







metaboprep: an R package for pre-analysis data description and processing of metabolomics data

D. A. Hughes ^{1,2}, K. Taylor^{1,2}, N. McBride^{1,2,3}, M. A. Lee^{1,2}, D. Mason⁴, D. A. Lawlor^{1,2,3}, N. J. Timpson^{1,2} & L. J. Corbin^{1,2}

- ¹ MRC Integrative Epidemiology Unit at the University of Bristol, UK
- ² Population Health Science, Bristol Medical School, University of Bristol, UK
- ³ NIHR Bristol Biomedical Research Centre, University of Bristol, Bristol, UK
- ⁴ Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK

metaboprep

Introduction

- Metabolomics is an increasingly common part of health research.
- > Even when working with high quality curated datasets, there is a need for pre-analytical data processing.
- While some pre-processing steps are common, we identified a lack of standardization and reporting transparency for these procedures.

sample outlier identification

Aim

To produce a **single**, **flexible** pre-processing pipeline to enable reproducible and transparent preprocessing of metabolomics data.

Estimate summary

statistics on

filtered data set

Data filtering

F) Replace xenobiotic features (unique to Metabolon data)

We created *metaboprep*, an openly available and

Conclusion

citable R package available on GitHub. https://github.com/MRCIEU/metaboprep

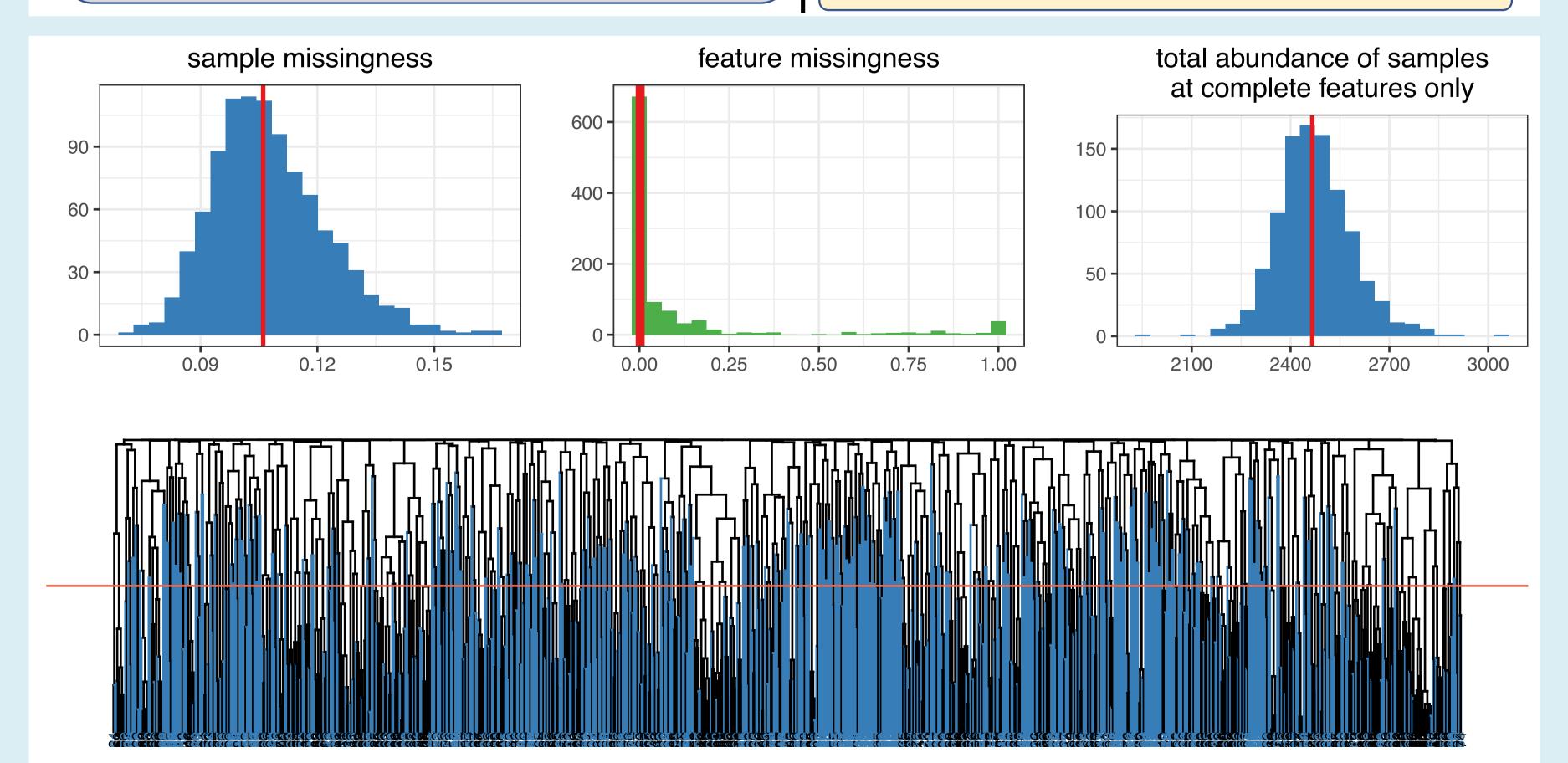
Study Pre-Laboratory Statistical design & processing data and data sample analyses generation collection filter

Generate

PDF

Report

Extract data from Estimate summary norm metaboprep Perform data commercial excel statistics on raw alizat pipeline filtering sheet ion data set Summary statistics A) Remove xenobiotic features (unique to Metabolon data) A) Sample summary statistics (dme) *dme* = derived sample missingness measures using all features B) Exclusion based on extreme missingness excluded to the exclusion of xenobiotics 1) sample missingness >= 80% (dme) total sum abundance feature missingness >= 80% using all features using features with no missingness C) Exclusions based on user defined missingness sample outlier occurrence across all features Re-estimated sample missingness >= 'X'% (dme) Re-estimated feature missingness >= 'X'% B) Feature summary statistics D) Sample exclusions based on user defined sample feature missingness a) using all samples characteristic (dme) total sum abundance >= +/- 'X' SD from the mean b) to the exclusion of samples >= 50% missingness feature descriptions (mean, sd, skewness W-stat, etc..) sample outlier incidence E) Data Reduction (dme) data reduction: (dme) build feature dendrogram build dendrogram identify representative features identify representative features G) Sample exclusion based on user define sample structure C) Sample structure using data from E2 principal component analysis (using data from B4) PC1-&-PC2 eigenvectors >= +/- 'X' SD from mean missing values imputed to median at minimum PC1 & PC2 or all significant PCs as ascertained by an acceleration analysis



Systems biology

description and processing

Advance Access Publication Date: 4 February 2022

metaboprep: an R package for preanalysis data



All feedback is welcome! d.a.hughes@bristol.ac.uk laura.corbin@bristol.ac.uk The *metaboprep* package:

- 1. Reads in and processes (un)targeted metabolomics data and exports in tab-delimited format;
- 2. Provides useful **summary data** in the form of tab-delimited text file and html report;
- 3. Performs data filtering using a standard pipeline and user-defined thresholds.

