Slovak University of Technology in Bratislava
Faculty of Mechanical Engineering
17th Conference on Applied Mathematics **APLIMAT 2018**

Proceedings

NORMALIZATION TECHNIQUES FOR UNIVARIATE BIOSTATISTICS ANALYSIS

ŠEDIVÁ Blanka (CZ), PELANTOVÁ Helena (CZ), BUGÁŇOVÁ Martina (CZ)

Abstract. The biostatistic processing of metabolomic data includes a number of mathematical and statistical methods. The series of preprocessing steps is necessary applied before application of the univariate or multivariate approaches for identifications of statistical significant factor.
The goal of this article is using simulation Monte Carlo approach for analysis of influence of normalizations processes on the results of univariate statistical methods. In our study four methods of normalization are compared - normalization by the area under the curve (AUC), normalization to creatinine, quantile normalization and probabilistic quotient normalization (PQN). The results of simulation experiments studies have shown that PQN, quantile normalization and creatinine normalization are more robust than AUC normalization, especially in case a small number of metabolites with a large fold change is presented. From a practical point of view PQN method is recomended as the robust normalization procedure with the broad application and easy data interpretation.

Keywords: data normalization, preprocessing, simulation, metabolomics

Mathematics subject classification: Primary 65C05; Secondary 62P10, 62H30

1 Introduction

The problem of multidimensional correlated data sets is very common in biostatistics. In these situations, mathematical statistical procedures focus on suitable data preprocessing methods, in particular on data standardization problems and other preprocessing methods that enable us to eliminate the effect of the large variability of biological samples on the one hand and to identify significant difference factors on the other. This problem is common for all so called "-omics" methods and metabolomics is no exception.

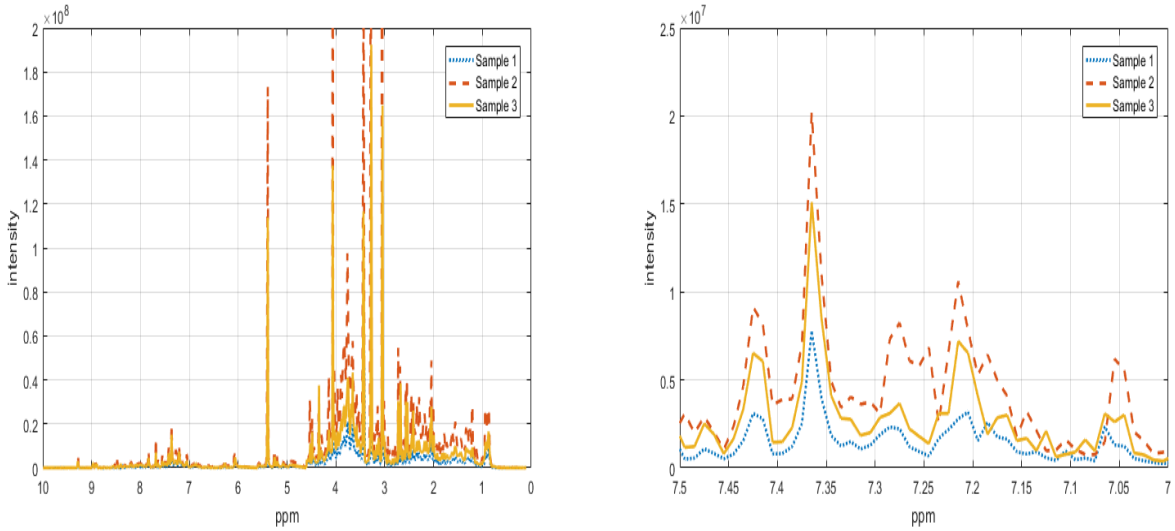
Metabolomics is focused on the comprehensive characterization of metabolites in biological systems (humans, animals, plants, bacteria, etc.); in humans all biological materials from biofluids (blood, urine) up to tissues can be analyzed (Alonso 2015 [1]). Therefore, it is increasingly being used in almost all fields of health science including pharmacology, pre-clinical drug trials, toxicology, newborn screening and many others. The metabolome of being as reflected in biofluids is defined

by genetic factors, but it can also be affected by diet, age, disease, etc. One of the most prominent analytical methods, which is applied for analysis of biological samples in metabolomic studies is nuclear magnetic resonance (NMR) spectroscopy.

It can provide one- or multidimensional spectra. For metabolomics are predominately used proton NMR spectra (^{1}H NMR), which consist of all hydrogen resonances of metabolites present in analyzed sample. The position and shape of detected signals reflect the structure of molecules and their intensity is proportional to concentration in the sample. Thus, we can both identify molecules constituting the sample (usually by comparison with databases) and directly determine their concentration.

The NMR spectra require extensive data pre-processing prior to data analysis. It starts with the Fourier transformation, baseline correction, binning, which allows to transform the continuous signal to discrete datasets and the normalization of the dataset. Once the complete metabolic sets are generated, the univariate or multivariate statistical methods focused directly on the identification of differences between the groups of samples can then be applied.

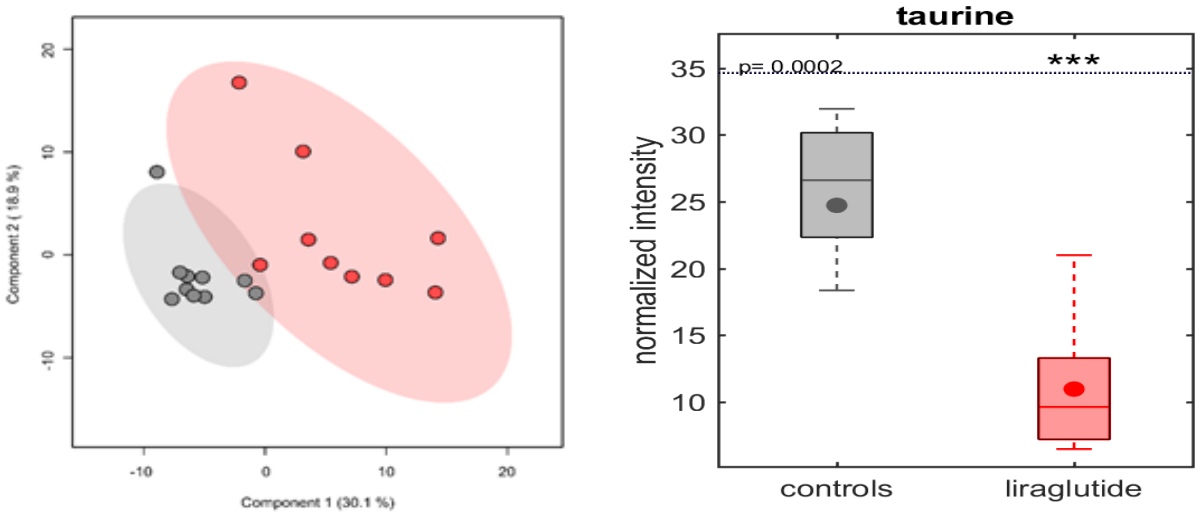
In Fig. 1 displays representative urine spectra of three samples. The result of the univariate and multivariate analysis can be presented, for example, as a boxplot for selected metabolites or 2D score plots of the first and second components after Principal Component Analysis (PCA) or Partial Least Squares - Discriminant Analysis (PLS-DA). In Fig. 2, we show the visualization of the final statistical results of the effects of liraglutide in mice with diet-induced obesity studied by metabolomics (Buganova 2017 [2]).



(a) An example of three sample spectra. (b) A zoomed view of the spectra in (a).

Fig. 1. Examples of spectra obtained with NMR spectroscopy technologies.

Urine is one of the most common biofluid used for metabolomics. Unfortunately, urine samples can substantially vary in their dilution, which is caused by different water consumption and other physiological factors. Consequently, the concentrations of metabolites are influenced. Therefore, proper normalization for these effects prior to further statistical processing is recommended (Alonso 2015 [1]).



(a) 2D scores plots as a result of PLS-DA.

(b) Boxplots for intensity of selected metabolite in control group and treated group.

Fig. 2. The visualization of univariate and multivariate statistical results.

The aim of this article was to study the influence of the normalization methods of raw datasets on the results and conclusions based on univariate statistics of identified significant features. Raw data of ^{1}H NMR spectra of urine from experiments described in (Pelantova 2016 [9]) were used as a background of the dataset.

We performed four methods of data normalization - normalization by the area under the curve (AUC normalization), normalization to creatinine, quantile normalization and probabilistic quotient normalization (PQN). The effect of these types of normalization on the results of the parametric t-test and non-parametric Mann-Whitney test was monitored and analyzed. The source data was artificially modified with the aim of assessing the effect of different combinations of normalization methods in relation to the response and fold change. We artificially modified peak intensities to introduce known metabolic differences between groups and this provided us with a target for discovering these artificial peak intensities using both univariate and multivariate statistics. The effects of the processing steps on our ability to re-discover these known metabolic differences were evaluated.

2 Methods

2.1 Data set, raw data preprocessing and simulations

Real datasets from studies (Pelantova 2016 [9]) were used as a background for simulation studies. In order to get as close as possible to real-world experiments, two sets (the number of samples in each set is $n = 8$) of observation were always generated. For the first step of the preprocessing approach, we applied the binning with the fix bin step 0.01 ppm.

The structure of generated samples was derived from real binned spectra using bootstrapping methods. For the second set (simulated treated group), spectral values in the area from 6.445 ppm to 6.485 ppm were artificially increased. The additional increasing of intensities was based on the intensity values from 10^5 to 10^8 and in percentages from 1% to 150%.

The resulting generated dataset was arranged into a data matrix with rows $1-8$ corresponding to unmodified samples and rows $9-16$ to an included part with artificial values. The amount of analyzed bins (corresponding to the number of columns of the data array) was 922. The simulation processes were repeated $1,000$ times.

2.2 Normalizations

The generated datasets were normalized by four methods recommended for normalizations of spectral datasets in the literature Di Guida 2016 [4], Euceda 2012 [5] or Kohl 2012 [8].

(a) AUC normalization (area under curve normalization, normalization by sum) is a method where each value in a row (sample) is divided by the total sum of the row (sample) and multiplied by 100; the unit is $\%$.

(b) Normalization to creatinine is related to the very specific metabolite, called creatinine, that is presented in all urine samples. Creatinine is a chemical waste product in blood (by-product of normal muscle contractions) that passes through the kidneys to be filtered and eliminated in urine. Under normal conditions, urinary creatinine output is relatively constant and measurable.

(c) Quantile normalization consists of two steps: in the first step a mapping between ranks and values is created. For rank 1, find the n values, one per array, that are the smallest value on the array, and save their averages. Similarly, for rank 2 and the second smallest values, and on up to the largest n values, one per array. In the second step, the actual values for each array is replaced with these averages.

(d) Probabilistic quotient normalization (Dieterle 2016 [3]) assumes that biologically interesting concentration changes influence only parts of the NMR spectrum, while dilution effects will affect all metabolite signals. PQN begins with an AUC normalization of each spectrum, followed by the calculation of a reference spectrum such as a median or mean spectrum. Next, for each variable of interest, the quotient of a given test spectrum and reference spectrum is calculated and the median of all quotients is estimated. Finally, all variables of the test spectrum are divided by the median quotient.

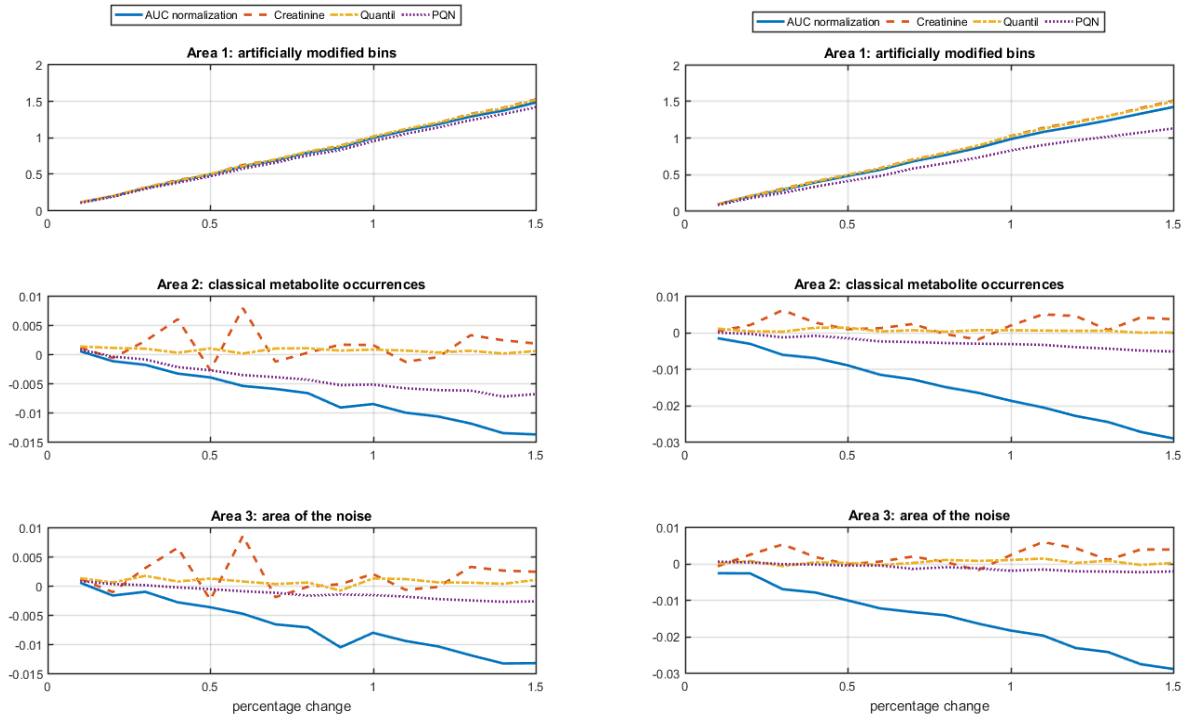
2.3 Statistical analyses

After the application of the preprocessing approach with one of the selected normalization procedures, two univariate methods of identification differences between metabolites in the "control" group and "simulated treated" group were calculated. The parametric t-test and the nonparametric robust test, the Mann-Whitney test, were chosen for testing the differences between groups, for more information about the advantages and disadvantages of univariate methods see Saccenti 2014 [10].

We focused on identifying the differences in the area of the added metabolite (area 1) in the area of common metabolite occurrences (area 2) as well as in the area of the noise (area 3). Appropriate standardization should reveal the diffusion in the area of the artificially enhanced spectrum (area 1) but at the same time not identify the differences in other areas (especially in area 2) because it would be the differences caused by the inappropriately chosen standardization.

3 Results and discussion

Our result from the influence of several types of standardization processes is confirmed by other simulation studies, Gardlo 2016 [6] and Hertel 2017 [7]. In Fig. 3 and Fig. 4 we show the typical results for artificial metabolite simulation with gradual increments of 1% to 150%. The sensitivity of the results is dependent on the intensity value of bins that is modified. To illustrate the impact of the intensity value, two situations are presented - the intensity value of bins (in the modified area 1) is equal to 10^6 or 10^7.



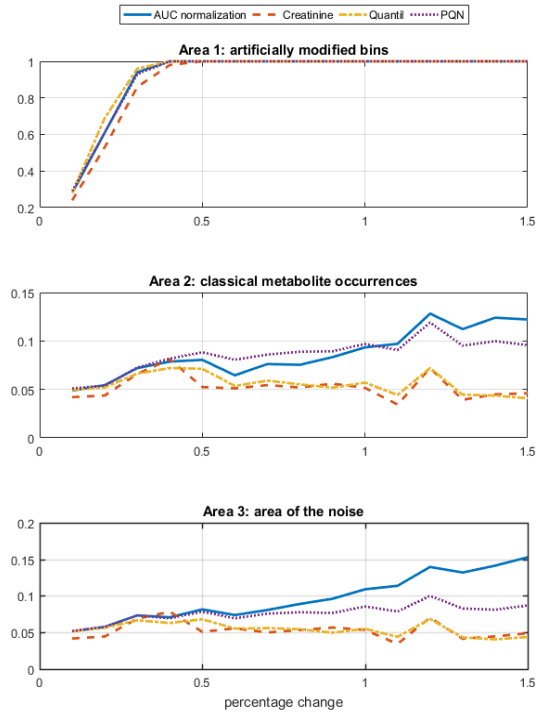
(a) The value of intensities of modified bins - 10^6 (b) The value of intensities of modified bins - 10^7

Fig. 3. The dependencies of changing values of the bins for four methods of the normalization and for three types of bin areas.

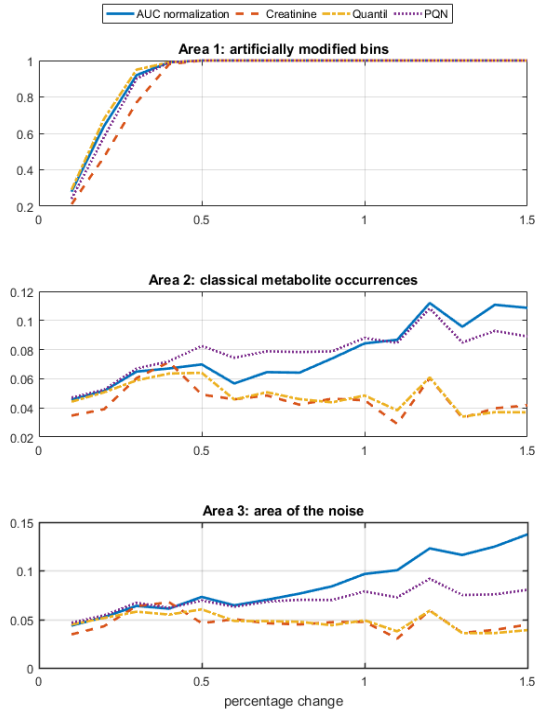
The x-axis in all graphs describes the percentage of the artificial bin changes. In Fig. 3, the dependencies of the changing values of the bins for four types of normalizations and for three types of bin areas are presented. It is clear that the changing of the values of bins in the modified area is proportionally increasing for all types of normalizations. Larger differences between normalization methods were observed in the second and third areas. The changes in area 1 also greatly influenced the bins after normalizations in area 2 and area 3. This effect is the biggest one for AUC normalization and PQN.

In Fig. 4, the results of parametric (left) and nonparametric tests (right) are presented. The curves capture the percentage of tests in which the hypothesis on the equality of the expected value in both groups of samples was rejected.

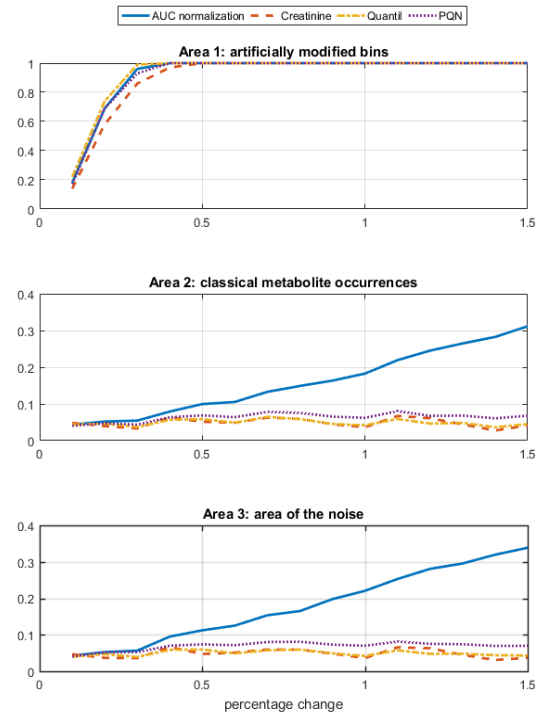
The AUC normalization shows the greatest problems in the field of the identification of fake differences in in all three areas under study. The problematic influence of the AUC normalization increases with the absolute value of modified bins before the normalization process. The percentage of false significant differences grew rapidly for this method of normalization. The PQN increases the number



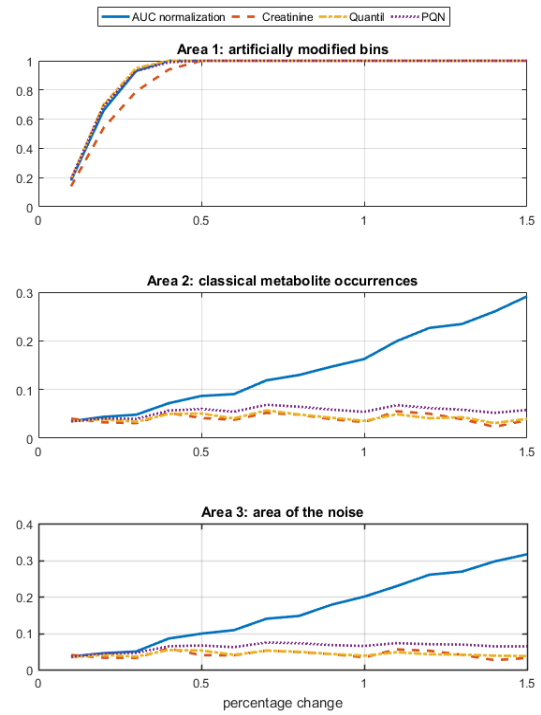
(a) T-test for the intensity value 10^6.



(b) M-W test for the intensity value 10^6.



(c) T-test for the intensity value 10^7.



(d) M-W test for the intensity value 10^7.

Fig. 4. The visualization of results for simulation additive artificial metabolite for addition value of intensities 10^6 and 10^7 (y-axis represents the percentage of simulations for which the null hypothesis of the equality of the mean is rejected).

of false identified differences for the smaller absolute value of modified bins. On the contrary, the quantile normalization and normalization based on creatinine did not show this negative behavior and the parametric and nonparametric tests in the area 2 and 3 did not cross the 5% limit.

4 Conclusions

The problem of choosing the best normalization procedure for a certain dataset can be solved empirically. After our simulation studies, we do not recommend using the AUC normalization process because this approach can increase the number of falsely identified differences of bins. On the other hand, we found that PQN, quantile and normalization based on creatinine provide comparable results. Datasets normalized by these approaches are robust and can be used for the explanation of differences between two groups of spectra. However, the considerable disadvantage of the quantile method is the difficult interpretation of value because the normalization process changes the intensities of the bin to the rank.

Thus, we suggest using creatinine normalization procedures or the PQN method on the data. However, compared to creatinine normalization PQN is more universal method, since it can be applied even in other biofluids (blood plasma and serum) or in metabolomic studies where the creatinine production is impaired.

Acknowledgement

The authors gratefully acknowledge the project LO1509 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] ALONSO A., MARSAL S., and JULIÀ A.. Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology*, 3:23, 2015.

[2] BUGÁŇOVÁ, M., PELANTOVÁ, H., HOLUBOVÁ, M., ŠEDIVÁ, B., MALETÍNSKÁ, L., ŽELEZNÁ, B., KUNEŠ, J., KAČER, P., KUZMA, M., HALUZÍK, M. The effects of liraglutide in mice with diet-induced obesity studied by metabolomics. *Journal of Endocrinology*, 2017, 233:1, pp. 93-104. ISSN: 0022-0795

[3] DIETERLE F., ROSS A., SCHLOTTERBECK G., SENN, H.: Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry* 78 (13), pp. 4281–4290, 2006.

[4] DI GUIDA R., ENGEL J., ALLWOOD J.W., WEBER R.J.M., JONES M.R., SOMMER U., VIANT M.R., and DUNN W.B.. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, 12(5):93, Apr 2016.

[5] EUCEDA L.R., GISKEØDEGÅRD G.F., and BATHEN F.T.. Preprocessing of NMR metabolomics data. *Scandinavian Journal of Clinical and Laboratory Investigation*, 75(3):193–203, 2015. PMID: 25738209.

[6] GARDLO A., SMILDE A.K., HRON K., HRDÁ M., KARLÍKOVÁ R., FRIEDECKÝ D., and ADAM T.. Normalization techniques for PARAFAC modeling of urine metabolomic data. *Metabolomics*, 12(7):117, Jun 2016.

[7] HERTEL J., VAN DER AUWERA S., FRIEDRICH N., WITTFELD K., PIETZNER M., BUDDE K., TEUMER A., KOCHER T., NAUCK M., and GRABE H.J. Two statistical criteria

to choose the method for dilution correction in metabolomic urine measurements. *Metabolomics*, 13(4):42, Feb 2017.

[8] KOHL S.K., KLEIN M.S., HOCHREIN J., OEFNER P.J., SPANG R., and GRONWALD W.. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8(1):pp. 146–160, Jun 2012.

[9] PELANTOVÁ, H., BUGÁŇOVÁ, M., HOLUBOVÁ, M., ŠEDIVÁ, B., ZEMENOVÁ, J., SÝKORA, D., KAVÁLKOVÁ, P., HALUZÍK, M., ŽELEZNÁ, B., MALETÍNSKÁ, L., KUNEŠ, J., KUZMA, M. Urinary metabolomic profiling in mice with diet-induced obesity and type 2 diabetes mellitus after treatment with metformin and vildagliptin and their combination. *Molecular and Cellular Endocrinology*, 2016, 431:15 August 2016, pp. 88-100. ISSN: 0303-7207

[10] SACCENTI E., HOEFSLOOT H.C.J., SMILDE A.K., WESTERHUIS J.A., and HENDRIKS M.M.W. B. . Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 10(3): pp.361–374, Jun 2014.

[11] XIA J., SINELNIKOV I.V., HAN B., and WISHART D.S.. Metaboanalyst 3.0—making metabolomics more meaningful. *Nucleic Acids Research*, 43(W1):W251–W257, 2015.

Current address

Šedivá Blanka, RNDr., Ph.D.
Institute of Microbiology of the Czech Academy of Sciences
Vídeňská 1083, 142 20, Prague 4, Czech Republic
E-mail: blanka.sediva@gmail.com

Pelantová Helena, RNDr., Ph.D.
Institute of Microbiology of the Czech Academy of Sciences
Vídeňská 1083, 142 20, Prague 4, Czech Republic
E-mail: pelantova@biomed.cas.cz

Bugáňová Martina, Mgr.
Institute of Microbiology of the Czech Academy of Sciences
Vídeňská 1083, 142 20, Prague 4, Czech Republic
E-mail: martina.buganova@gmail.com