MetaboDiff: an R package for differential metabolomic analysis.

Andreas Mock and Christel Herold-Mende

Division of Experimental Neurosurgery, Department of Neurosugery, Heidelberg University Hospital

Introduction Comparative metabolomics comes of age by an increasing list of commercial vendors (i.e. Metabolon®) offering reproducible high-quality metabolomic data for translational researchers outside the mass spectrometry field. This R packages aims to provide a low-level entry to differential metabolomic analysis by starting off with the table of relative metabolite quantifications provided by commercial vendors.

Installation of *R* package

```
library(devtools)
install_github("andreasmock/MetaboDiff")
```

The package vignette can be found on Github at
https://github.com/andreasmock/MetaboDiff.

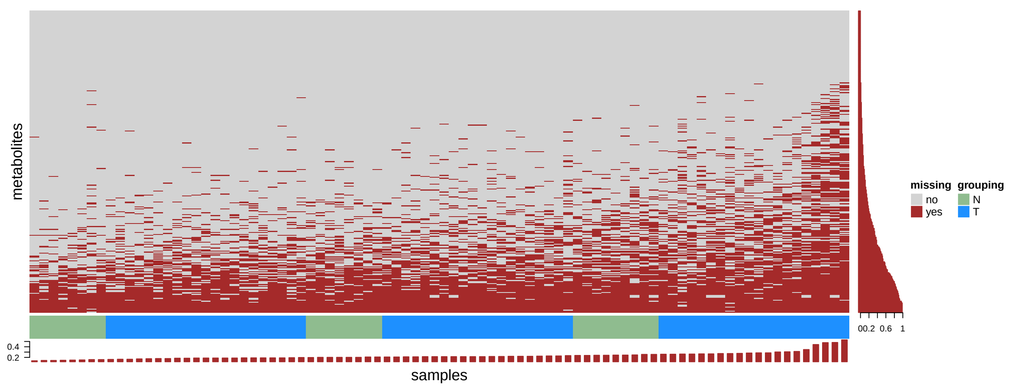
Example data The example data is derived from a study by Priolo and colleagues in which they used the service of the Metabolon company to compare the tissue metabolome of 40 prostate cancers with 16 normal prostate specimens.

Data representation The metabolomic data within `MetaboDiff` is stored as a `MultiAssayExperiment` class (Sig, 2015). This framework enables the coordinated representation of multiple experiments on partially overlapping samples with associated metadata and integrated subsetting across experiments. In the context of metabolomic data analysis, multiple assays are needed to store raw data and imputed data.

The core components of the `MultiAssayExperiment` class are:

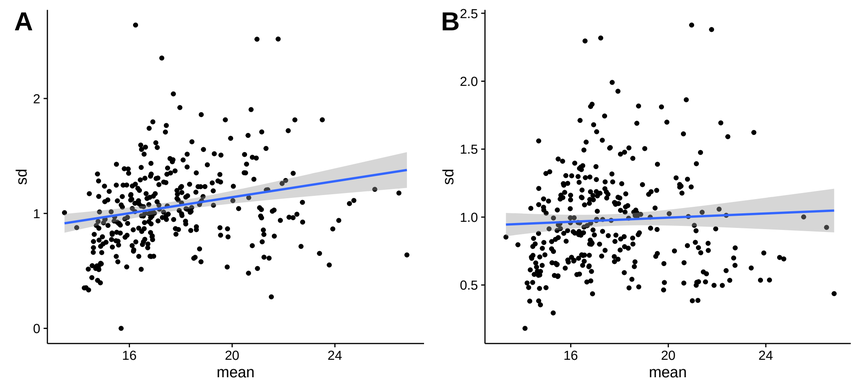
- `ExperimentList` - a slot of class ExperimentList containing data for each experimental assay. Within the ExperimentList slot, the metabolomic data is stored as a SummarizedExperiment object consisting of:
  - `assay` - a matrix containing the relative measurements.
  - `rowData` - a dataframe containing the metabolite annotation.
- `colData` - a slot of class data frame describing the sample metadata available across all experiments.
- `sampleMap` - a slot of class data frame relating clinical data to experimental assay.

Imputation of missing values In contrast to microarrays, missing values are common in quantitative metabolomic datasets. The following heatmap shows the missing values across the example data. K-nearest neighbor imputation was used to minimize effects on the normality and variance of the data (Armitage et al., 2015).

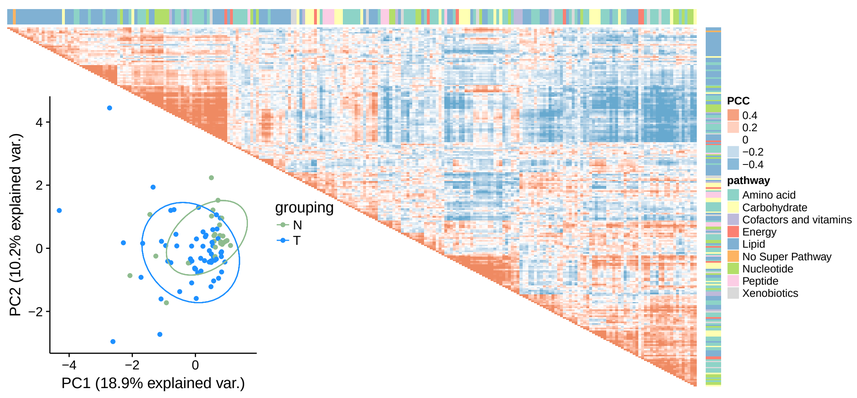


About this poster This poster was created using a R markdown template of rOpenGov (http://ropengov.github.io). It is fully reproducible; the full source code of this poster is available at https://github.com/andreasmock/MetaboDiff_poster.

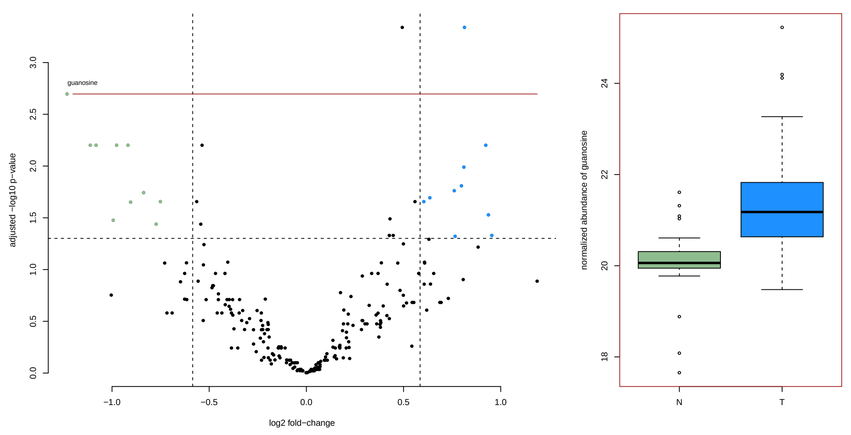
Normalization Variance stabilizing normalization (vsn) was originally developed for microarrys to ensure that the variance remains nearly constant over the whole intensity spectrum (Huber et al., 2002). It could be shown that vsn normalization performs also very well for metabolomic data (Kohl et al., 2012). The following plots illustrate the variance stabilization before (A) and after normalization (B) in the example data:



Unsupervised analysis A number of unsupervised analysis and visualizations are at offer within `MetaboDiff` including correlation heatmaps and PCA plots:



Differential analysis Metabolites are compared using Student T-Tests. Correction for multiple testing is performed by independent hypothesis weighting (IWH; Ignatiadis et al., 2016) with variance as a covariate. Differential pathways are identified by enrichment analyses.



To do

- Implementation of HotNet2 algorithm to identify significantly altered subpathways (Pleiserson et. al, 2014)
- Hive plot of metabolic network (Krzywinski et al., 2012)

References

1. Armitage, EG et al. (2015). Missing value imputation strategies for metabolomics data. Electrophoresis, 36(24), 3050–3060.
2. Huber, W et al. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics, 18 Suppl 1, S96–104.
3. Ignatiadis, N et al. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nature Methods, 13(7), 577–580.
4. Kohl, SM et al. (2012). State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics, 8(Suppl 1), 146–160.
5. Krzywinski, M et al. (2012). Hive plots–rational approach to visualizing networks. Briefings in Bioinformatics, 13(5), 627–644.
6. PLeiserson, MDM et al. (2014). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature Genetics, 47(2), 106–114.
7. Sig, M (2017). MultiAssayExperiment: Software for the integration of multi-omics experiments in Bioconductor. R package version 1.2.1