MetaboPrep Data Preparation Summary Report

05 June, 2025

Contents

1 Project Information

This `metaboprep` report summarizes the data preparation steps for:

- **Project:** Project

1.1 Overview

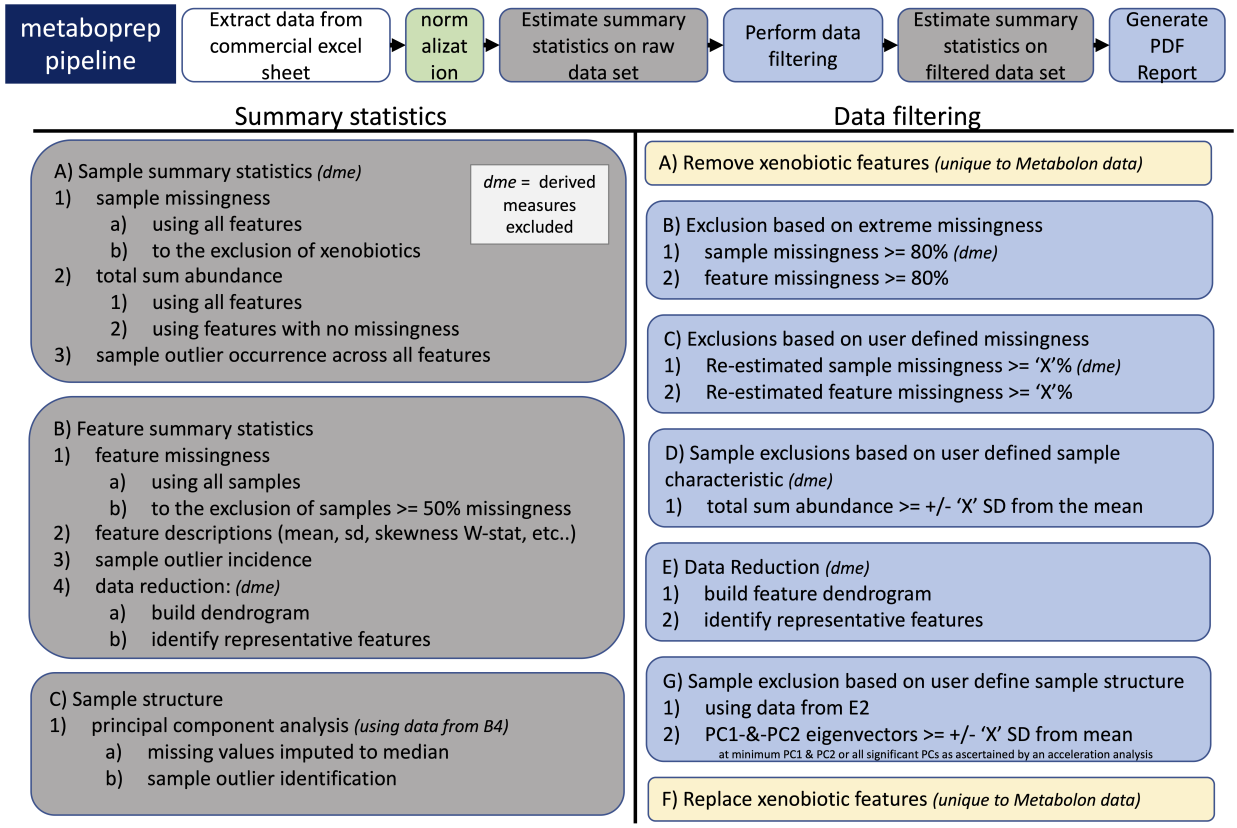
The `metaboprep` R package performs three key operations:

1. **Assessment & Summary Statistics**: Provides an initial assessment of raw metabolomics data.
2. **Data Filtering**: Applies filtering techniques to clean the dataset.
3. **Post-Filtering Assessment**: Evaluates the filtered dataset, particularly in relation to batch variables when available.

This report contains descriptive information on both raw and filtered metabolomics data for **Project**.

Please raise any issues on the GitHub issues page.

1.2 Data preparation workflow:



2 Summary of raw data

2.1 Sample size of Project data set:

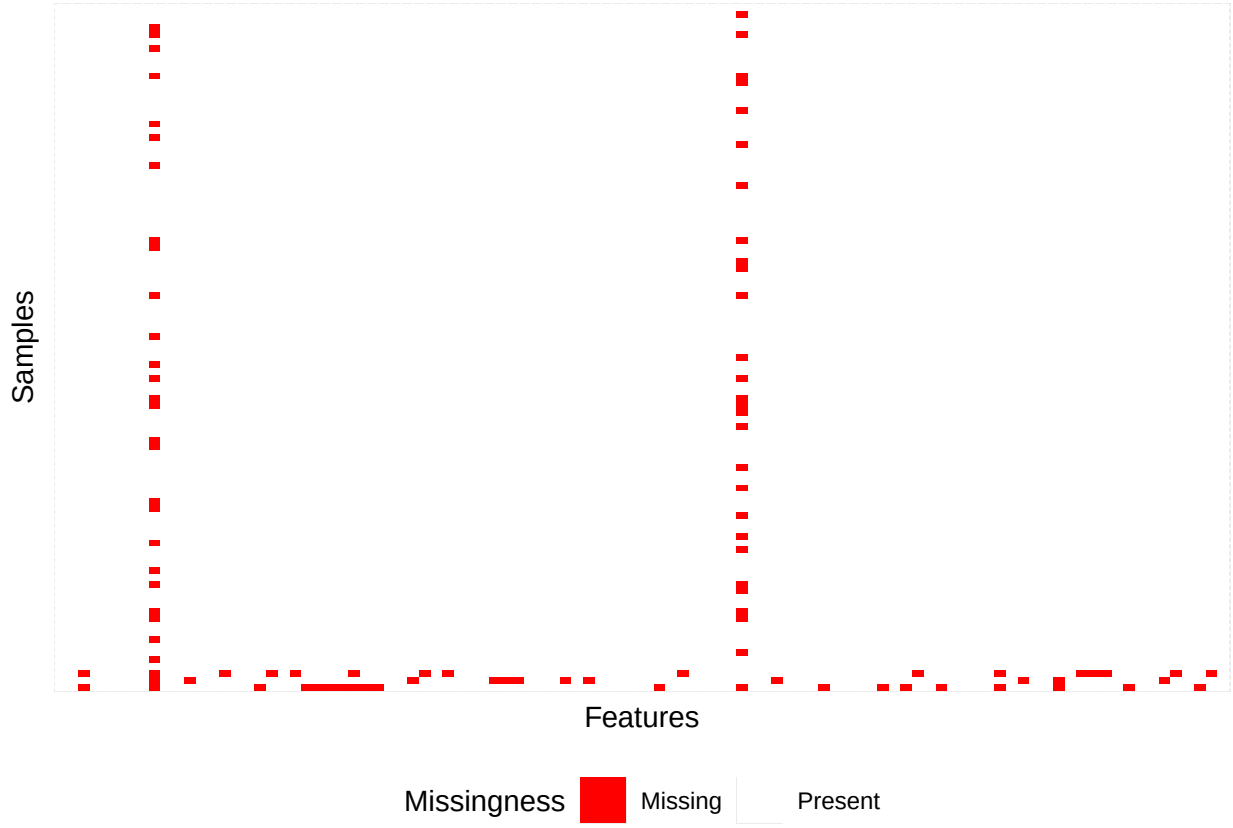
Table 1: Sample Size Summary

Dataset	Samples	Features
INPUT	100	100
QC	100	100

3 Missingness

Missingness is evaluated across samples and features using the original/raw data set.

3.1 Visual structure of missingness in the raw data



3.2 Summary of sample and feature missingness

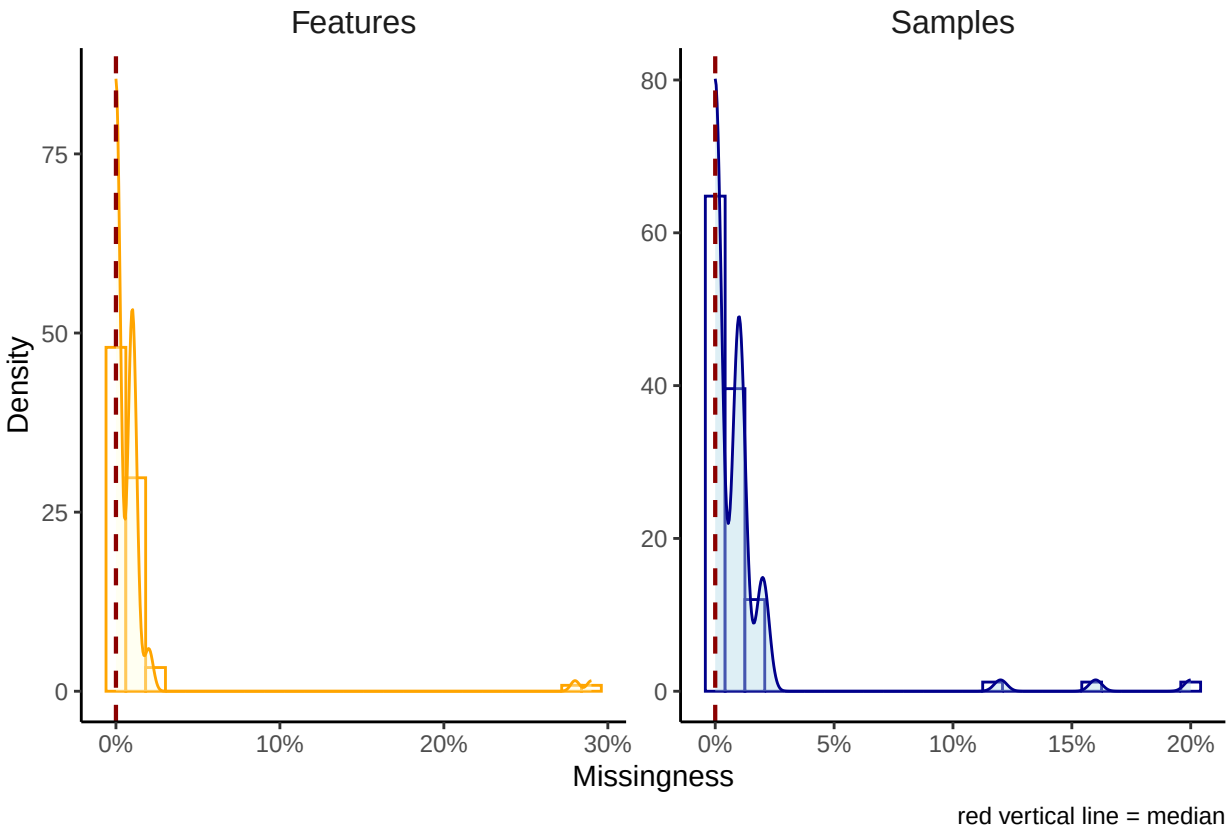


Table 2: Sample and feature missingness percentiles.

	Percentile	Features	Samples
0%	0.00	0.00	0.00
25%	0.25	0.00	0.00
50%	0.50	0.00	0.00
75%	0.75	0.01	0.01
100%	1.00	0.29	0.20

Table 3: Estimates of samples sizes under various levels of feature missingness.

Missingness....	Sample.Size
10%	90
20%	80
30%	70
40%	60
50%	50

4 Data Filtering

4.1 Exclusion summary

Table 4: Sample and Feature Exclusions

Reason	Count
Features	
- user excluded	0
- extreme feature missingness	0
- user defined feature missingness	*0*
Samples	
- user excluded	0
- extreme sample missingness	0
- user defined sample missingness	0
- user defined sample totalpeakarea	0
- user defined sample pca outlier	0

Note:
Six primary data filtering exclusion steps were made during the preparation of the data. User-defined thresholds indicated with an asterisk. In addition, the number of outlier datapoints modified during QC are presented.

[1] Samples with missingness >=80
[2] Sample exclusions based on the user defined threshold of *>=80
[3] Features with missingness >=80
[4] Feature exclusions based on user defined threshold of *>=80
[5] Samples with a total-peak-area or total-sum-abundance that is *>=5 SD from the mean.
[6] Samples that are *>=5 SD from the mean on principal component axis 1 and 2

4.2 Metabolite or feature reduction and principal components

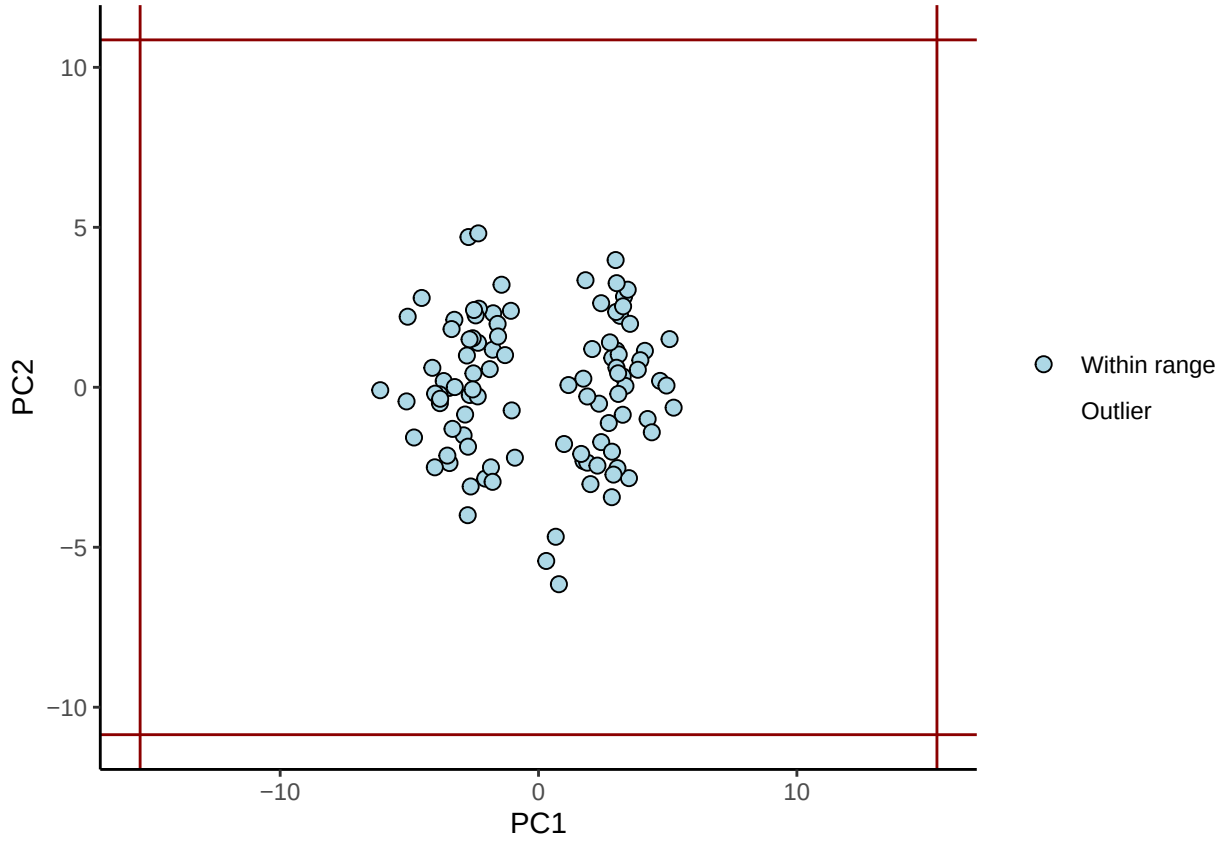
A data reduction was carried out to identify a list of representative features for generating a sample principal component analysis. This step reduces the level of inter-correlation in the data to ensure that the principal components are not driven by groups of correlated features.

The data reduction table presents the number of metabolites at each phase of the data reduction (Spearman's correlation distance tree cutting) analysis.

Table 5: Feature summary

Data reduction	Count
Total metabolite count	100
Metabolites included in data reduction	98
Number of metabolite clusters	95
Number of representative metabolites	95

The following plot respresents principal components 1 and 2 using 98 representative metabolites. The red vertical and horizontal lines indicate the standard deviation cutoffs for identifying individual outliers. Outliers are those >=5 SD from the mean of PCs 1-2.



5 Summary of filtered data

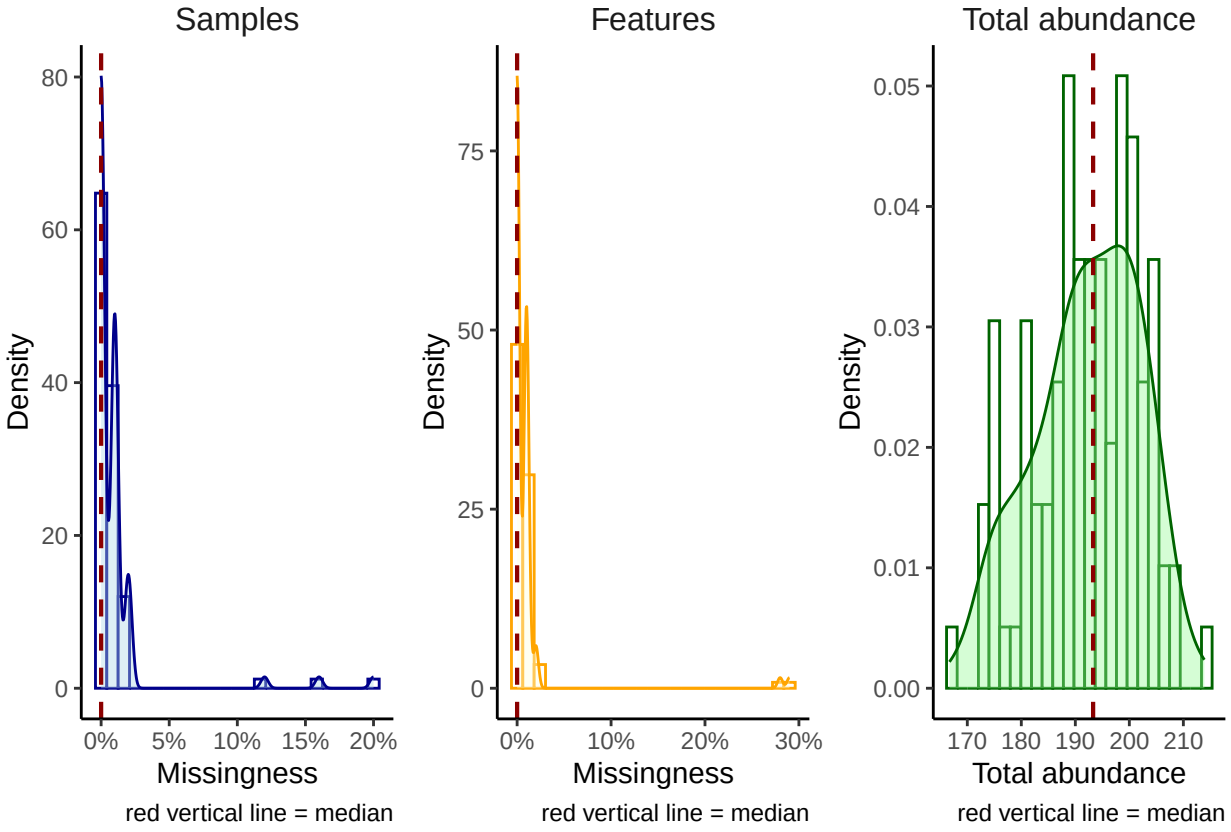
5.1 Sample size (N)

- The number of samples in data = 100
- The number of features in data = 100

5.2 Relative to the raw data

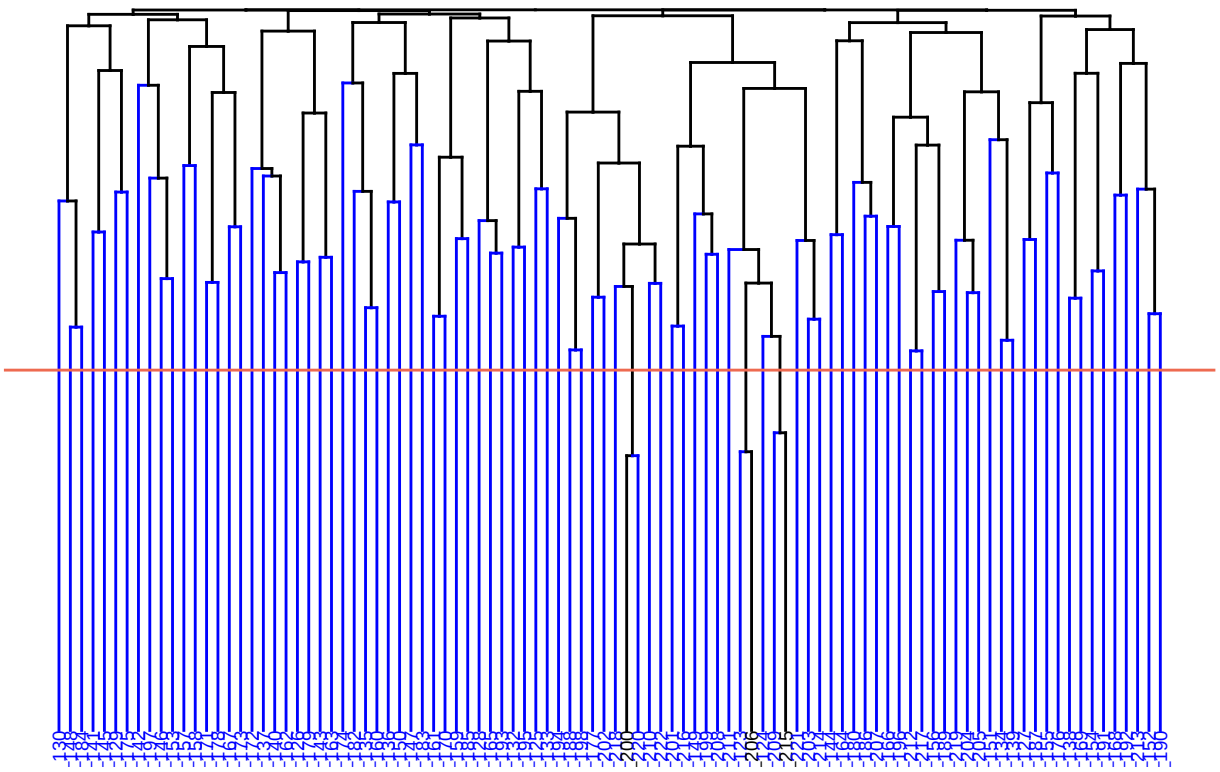
- 0 samples were filtered out, given the user's criteria.
- 0 features were filtered out, given the user's criteria (also excluding xenobiotics)
- Please review details above and your log file for the number of features and samples excluded and why.

5.2.1 Distributions for sample and feature missingness



5.2.2 Clustering dendrogram of representative features

Spearman's correlation distance clustering dendrogram highlighting the metabolites used as representative features in blue, the clustering tree cut height is denoted by the horizontal line.



5.2.3 Summary of the QC (filtered) metabolite data

The data reduction table presents the number of metabolites at each phase of the data reduction (Spearman's correlation distance tree cutting) analysis.

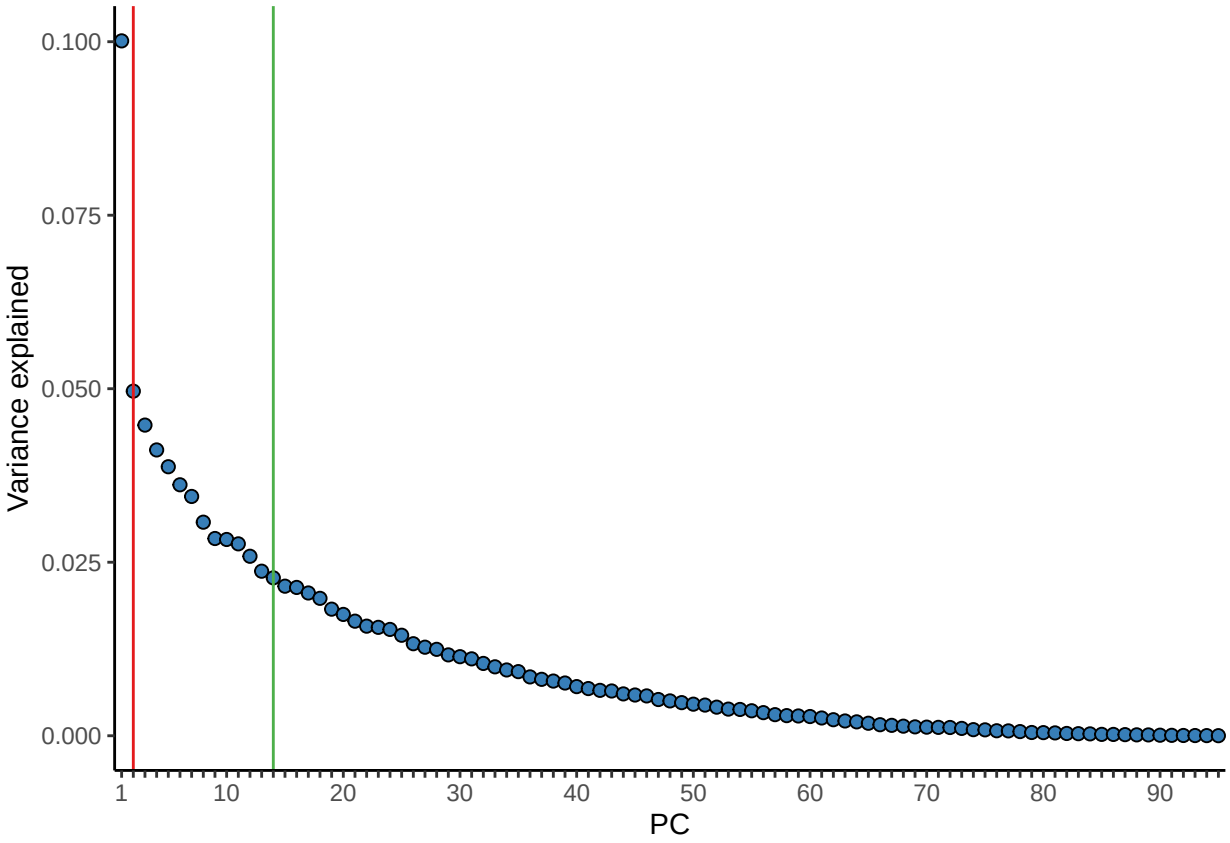
Table 6: QC Feature Summary

	Count
Total metabolite count (incl. xenbiotics)	100
Metabolites included in data reduction	98
Number of metabolite clusters	95
Number of representative metabolites	95

5.2.4 Scree plot

Scree plot of the variance explained by each PC (limited to 100 for plotting) and a plot of principal component 1 and 2, as derived from the representative metabolites. The Scree plot also identifies the number of PCs

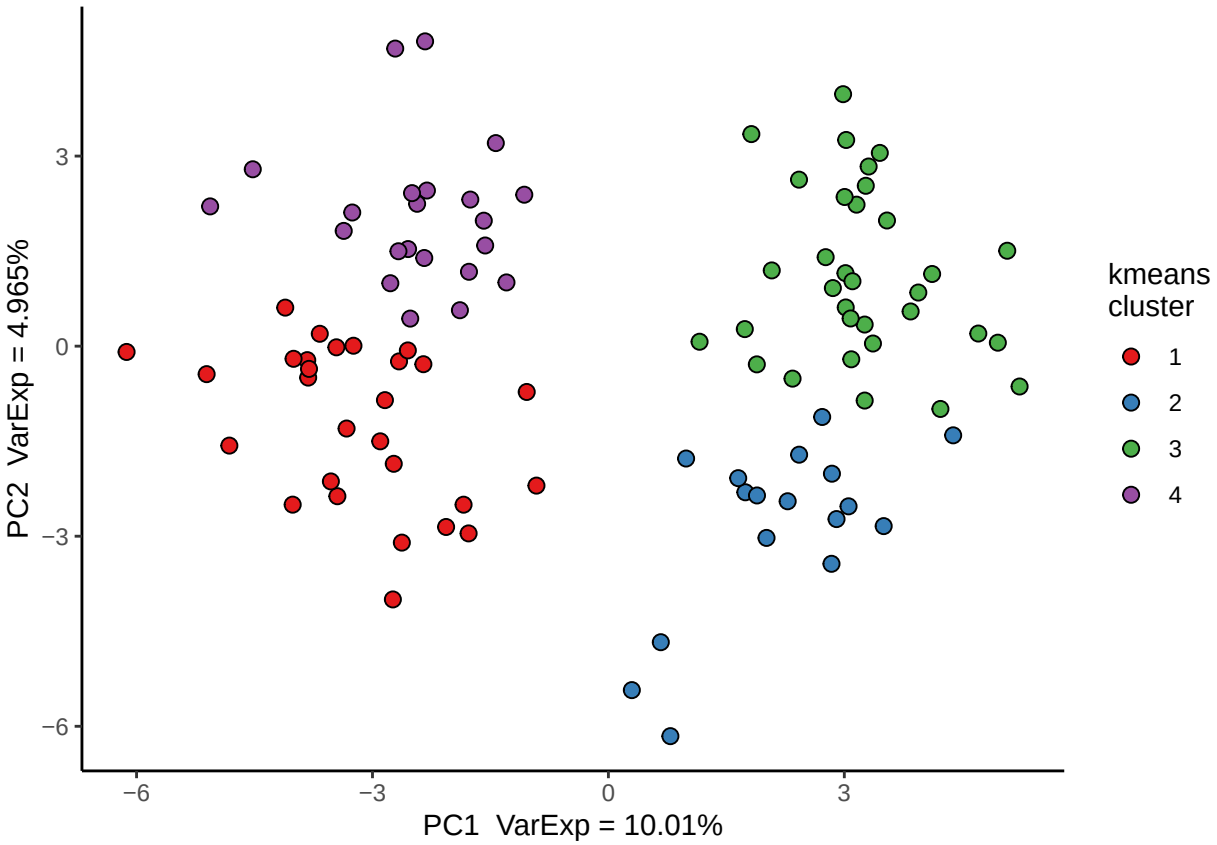
estimated to be informative (vertical lines) by the Cattel's Scree Test acceleration factor (red, n = 2) and Parallel Analysis (green, n = 14).



5.2.5 PC plot

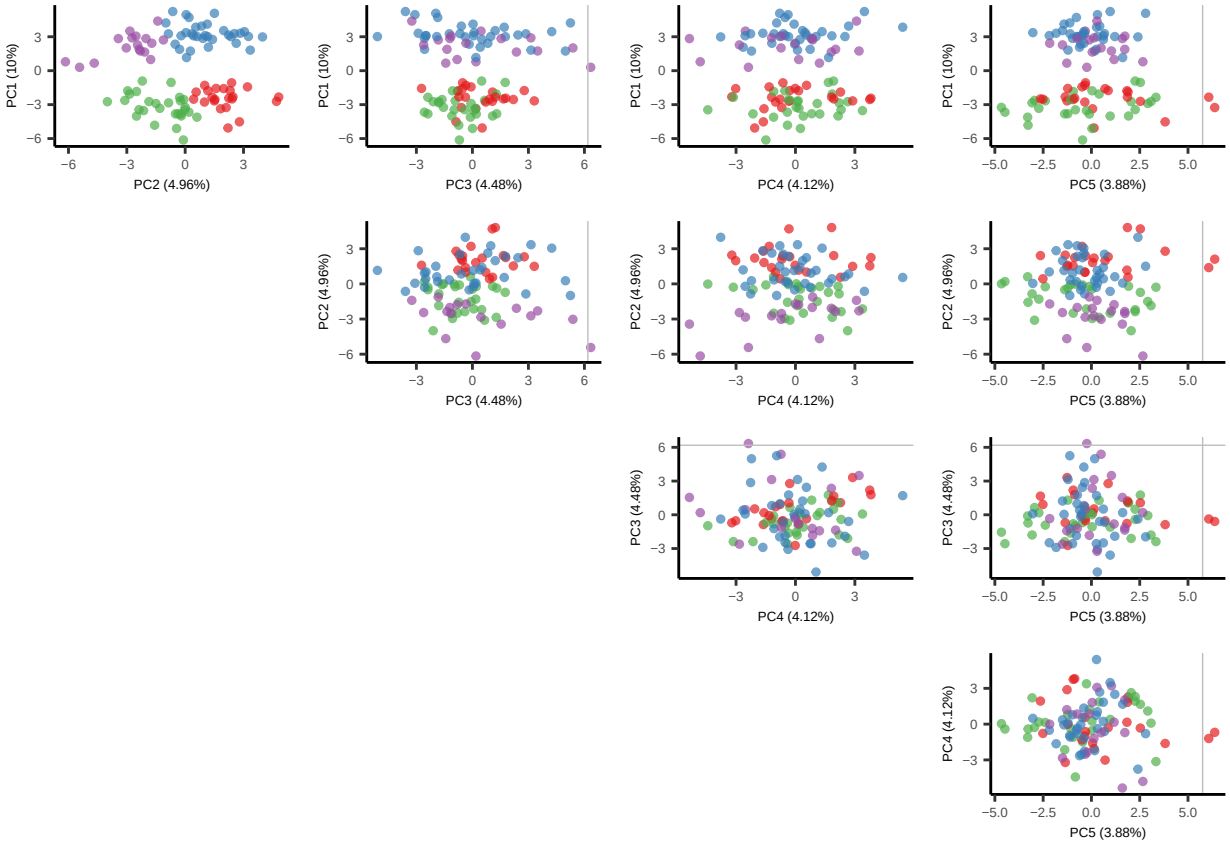
Individuals in the PC plot were clustered into 4 kmeans (k) clusters, using data from PC1 and PC2. The kmeans clustering and color coding is strictly there to help provide some visualization of the major axes of variation in the sample population(s).

The plot presents principal components 1 & 2 using 95 representative metabolites.


PC2 VarExp = 4.965%
PC1 VarExp = 10.01%
3
0
−3
−6
−6
−3
0
3
kmeans cluster
1
2
3
4

5.3 Structure among samples

A matrix (pairs) plot of the top five principal components including demarcations of the 3rd (grey), 4th (orange), and 5th (red) standard deviations from the mean. Samples are color coded as in the summary PC plot above using a kmeans analysis of PC1 and PC2 with a k (number of clusters) set at 4. The choice of k = 4 was not robustly chosen it was a choice of simplicity to help aid visualize variation and sample mobility across the PCs.



5.4 Feature Distributions

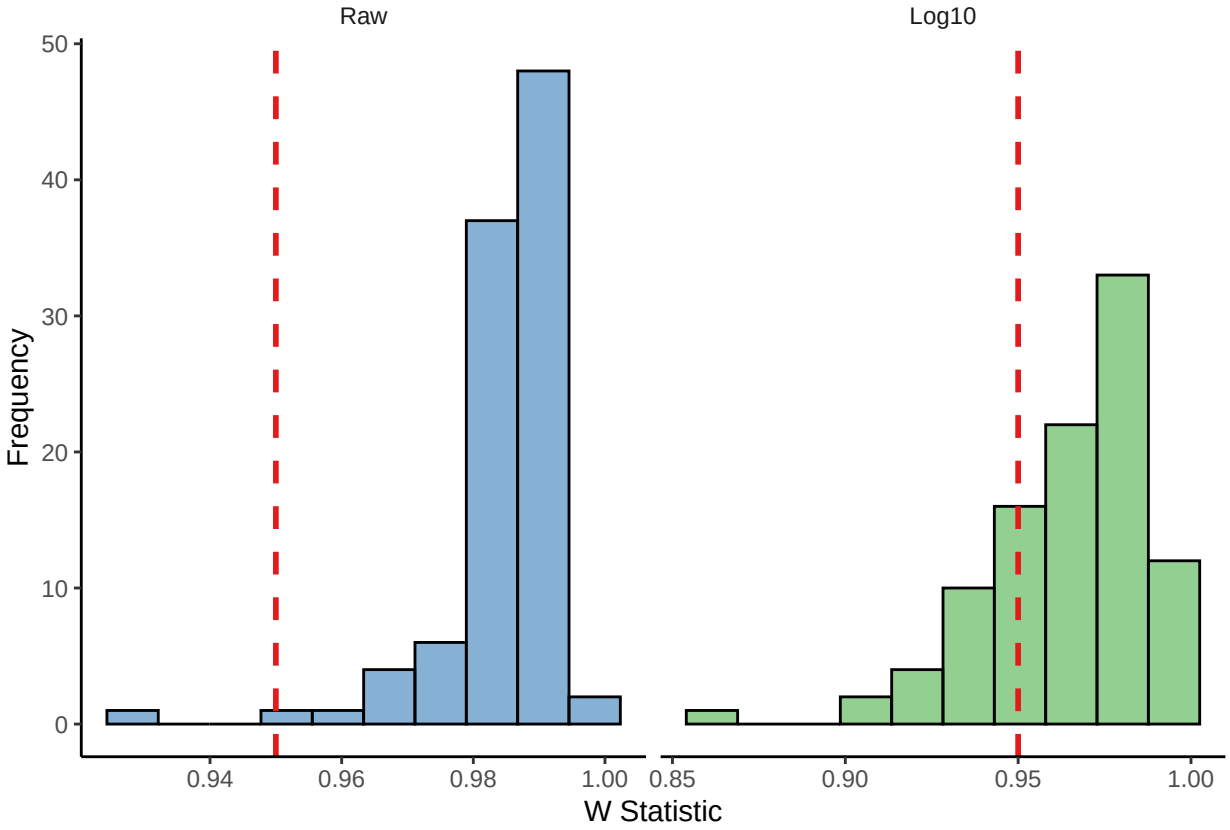
5.4.1 Estimates of normality: W-statistics for raw and log transformed data

Of the 100 features in the data 0 features were excluded from this analysis because of no variation or too few observations (n < 40). Of the remaining 100 metabolite features, a total of 98 may be considered normally distributed given a Shapiro W-statistic >= 0.95.

5.4.2 Distribution of W Statistics on Raw and Log10 Metabolite Abundances

Histogram plots of Shapiro W-statistics for raw and log transformed data distributions. A W-statistic value of 1 indicates the sample distribution is perfectly normal and value of 0 indicates it is perfectly uniform. Please note that log transformation of the data *may not* improve the normality of your data.

98 of the metabolites exhibit distributions that may declared normal, given a W-stat >= 0.95. In 81 instances (81%) of the tested metabolites the log10 data W-stat is < raw data W-stat.



5.5 Outliers

Evaluation of the number of samples and features that are outliers across the QC data. The below table presents the average number of outlier values for samples and features in the QC data set.

Table 7: Outlier Summary

	Min.	25th	Median	Mean	75th	Max.
Features	0	0	0	0	0	0
Samples	0	0	0	0	0	0

5.5.1 Notes on outlying samples at each metabolite|feature

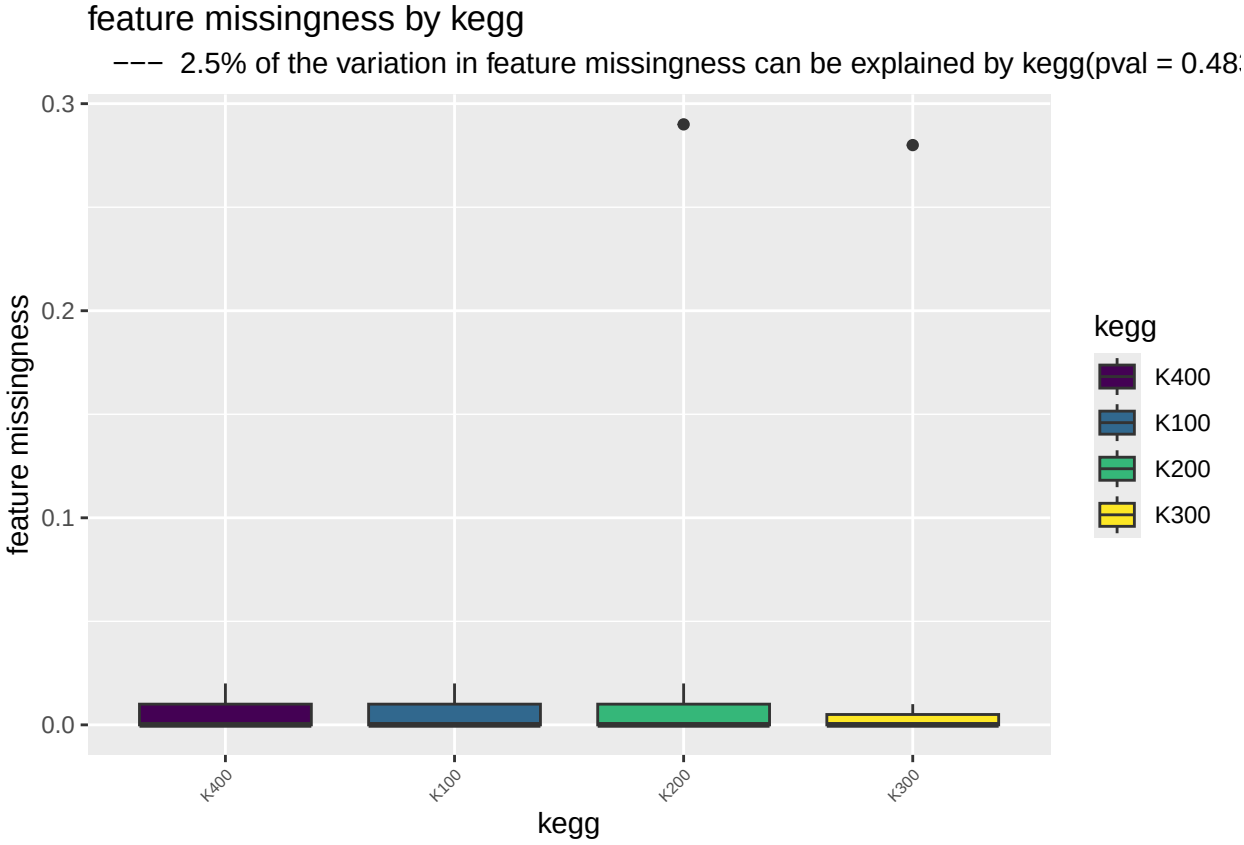
There may be extreme outlying observations at individual metabolites|features that have not been accounted for. You may want to:

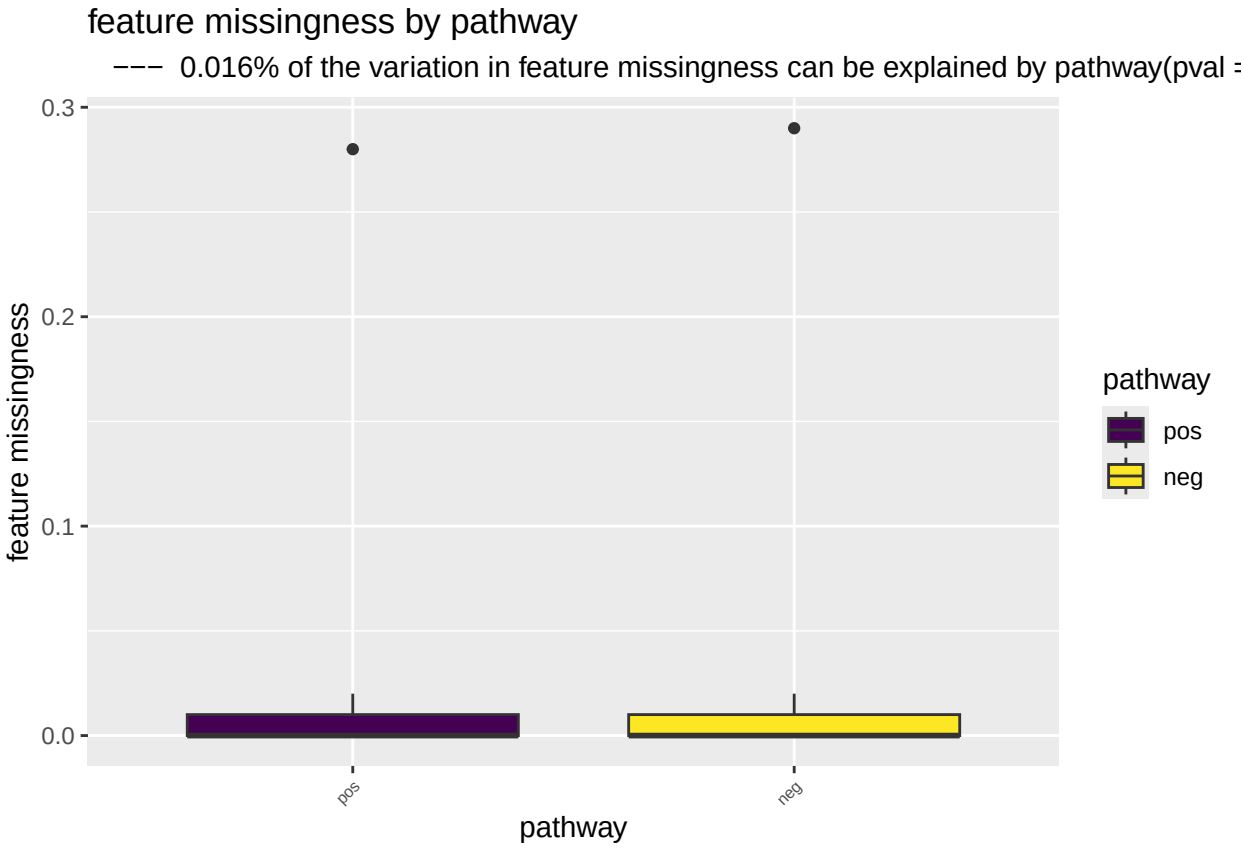
1. Turn these observations into NAs.
2. Winsorize the data to some maximum value.
3. Rank normalize the data which will place those outliers into the top of the ranked standard normal distribution.
4. Turn these observations into NAs and then impute them along with other missing data in your data set.

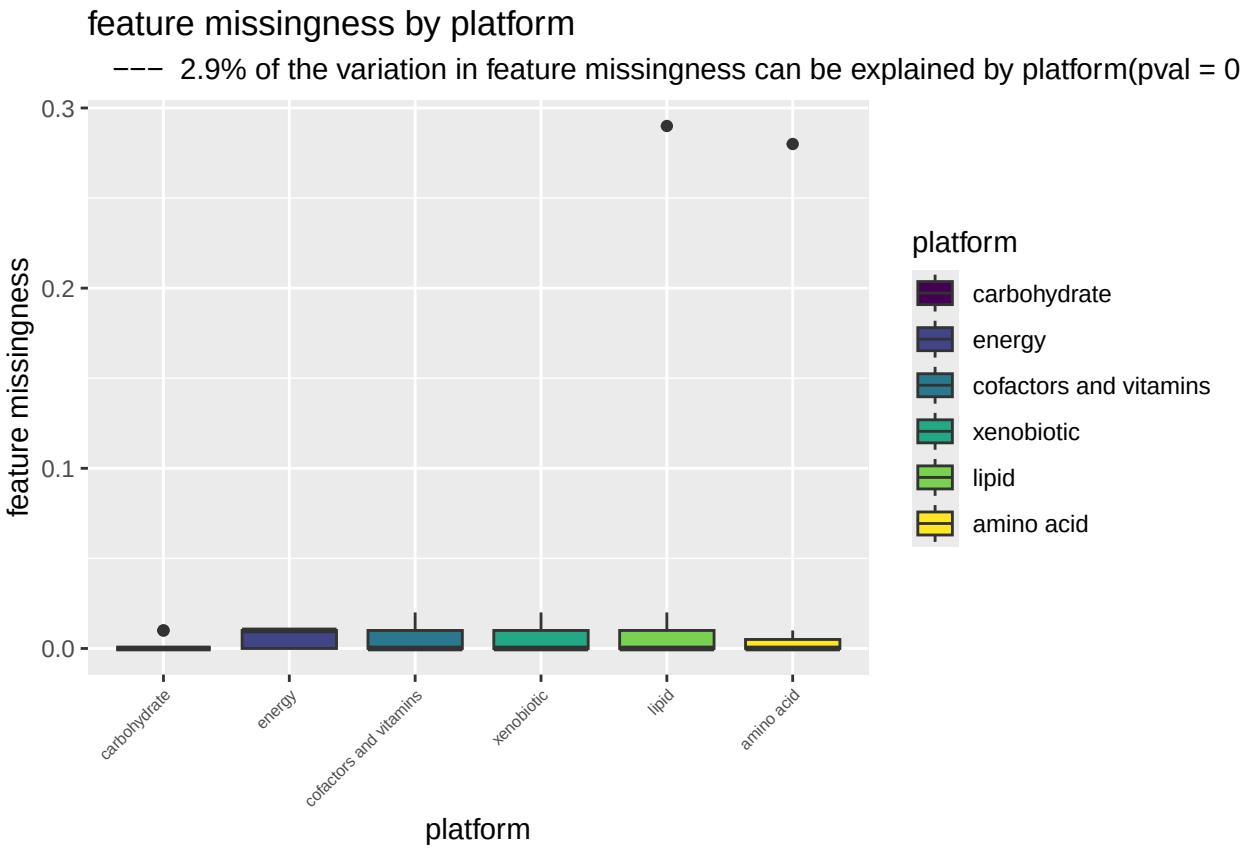
6 Variation in filtered data by available variables

6.1 Feature missingness

Feature missingness may be influenced by the metabolites' (or features') biology or pathway classification, or your technologies methodology. The figure(s) below provides an illustrative evaluation of the proportion of *feature missingness* as a product of the variable(s) available in the raw data files.




feature missingness by pathway
−−− 0.016% of the variation in feature missingness can be explained by pathway(pval =
feature missingness
0.3
0.2
0.1
0.0
pos
neg
pathway
pathway
pos
neg

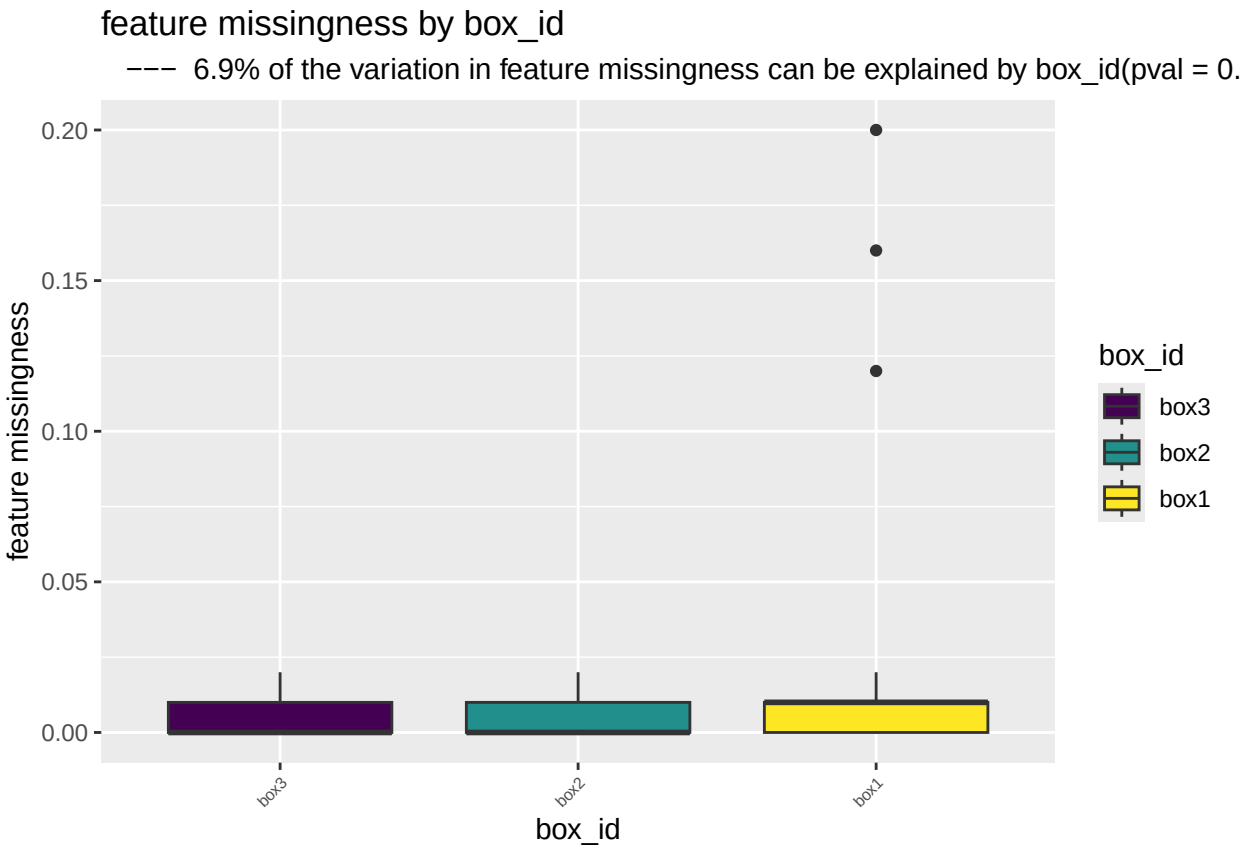


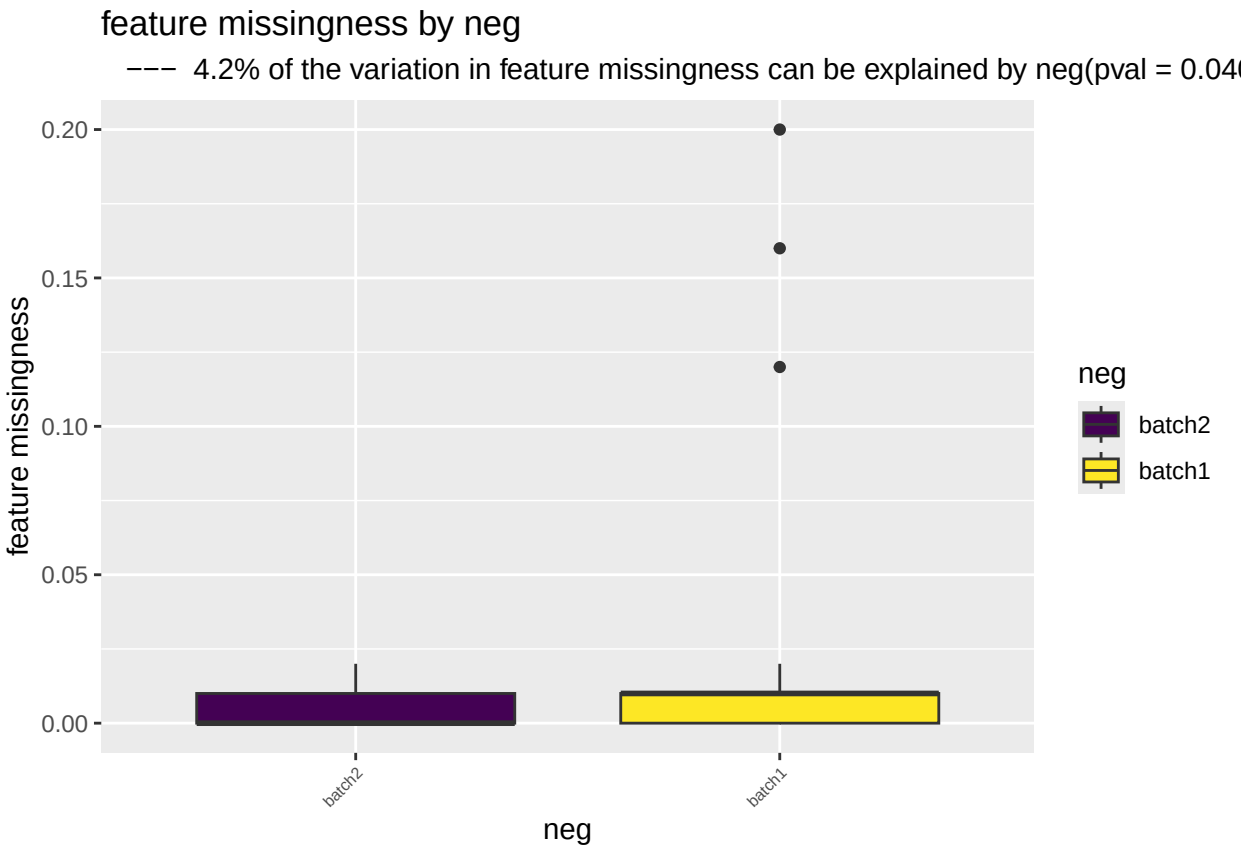
6.2 Sample missingness

The figure provides an illustrative evaluation of the proportion of *sample missingness* as a product of sample batch variables provided by your supplier. This is the univariate influence of batch effects on *sample missingness*. Box plot illustration(s) of the relationship that available batch variables have with sample missingness.

```
-- After filtering a total of 4 feature level batch variables were identified. --
-- They are:
   box_id
   neg
   pos
   run_day

-- After testing for redundancies a total of 2 feature level batch variables remain. --
-- They are:
   box_id
   neg
```


feature missingness by box_id
−−− 6.9% of the variation in feature missingness can be explained by box_id(pval = 0.
0.20
0.15
0.10
0.05
0.00
feature missingness
box3
box2
box1
box_id
box_id
box3
box2
box1



6.3 Multivariate evaluation: batch variables

TypeII ANOVA: the eta-squared (eta-sq) estimates are an estimation of the percent of variation explained by each independent variable, after accounting for all other variables, as derived from the sum of squares. This is a multivariate evaluation of batch variables on *sample missingness*. Presence of NA's would indicate that the model is inappropriate.

batch.variable	etasq.var.exp	pvalue
box.id	2.81	2.5413e−01
neg	0.06	8.1352e−01
residuals	97.13	NA

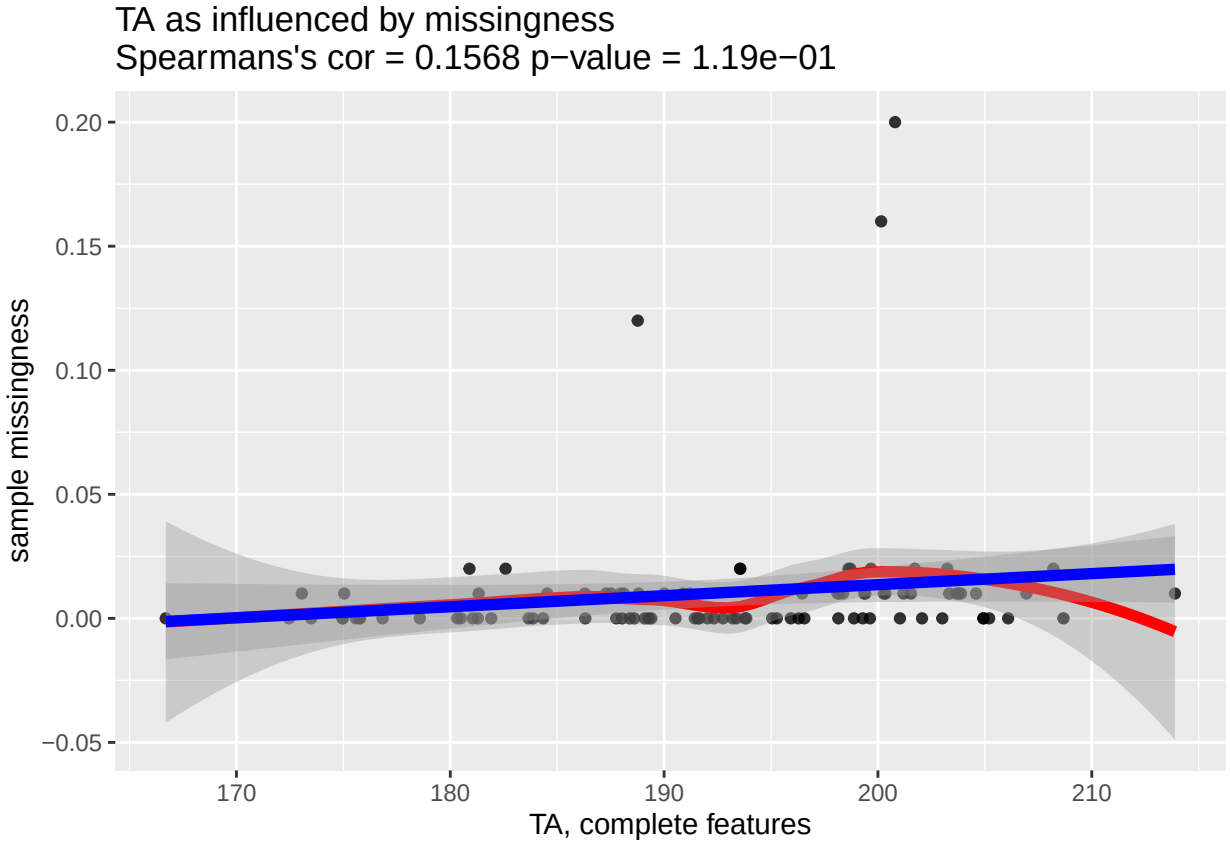
7 Total peak or abundance area (TA) of samples:

The total peak or abundance area (TA) is simply the sum of the abundances measured across all features. TA is one measure that can be used to identify unusual samples given their entire profile. However, the level of missingness in a sample may influence TA. To account for this we:

1. Evaluate the correlation between TA estimates across all features with PA measured using only those features with complete data (no missingness).
2. Determine if the batch effects have a measurable impact on TA.

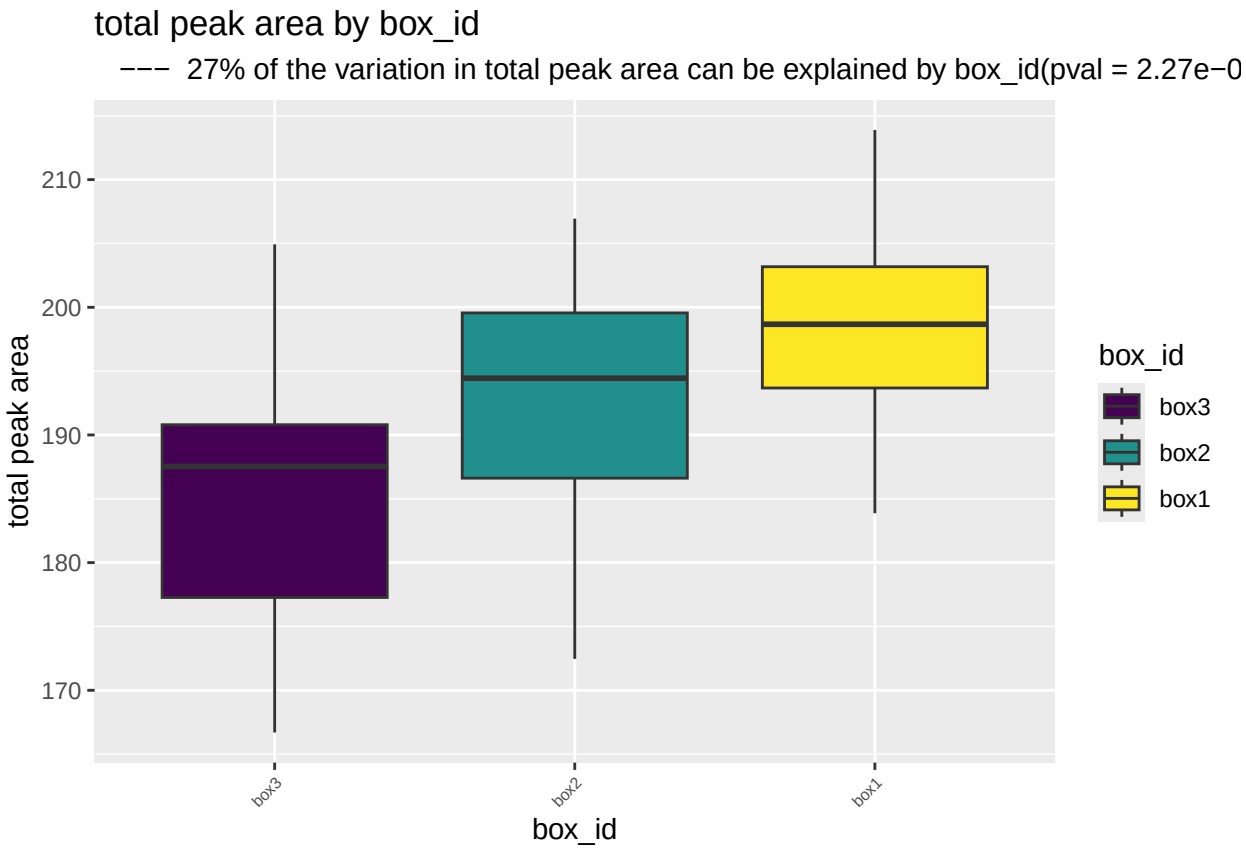
7.1 Relationship with missingness

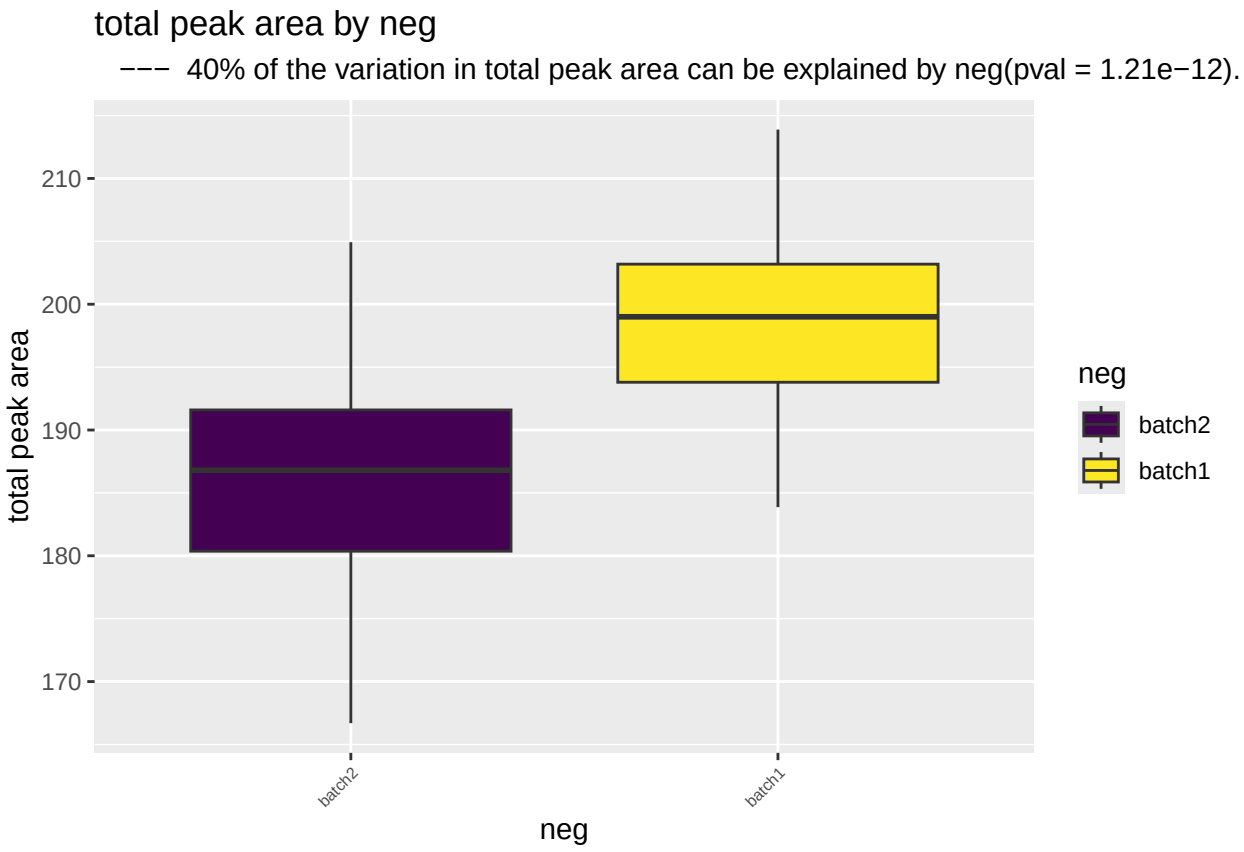
Correlation between total abundance (TA; at complete features) and missingness. Relationship between total peak area at complete features (x-axis) and sample missingness (y-axis).



7.2 Univariate evaluation: batch effects

The figure below provides an illustrative evaluation of the *total abundance* (at complete features) as a product of sample batch variables provided by your supplier. Violin plot illustration(s) of the relationship between total abundance (TA; at complete features) and sample batch variables that are available in your data.


total peak area by box_id
−−− 27% of the variation in total peak area can be explained by box_id(pval = 2.27e−0
210
200
190
180
170
total peak area
box3
box2
box1
box_id
box_id
box3
box2
box1



7.3 Multivariate evaluation: batch variables

TypeII ANOVA: the eta-squared (eta-sq) estimates are an estimation on the percent of variation explained by each independent variable, after accounting for all other variables, as derived from the sum of squares. This is a multivariate evaluation of batch variables on *total peak|abundance area* at complete features.

batch.variable	etasq.var.exp	pvalue
box.id	0.28	8.5020e−01
neg	18.52	9.4348e−06
residuals	81.2	NA

8 Power analysis

Exploration for case/control and continuous outcome data using the filtered data set

Analytical power analysis for both continuous and imbalanced presence/absence correlation analysis.

Simulated effect sizes (standardized by trait SD) are illustrated by their color in each figure. Figure (A) provides estimates of power for continuous traits with the total sample size on the x-axis and the estimated power on the y-axis. Figure (B) provides estimates of power for presence/absence (or binary) traits in an imbalanced design. The estimated power is on the y-axis. The total sample size is set to 99 and the x-axis depicts the number of individuals present (or absent) for the trait. The effects sizes illustrated here were chosen by running an initial set of simulations which identified effects sizes that would span a broad range of power estimates given the sample population's sample size.