simple in untargeted metabolomics, where the LoD is not determined for each metabolite
- A signal-to-noise ratio > 5 is a better measure of detection than the presence or absence of a signal
- Run solvent blanks to distinguish actual metabolites from background ions

Statistics in notame



Additional flagging and demonstration in RStudio



Feature data has a lot of information already

MS_MS_spectrum	Flag	Detection_rate_Tissue_Heart	Detection_rate_Tissue_Liver	Detection_rate_QC	DC_note	RSD	RSD_r	D_ratio	D_ratio_r
NA	Low_group_detection	0.2222222	0.1111111	0.75	Missing_QCS	1.66432128	0.741300000	1.391449605	0.45120798
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
28.769:53 44.933:42 52.106:89 69.518:140	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
72.938:159	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_qc_detection	0.2222222	0.1111111	0.25	Missing_QCS	NA	0.000000000	NA	0.00000000
NA	Low_quality	1.0000000	1.0000000	1.00	Drift_corrected	0.07742367	0.073188463	0.522562935	0.50289175
NA	Low_quality	1.0000000	1.0000000	1.00	Drift_corrected	0.07239618	0.079319374	0.774161914	0.91414465
NA	Low_quality	1.0000000	1.0000000	1.00	Drift_corrected	0.11338674	0.056372926	0.402303322	0.43910808
NA	Low_quality	1.0000000	1.0000000	1.00	Drift_corrected	0.17490598	0.205731801	0.235147161	0.65299903

Additional flagging

- Flag contaminants
  - Based on signal to noise ratio
- Flag based on mean abundance
- Add flag based on minimum abundance
- Detect and flag absence of msms information
- Create “goodness” calculators

Flag contaminants

- Default for flag_thresh = 0.05
- Signal-to-noise > 5
  - flag_thresh = 0.2

```
flag_contaminants(
  object,
  blank_col,
  blank_label,
  flag_thresh = 0.05,
  flag_label = "Contaminant"
)
```

Flag (low) abundance

- Ready-made functions to manipulate data
- exprs(metaboset) fetch the abundance table
- flag(metaboset) access or set the current stage of flagging

```
211  # threshold for mean expression
212  threshhold <- 5000
213
214  means <- apply(exprs(imputed), 1, finite_mean)
215  condition <- (means < threshhold) & is.na(flag(imputed))
216  summary(condition)
217  flag(imputed)[condition] <- "Below 5K abundance"
218
```

Flag based on minimum abundance

- Same methods used
  - flag()
  - exprs()
- Freely set the name of the condition

```
219  # threshold for smallest value
220  threshhold <- 500
221  means <- apply(exprs(imputed), 1, min)
222  condition <- (means < threshhold) & is.na(flag(imputed))
223  summary(condition)
224  flag(imputed)[condition] <- "Missing values"
```

Flag absence of ms(2)

- Specific columns can be selected to be the definition of flagging
- In this case "MS_MS_spectrum"

```
229 #10 No ms2
230 # Check the name of msms column
231 colnames(fData(imputed))
232
233 condition <- is.na(fData(imputed)$"MS_MS_spectrum") &
234   is.na(flag(imputed))
235 summary(condition)
236 flag(imputed)[condition] <- "No MSMS"
```

“Goodness” calculators

- Especially useful for the untargeted method and the list of unknown features
- Can be used with any condition
  - Fold change
  - Anova
  - Paired comparisons

```
273   # count if fold change or cohen_D
274   if (sel_test %in% c("Cohen_d")){
275     thresh2 <- -thresh
276     temp <- select(fData(object), contains(sel_test))
277     countt <- apply(temp, 1, function(x) sum(x>thresh|x<thresh2, na.rm = T))
278   }
279   else if (sel_test == c("FC")){
280     thresh2 <- 1/thresh
281     temp <- select(fData(object), contains(sel_test))
282     countt <- apply(temp, 1, function(x) sum(x>thresh|x<thresh2, na.rm = T))
283   }
284
```

Now, let's go to R...

UNIVERSITY OF TURKU

How to view a feature data?

- View(fData(imputed))

	Feature_ID	Split	Alignment_ID	Average_Rt_min_	Average_Mz	Metabolite_name	Adduct_ion_name	Post_curation_result
HILIC_neg_61_9906a0_95	HILIC_neg_61_9906a0_95	HILIC_neg	2	0.950	61.9906	Unknown	[M-H]-	NA
HILIC_neg_68_9956a2_89	HILIC_neg_68_9956a2_89	HILIC_neg	3	2.890	68.9956	Unknown	[M-H]-	NA
HILIC_neg_73_9719a0_51	HILIC_neg_73_9719a0_51	HILIC_neg	6	0.510	73.9719	Unknown	[M-H]-	NA
HILIC_pos_84_0445a6_13	HILIC_pos_84_0445a6_13	HILIC_pos	14	6.130	84.0445	Unknown	[M+H]+	NA
HILIC_neg_84_0104a4_18	HILIC_neg_84_0104a4_18	HILIC_neg	11	4.180	84.0104	Unknown	[M-H]-	Highly correlated with 348(0.96)
HILIC_neg_89_0243a1_33	HILIC_neg_89_0243a1_33	HILIC_neg	15	1.330	89.0243	Unknown	[M-H]-	NA
HILIC_pos_60_0444a6_17	HILIC_pos_60_0444a6_17	HILIC_pos	3	6.170	60.0444	w/o MS2:N-Methylformamide	[M+H]+	NA
HILIC_pos_86_0963a6_3	HILIC_pos_86_0963a6_3	HILIC_pos	17	6.300	86.0963	w/o MS2:Piperidine	[M+H]+	NA
HILIC_pos_86_0964a3_88	HILIC_pos_86_0964a3_88	HILIC_pos	18	3.880	86.0964	Unknown	[M+H]+	NA
HILIC_pos_89_1074a6_52	HILIC_pos_89_1074a6_52	HILIC_pos	20	6.520	89.1074	w/o MS2:1,2-Diethylhydrazine	[M+H]+	NA
RP_neg_78_9588a1_11	RP_neg_78_9588a1_11	RP_neg	6	1.110	78.9588	Unknown	[M-H2O-H]-	May be [M-H2O-H]- of Alignment ID: 10;
RP_neg_96_9675a10_88	RP_neg_96_9675a10_88	RP_neg	9	10.880	96.9675	w/o MS2:Sulfuric acid	[M-H]-	NA

						A	A	B	B	B
File type	QC	Sample	Sample	Sample	Sample	Sample	Sample	Sample
Injection order	1	2	3	4	5	6	7
Batch ID	1	1	1	1	1	1	1
Alignment ID	Average Rt(min)	Average Mz	Metabolite name	Split	File_1	File_2	File_3	File_4	File_5	File_6	File_7
1	14.05	100.0793	Metabolite 1	RP pos	15389	16004	67816	37654	2976	2867	1085
2	8.78	112.1583	Metabolite 2	RP pos	1654516	6549684	4352415	1986545	6854646	4632184	7874546
3	2.16	132.0428	Metabolite 3	RP pos	4594146	8242334	3261199	1728190	2475302	3164750	9225945
4	3.53	140.0467	Metabolite 4	RP pos	4821951	113017	690884	2961649	6238804	3533544	1099746
5	8.35	153.9692	Metabolite 5	RP pos	8701229	9870760	91094	4073828	6291024	5607944	710761
6	10.37	165.0105	Metabolite 6	RP pos	8029117	6695877	910399	3880946		1644123	5253212
7	9.01	165.0351	Metabolite 7	RP pos	419517	120	7489125	792169	5787382	952883	8570666
8	4.77	177.0273	Metabolite 8	RP pos	3664172	9280985	8763212	2334167	6399593	7653679	9248892

UNIVERSITY OF TURKU

How to manipulate feature data

- It can be extracted
  - extracted_fdata<- fData(object)
- fData manipulation is possible similarly than any data frame
  - fData(object)$good_flag <- 0
- Build in functions will do the job in most cases
  - join_fdata(Metaboset, new_information)

Wheat_vs_Rye_FC	Rye_vs_Wheat_t_test_P	1_vs_2_t_test_q	Wheat_vs_Rye_Cohen_d
0.52100259	1.340075e-01	NA	-0.65733789
0.61733333	6.194331e-03	NA	-1.36651127
1.08075121	6.887459e-01	NA	0.16571948
1.11928154	4.188084e-01	NA	0.33675272
0.97717978	7.283208e-01	NA	-0.14369190
1.14187697	2.586829e-01	NA	0.47523618
1.13529758	4.976321e-01	NA	0.28162054
0.77600707	5.242789e-01	NA	-0.26805607
1.51240165	4.207693e-02	NA	0.88370016
0.98388013	7.052554e-01	NA	-0.15675439
0.76922886	1.219481e-01	NA	-0.65954857
1.18542902	3.658346e-01	NA	0.37710843
1.37708841	2.255556e-01	NA	0.51739252

UNIVERSITY OF TURKU

How to get the abundance table?

- View(exprs(Metaboset))
- abundance_table <- exprs(Metaboset)
- Can be used for example as an input for machine learning tools

				Class		A	A	A	B	B	B
				File type	QC	Sample	Sample	Sample	Sample	Sample	Sample
				Injection order	1	2	3	4	5	6	7
				Batch ID	1	1	1	1	1	1	1
Alignment ID	Average Rt(min)	Average Mz	Metabolite name	Split	File_1	File_2	File_3	File_4	File_5	File_6	File_7
1	14.05	100.0793	Metabolite 1	RP pos	15389	16004	67816	37654	2976	2867	1085
2	8.78	112.1583	Metabolite 2	RP pos	1654516	6549684	4352415	1986545	6854646	4632184	7874546
3	2.16	132.0428	Metabolite 3	RP pos	4594146	8242334	3261199	1728190	2475302	3164750	9225945
4	3.53	140.0467	Metabolite 4	RP pos	4821951	113017	690884	2961649	6238804	3533544	1099746
5	8.35	153.9692	Metabolite 5	RP pos	8701229	9870760	91094	4073828	6291024	5607944	710761
6	10.37	165.0105	Metabolite 6	RP pos	8029117	6695877	910399	3880946		1644123	5253212
7	9.01	165.0351	Metabolite 7	RP pos	419517	120	7489125	792169	5787382	952883	8570666
8	4.77	177.0273	Metabolite 8	RP pos	3664172	9280985	8763212	2334167	6399593	7653679	9248892

UNIVERSITY OF TURKU

Pheno data

- pData(Metaboset)
- Other measurements can be added
  - For example, Clinical measurements or other omics data
- join_pdata(Metaboset, new information)

				Class		A	A	A	B	B	B
				File type	QC	Sample	Sample	Sample	Sample	Sample	Sample
				Injection order	1	2	3	4	5	6	7
				Batch ID	1	1	1	1	1	1	1
Alignment ID	Average Rt(min)	Average Mz	Metabolite name	Split	File_1	File_2	File_3	File_4	File_5	File_6	File_7
1	14.05	100.0793	Metabolite 1	RP pos	15389	16004	67816	37654	2976	2867	1085
2	8.78	112.1583	Metabolite 2	RP pos	1654516	6549684	4352415	1986545	6854646	4632184	7874546
3	2.16	132.0428	Metabolite 3	RP pos	4594146	8242334	3261199	1728190	2475302	3164750	9225945
4	3.53	140.0467	Metabolite 4	RP pos	4821951	113017	690884	2961649	6238804	3533544	1099746
5	8.35	153.9692	Metabolite 5	RP pos	8701229	9870760	91094	4073828	6291024	5607944	710761
6	10.37	165.0105	Metabolite 6	RP pos	8029117	6695877	910399	3880946		1644123	5253212
7	9.01	165.0351	Metabolite 7	RP pos	419517	120	7489125	792169	5787382	952883	8570666
8	4.77	177.0273	Metabolite 8	RP pos	3664172	9280985	8763212	2334167	6399593	7653679	9248892

UNIVERSITY OF TURKU

Pheno data

- pData(Metaboset)
- Other measurements can be added
  - For example, Clinical measurements or other omics data
- join_pdata(Metaboset, new information)

Sample_ID	Class	Tissue	Diet	QC	Injection_order
Sample_2	Heart_Control	Heart	Control	Sample	2
Sample_3	Liver_Control	Liver	Control	Sample	3
Sample_4	Heart_Wheat	Heart	Wheat	Sample	4
Sample_5	Liver_Wheat	Liver	Wheat	Sample	5
Sample_6	Heart_Rye	Heart	Rye	Sample	6
Sample_7	Liver_Rye	Liver	Rye	Sample	7
Sample_8	Heart_Control	Heart	Control	Sample	8
Sample_9	Liver_Control	Liver	Control	Sample	9
Sample_10	Heart_Wheat	Heart	Wheat	Sample	10
Sample_11	Liver_Wheat	Liver	Wheat	Sample	11
Sample_12	Heart_Rye	Heart	Rye	Sample	12
Sample_13	Liver_Rye	Liver	Rye	Sample	13
Sample_15	Heart_Control	Heart	Control	Sample	15
Sample_16	Liver_Control	Liver	Control	Sample	16

		Class		A	A	A	B	B	B
		File type	QC	Sample	Sample	Sample	Sample	Sample	Sample
		Injection order	1	2	3	4	5	6	7
		Batch ID	1	1	1	1	1	1	1
Alignment ID	Average Rt(min)	Average Mz	Metabolite name	Split	File_1	File_2	File_3	File_4	File_5	File_6	File_7
1	14.05	100.0793	Metabolite 1	RP pos	15389	16004	67816	37654	2976	2867	1085
2	8.78	112.1583	Metabolite 2	RP pos	1654516	6549684	4352415	1986545	6854646	4632184	7874546
3	2.16	132.0428	Metabolite 3	RP pos	4594146	8242334	3261199	1728190	2475302	3164750	9225945
4	3.53	140.0467	Metabolite 4	RP pos	4821951	113017	690884	2961649	6238804	3533544	1099746
5	8.35	153.9692	Metabolite 5	RP pos	8701229	9870760	91094	4073828	6291024	5607944	710761
6	10.37	165.0105	Metabolite 6	RP pos	8029117	6695877	910399	3880946		1644123	5253212
7	9.01	165.0351	Metabolite 7	RP pos	419517	120	7489125	792169	5787382	952883	8570666
8	4.77	177.0273	Metabolite 8	RP pos	3664172	9280985	8763212	2334167	6399593	7653679	9248892

UNIVERSITY OF TURKU

Pheno data manipulation in a separate data frame

- dframe <- pData(Metaboset)
- Do your things for example remove rows that aren't needed
  - dframe <- dframe[-1,]
- Set it back to Metaboset
  - pData(Metaboset) <- dframe

Manipulating data and some statistics

- Select data based on groups similarly as in any data frame in R
- Basic statistics tools are included in Notame package
- For example, fold change can be calculated easily
  - sel_sample<- imputed$Diet %in% c("Rye" , "Wheat")
  - temp <- imputed[,sel_sample]
  - f_change <- fold_change(temp, group = "Diet")

Sample_ID	Class	Tissue	Diet	QC	Injection_order
Sample_2	Heart_Control	Heart	Control	Sample	2
Sample_3	Liver_Control	Liver	Control	Sample	3
Sample_4	Heart_Wheat	Heart	Wheat	Sample	4
Sample_5	Liver_Wheat	Liver	Wheat	Sample	5
Sample_6	Heart_Rye	Heart	Rye	Sample	6
Sample_7	Liver_Rye	Liver	Rye	Sample	7
Sample_8	Heart_Control	Heart	Control	Sample	8
Sample_9	Liver_Control	Liver	Control	Sample	9
Sample_10	Heart_Wheat	Heart	Wheat	Sample	10
Sample_11	Liver_Wheat	Liver	Wheat	Sample	11
Sample_12	Heart_Rye	Heart	Rye	Sample	12
Sample_13	Liver_Rye	Liver	Rye	Sample	13
Sample_15	Heart_Control	Heart	Control	Sample	15
Sample_16	Liver_Control	Liver	Control	Sample	16

Take home message

- Garbage in, garbage out—data preprocessing matters
- Some remaining pitfalls include a lack of quality assessment and challenges in imputation

Workshop material GitHub link

Publication

Notame GitHub link

UNIVERSITY OF TURKU

UNIVERSITY
OF TURKU