Bioinformatics, 37(15), 2021, 2218–2220
doi: 10.1093/bioinformatics/btaa967
Advance Access Publication Date: 16 November 2020
Applications Note

Systems biology

Metabolite-Investigator: an integrated user-friendly workflow for metabolomics multi-study analysis

Carl Beuchel[1], Holger Kirsten [1,2], Uta Ceglarek[3] and Markus Scholz[1,2,4,*]

[1]Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, 04107 Leipzig, Germany, [2]LIFE – Leipzig Research Center for Civilization Diseases and [3]Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, Leipzig University, 04103 Leipzig, Germany and [4]IFB AdiposityDiseases, University Hospital Leipzig, 04103 Leipzig, Germany

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on June 16, 2020; revised on October 13, 2020; editorial decision on November 1, 2020; accepted on November 4, 2020

Abstract

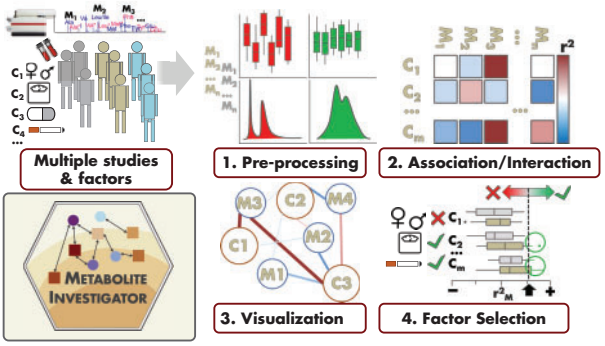
Motivation: Many diseases have a metabolic background, which is increasingly investigated due to improved measurement techniques allowing high-throughput assessment of metabolic features in several body fluids. Integrating data from multiple cohorts is of high importance to obtain robust and reproducible results. However, considerable variability across studies due to differences in sampling, measurement techniques and study populations needs to be accounted for.

Results: We present *Metabolite-Investigator*, a scalable analysis workflow for quantitative metabolomics data from multiple studies. Our tool supports all aspects of data pre-processing including data integration, cleaning, transformation, batch analysis as well as multiple analysis methods including uni- and multivariable factor-metabolite associations, network analysis and factor prioritization in one or more cohorts. Moreover, it allows identifying critical interactions between cohorts and factors affecting metabolite levels and inferring a common covariate model, all via a graphical user interface.

Availability and implementation: We constructed *Metabolite-Investigator* as a free and open web-tool and standalone Shiny-app. It is hosted at https://apps.health-atlas.de/metabolite-investigator/, the source code is freely available at https://github.com/cfbeuchel/Metabolite-Investigator.

Contact: markus.scholz@imise.uni-leipzig.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.



1 Introduction

Many diseases have a metabolic background, which is increasingly recognized in basic and medical research (Liu *et al.*, 2019). Accordingly, metabolomics measurements were commenced in large cohorts to unravel disease mechanisms and to develop effective treatment concepts (Abbiss *et al.*, 2019; Iida *et al.*, 2019; Souza *et al.*, 2019). Specifically, metabolomics by mass-spectrometry allows analysis of defined metabolomics features in high-throughput and with high sensitivity in several body fluids and cellular compartments (Cambiaghi *et al.*, 2017).

We here propose a general analysis framework to analyse large-scale metabolic data of several studies in parallel, considering typical data analysis issues of LC-MS-based metabolomics data. Particularly, the tool allows for (i) Metabolite pre-processing accounting for outliers, zero-inflation (an excess of zero values in the data, resulting in skewed distributions) and technical batch-effects, (ii) Identification of uni- and multivariable effects of individual factors influencing the metabolome and analysis of heterogeneity of these effects across studies, (iii) Selection and prioritization of factors as covariates for metabolite analyses. With these features, our tool is particularly well suited to verify, e.g. the effect of factors on metabolites across independent studies.

We optimized these methods and analysis steps on the basis of a large simulation study to propose a general workflow applicable to many situations, combining streamlined pre-processing with a comprehensive multi-study analysis not yet available in other applications. The workflow was implemented as an interactive Shiny-application (Winston Chang *et al.*, 2019) called *Metabolite-Investigator*, providing an intuitive graphical-user-interface. We here present the implemented methods, workflow and an example application based on a real-world example.

2 Implemented methods and workflow

The general analysis workflow of *Metabolite-Investigator* is depicted in Figure 1 and a detailed description of the methods, as well as a step-by-step presentation of the workflow can be found in Supplementary Methods. Part of the workflow was also applied previously in a study with real data (Beuchel *et al.*, 2019). Briefly, the tool starts with tabular, raw quantitative metabolome data, also allowing upload of data using standardized file formats mwTab and mzTab-m 2 and performs data integration and further pre-processing considering a number of typical issues of these kind of data (Hoffmann *et al.*, 2019). The primary function of the tool is to identify and visualize relationships between individual factors such as age, sex or laboratory parameters like blood cell counts with metabolomics features and to select a suitable covariate model for joint regression analysis of data from one or multiple cohorts.

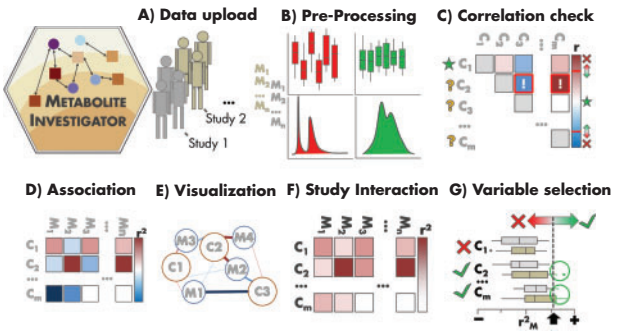


Fig. 1. Schematic workflow of the tool *Metabolite-Investigator*. The application identifies and characterizes relationships between influencing factors and metabolomics features in single or multiple studies. The workflow is structured as follows: (**A**) available metabolite (M_1, M_2, …, M_n) and factor data (C_1, C_2, …, C_m) of all cohorts are loaded into the application and merged. (**B**) Data are pre-processed by filtering outlier measurements, inverse-normal-transformation and batch-adjustment according to the requirements of the user. (**C**) Pearson's correlations of factors are computed and visualized. Highly correlating factors (marked with a red box and exclamation mark) can be excluded from the multivariable association step to avoid multi-collinearity issues. (**D**) Factor-metabolite relationships are analysed in uni- and multivariable association steps separately in each study. (**E**) Visualizations of the strength of identified relationship between factors and metabolites as interactive bi-partite network. (**F**) Analysis and visualization of effect heterogeneity across multiple studies based on testing for uni- and multivariable interaction effects of the factors with the study identifier. (**G**) Factors are prioritized to assist in the identification of a confounding model applicable to all cohorts for follow up analyses. This is done via backward-selection of factors that explain a user-specified minimum amount of variance in all cohorts. Visualizations and results are available for download

The users can follow a structured workflow comprising seven steps shown in Figure 1 and explained in detail in Supplementary Methods. Results are available for download after each step. Thereby, single measurements larger than 5*SD of log-transformed values can be removed as outliers. Possibly skewed data are optionally mitigated by inverse-normal transformation (Beasley *et al.*, 2009). Known technical batch-effects can be addressed by an empirical Bayes method (Leek *et al.*, 2012; Johnson *et al.*, 2007). Univariable linear regression models are fit for each study with metabolites treated as responses and factors as possible predictors. Highly correlating factors are identified and can be removed from further analysis. Remaining factors are fit as joint predictors in multivariable linear regression models of the metabolites. Visualizations of the distribution of effect sizes of factors on metabolites per study as well as factor correlations are presented. The heterogeneity of factor effects on metabolite levels across cohorts is assessed by fitting a factor-by-cohort interaction term in addition to the main effects in the pooled dataset.

Backward-selection of factors is implemented by fitting all remaining factors as predictors in multivariable regression models of metabolites and removing factors with small explained variance. This procedure is repeated iteratively until a subset of factors is obtained meeting the user-specified threshold of explained variance for at least one of the metabolites in at least one of the studies (see Supplementary Methods for details). For the user's convenience, a pre-formulated methods description is available after analyses.

3 Case study

For illustration purposes, we analyse a public dataset of 66 acylcarnitines quantified using LC-MS/MS and ten experimental factors collected from 366 samples after six weeks of atenolol treatment, that is freely available at http://dx.doi.org/10.21228/M8PC7J. Our tool allows for quick parsing of the corresponding mwTab format and recognition of analysis-ready columns. After pre-processing and feature annotation, high zero-inflation of all but two metabolites (Acetylcarnitine and Malonylcarnitine) was detected. The backward selection of factors influencing metabolites resulted in a model of three factors, namely race (for Malonylcarnitine), age and baseline glucose (for Acetylcarnitine) (see Supplementary Methods for a more detailed description).(Metabolomics Workbench, 2016)

4 Performance

Metabolite-Investigator is scalable to thousands of samples and hundreds of variables. A public dataset ($N = 366$, 66 metabolites, 10 factors) was analysed in under 1 min. Completing all analysis steps for a larger test dataset of 63 metabolites and 10 factors with $N = 15\,260$ ran for 12 min using all standard settings on our test machine (16 GB RAM, 3.4 GHz CPU).

5 Comparison with existing tools

We compared our tool with other graphical user interface-based tools intended for pre-processing and analysis of metabolite data. Results are presented in Supplementary Table S1. Our tool addresses a number of issues not considered in available tools. These comprise appropriate dealing with typical LC-MS/MS data issues including skewness of data and zero-inflation, technical batches, identification of external sources of

variation thereby accounting for correlation, scalable identification and prioritization of factors affecting the metabolome (especially when integrating data from multiple studies), and finally, visualization of results, including interactive network analysis.

6 Conclusion

Metabolite-Investigator is an easy-to-use application, offering a standardized workflow for analysis and prioritization of factors affecting metabolite data. Local deployment and availability on a web server allow for online and offline use with no programming skills required. It offers a convenient way for (multi-)study preprocessing and subsequent graphical and table-based exploration of results including study heterogeneity and multivariate effect size based model selection currently not available elsewhere.

Funding

M.S. received funding from the Federal Ministry of Education and Research, Germany, FKZ: 01EO1501.AD2-7117. H.K. was supported by the Leipzig Research Center for Civilization Diseases (LIFE). LIFE is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by funds of the Free State of Saxony within the framework of the excellence initiative. The Leipzig Health Atlas is funded by the German Ministry of Education and Research (reference number: 031L0026, program: i: DSem – Integrative Datensemantik in der Systemmedizin).

Conflict of Interest: none declared.

References

Abbiss,H. *et al.* (2019) Metabolomics approaches for the diagnosis and understanding of kidney diseases. *Metabolites*, **9**, 34.

Beasley,T.M. *et al.* (2009) Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav. Genet.*, **39**, 580–595.

Beuchel,C. *et al.* (2019) Clinical and lifestyle related factors influencing whole blood metabolite levels – a comparative analysis of three large cohorts. *Mol. Metab.*, **29**, 76–85.

Cambiaghi,A. *et al.* (2017) Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Brief. Bioinf.*, **18**, 498–510.

Chang,W. *et al.* (2019) shiny: Web Application Framework for R. https://CRAN.R-project.org/package=shiny.

Hoffmann,N. *et al.* (2019) mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal. Chem.*, **91**, 3302–3310.

Iida,M. *et al.* (2019) Application of metabolomics to epidemiological studies of atherosclerosis and cardiovascular disease. *J. Atheroscl. Thromb.*, **26**, 747–757.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, **8**, 118–127.

Leek,J.T. *et al.* (2012) The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, **28**, 882–883.

Liu,X. *et al.* (2019) New advances in analytical methods for mass spectrometry-based large-scale metabolomics study. *TrAC Trends Anal. Chem.*, **121**, 115665.

Souza,R.T. *et al.* (2019) Use of metabolomics for predicting spontaneous preterm birth in asymptomatic pregnant women: protocol for a systematic review and meta-analysis. *BMJ Open*, **9**, e026033.