Review

Data Normalization in NMR-based Metabolomics

Helena U. Zacharias[1], Michael Altenbuchinger[2] and Wolfram Gronwald[3]*

[1] Institute of Computational Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany; helena.zacharias@helmholtz-muenchen.de

[2] Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, Am Biopark 9, 93053 Regensburg, Germany; michael.altenbuchinger@ukr.de

[3] Institute of Functional Genomics, University of Regensburg, Am Biopark 9, 93053 Regensburg, Germany; wolfram.gronwald@ukr.de

* Correspondence: wolfram.gronwald@ukr.de.: +49-941-943-5015

Abstract: The aim of this article is to summarize recent bioinformatic and statistical developments applicable to NMR-based metabolomics. Extracting relevant information from large multivariate datasets by statistical data analysis strategies may be of considerable complexity. Typical tasks comprise for example classification of specimens, identification of differentially produced metabolites, and estimation of fold changes. In this context it is of prime importance to minimize contributions from unwanted biases and experimental variance prior to these analyses. This is the goal of data normalization. Therefore, special emphasize is given to different data normalization strategies. In the first part, we will discuss the requirements and the pros and cons for a variety of commonly applied strategies. In the second part, we will concentrate on possible solutions in case that the requirements for the standard strategies are not fulfilled. In the last part, very recent developments will be discussed that allow reliable estimation of metabolic signatures for sample classification without prior data normalization. In this contribution special emphasis will be given to techniques that have worked well in our hands.

Keywords: data normalization; data scaling; zero-sum; metabolic fingerprinting; NMR; statistical data analysis

1. Introduction

In general, the aim of metabolomics is the comprehensive study of the flow of small organic compounds through bioenergetic and biosynthetic pathways. These compounds, so-called metabolites, are qualitatively and quantitatively analyzed in cells, tissues, organs, body fluids, and whole organisms. Typical metabolites are amino acids, sugars, organic acids, bases, lipids, vitamins, and various conjugates of absorbed substances of exogenous origin. Metabolomics finds widespread application including such diverse topics as the screening of milk of dairy cows milk [1] or the investigation of acute kidney injury following heart surgery [2]. Metabolomic investigations are mainly conducted employing hyphenated mass spectrometry or nuclear magnetic resonance (NMR) spectroscopy. Here, we will focus mainly on the application of solution NMR spectroscopy to biological fluids as well as tissue and cell extracts in an academic setting, although many of the described approaches are not limited to these examples. Following metabolomic measurements, statistical data analysis is often key to extract meaningful information from the generally high-dimensional data. Here, proper pre-processing of the data is one of the most important steps influencing all downstream analyses. Therefore, the main focus of this review will be on data preprocessing and its effects on statistical analysis.

2. Methods

2.1 NMR Measurements

The use of automated data acquisition schemes is recommended to ensure high reproducibility of measurements by avoiding biases from human interference. Prior to measurement, the temperature unit of the spectrometer should be carefully calibrated employing, for example, a deuterated methanol sample. Typically, a sample temperature between 298 and 300 K is used. When possible, samples should be automatically locked, tuned, matched, and shimmed. Automatic shimming procedures should start from a standard shim file optimized for the current sample matrix. Also calibration of pulse lengths should be done in an automated fashion followed by automated data acquisition. The most widely used experiments in NMR-based metabolomics are 1D ^{1}H measurements [3], although 2D measurements are also applicable [4], especially when schemes for fast data acquisition like non-uniform sampling are applied [5]. Mostly, 1D NOESY pulse sequences with presaturation during relaxation and mixing time together with additional spoil gradients for water suppression are used. If biofluid specimens such as plasma or serum contain large amounts of macromolecules, they should be either ultra filtrated prior to measurement [6] or CPMG-type pulse sequences are recommended for the attenuation of macromolecular signals [7, 8].

2.2 Data Analysis

2.2.1 Spectra Processing and Data Preprocessing

To reduce biases due to human interference automated routines for Fourier transform and baseline and phase correction should be used whenever possible.

The subsequent statistical data analyses can be either based on targeted data where certain pre-selected metabolites were quantified, or on metabolic fingerprinting data. For this review we will mainly focus on the latter, which take the whole spectral information into account. Generally, they have to be corrected for variations in signal position due to differences in pH, salt, or temperature. A widely used and robust method to at least partially compensate for these effects is spectral binning, where spectra are split into a number bins. Equal-spaced bins/buckets are frequently used, although other schemes such as adaptive binning or spectral alignment are also employed. Data points inside every bucket are summed up or integrated. The whole spectrum is then represented as a vector of bucket integrals. Alternatively, peak alignment approaches such as icoshift may be used [9].

Resulting data are usually analyzed employing multivariate statistics [10, 11] that exploit the joint distribution of the data including the variance of individual features and their joint covariance structure. However, metabolomic datasets are prone to unwanted experimental and/or biological variances and biases. To minimize these contributions, data scaling and variance adjustment approaches may be used. The aim of scaling approaches is to minimize unwanted sample to sample variations. They involve, for instance, scaling relative to the signal of creatinine for urinary data, exemplified in Figure 1, scaling of every spectrum to a total sum of one, or probabilistic quotient normalization (PQN) [12] to name only a few. Previously, we systematically analyzed a variety of different scaling methods [13]. In this context, it is important to note that all of these different strategies have their pros and cons. For example, choosing creatinine as a standard for urine assumes the absence of interindividual differences in production and renal excretion of creatinine [14]. However, creatinine production and excretion depend on sex, age, muscle mass, diet, pregnancy as well as renal pathology of the examined individual [15, 16]. Normalization to a constant total spectral area or spectral mapping to a reference spectrum assumes that only a relatively small amount of metabolites is regulated in approximately equal shares up and down while all others remain more or less constant. However, this prerequisite is often not fulfilled, especially in case of kidney diseases where generally higher blood metabolite levels in diseased than in healthy patients are observed [2]. In this case, it is possible to learn the scaling parameters on a subset of non-regulated features only [17]. However, it is important to note that all of the different scaling strategies impact the following analysis steps such as screening for differential metabolites or multivariate metabolic signatures [18–21].

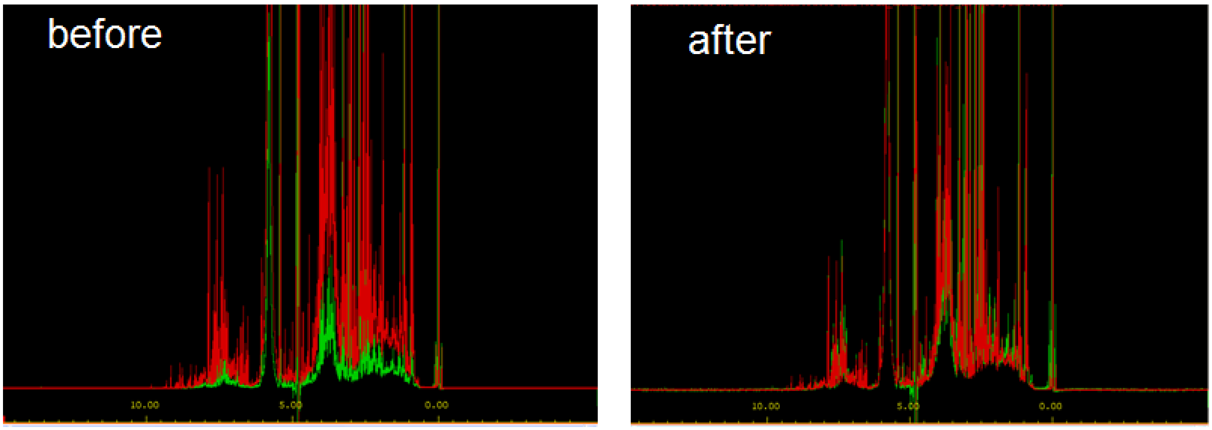

Figure 1 Scaling of two different urine spectra with respect to creatinine. (Left) before scaling and (right) after scaling.

Following data scaling it is often necessary to adjust the variance of the data. The variance of non-induced biological variation often correlates with the corresponding mean abundance of metabolites leading to considerable heteroscedasticity in the data, which will also impact subsequent data analysis. Approaches to reduce heteroscedasticity include simple logarithmic transformation or variance stabilization normalization [22, 23]. In our experience especially variance stabilization performed quite well in this regard [13].

A convenient way of performing the above scaling and variance adjustment approaches as well as subsequent multivariate data analysis is facilitated by the statistical programming environment *R* [24, 25]. Other common tools to perform these tasks or parts thereof include, for example, the numerical programming environment MATLAB (The MathWorks Inc., Natick, USA), the online server MetaboAnalyst [25], and the data analysis software SIMCA (Umetrics, Umeå, Sweden).

In case of NMR fingerprints, typically several hundred features are extracted from a single NMR experiment. In contrast, often only a relatively small number of different experiments are available to analyze the high-dimensional data space, rendering proper statistical analysis and visualization a challenging task. In the following sections we will summarize typical approaches that are used in this regard. They may be separated in unsupervised and supervised methods.

2.2.2 Unsupervised Data Analysis

In unsupervised methods, no information about underlying groups is used. Therefore, group separations observed are purely data-driven. Unsupervised algorithms are often employed initially to check for group separation prior to classification of data or in cases where too few samples are available for classification with rigid cross-validation. Commonly used methods include clustering approaches such as hierarchical clustering [26], non-hierarchical clustering employing the *k*-means method [27], and clustering by affinity propagation [28]. Principal component analysis (PCA) is a dimension reduction approach where new coordinate axes in the directions of maximal variances are drawn. It allows easy visualization of high-dimensional data. Closely related to PCA is independent component analysis (ICA) which has been shown to provide good results for metabolic data [29, 30]. Self organizing maps are a widely used method for two-dimensional data visualization [31].

2.2.3 Supervised Data Analysis

In supervised data analysis information about the class labels of the individual samples is included. One typical aim is to test whether a given hypothesis can be rejected or not. Here many different statistical tests are available from the literature, of which a Student's *t*-test [32] is probably the most common choice for detecting significant metabolic differences between two distinct sample groups. However, the results of this data analysis strategy exhibit a distinct dependency on the a priori chosen normalization method, as investigated by [18]. Figure 2 illustrates these observations for a *t*-test analysis of urinary NMR fingerprints of acute kidney injury (AKI) vs. healthy patients. The number as well as the identity of statistically significant NMR buckets strongly depends on the employed normalization strategy. This finding points to an inherent problem of standard statistical data analysis in metabolomics studies: the respective results are always dependent on the, often arbitrarily chosen, normalization strategy and findings can probably only be reproduced if the initial choice of normalization is used.

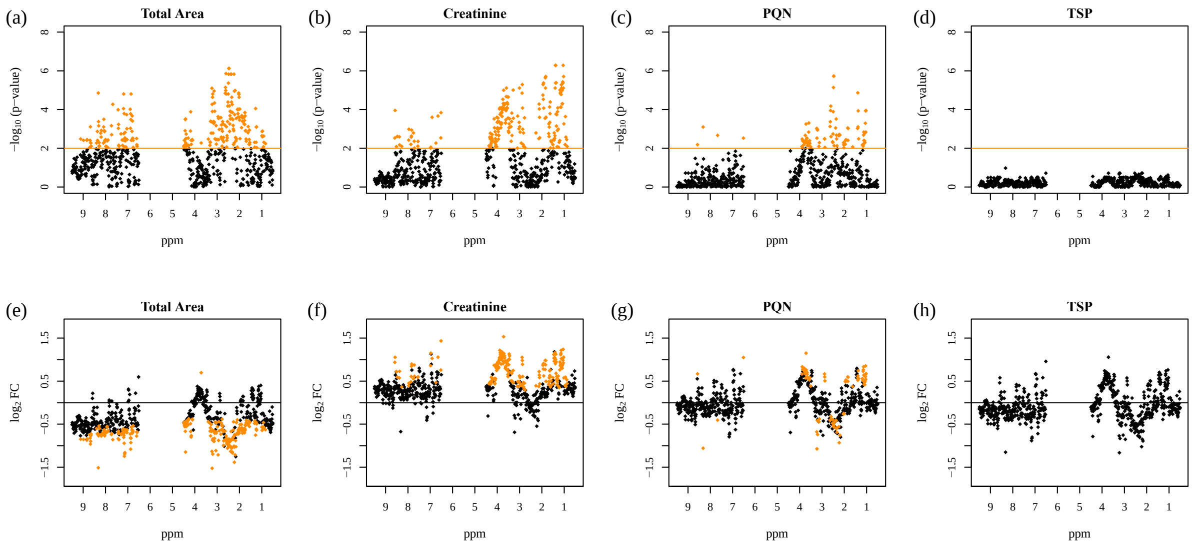

Fig. 2 Test for differentially regulated metabolites in 1D ^{1}H urinary NMR fingerprints between acute kidney injury (AKI) and healthy patients with respect to different normalization strategies. $-\text{Log}_{10}$(*p*-values) of moderated *t*-test analysis are shown after preprocessing with four different normalization methods: scaling to (a) equal total spectral area, (b) scaling to creatinine, (c) PQN, and (d) scaling to the internal reference TSP, plotted versus the ppm regions of the corresponding NMR buckets (upper panels). The significance level for Benjamini–Hochberg (B/H) adjusted *p*-values below 0.01, corresponding to a false discovery rate (FDR) below 1%, is marked by an orange line and the significant NMR features are indicated as orange diamonds. The corresponding log$_2$ fold changes (log$_2$ FC) plotted versus the ppm regions are shown in the lower panels (e–h). Since log$_2$ FCs were calculated as AKI minus non-AKI, positive log$_2$ FCs correspond to higher values in AKI than in non-AKI samples. Figure adapted from (Zacharias et al. 2017).

For the remaining of this review we will concentrate on another common application of supervised data analysis, namely sample classification.

2.2.3.1 Sample Classification

Classification of an unknown sample to two or more known phenotypic classes (e.g. healthy and diseased) is a common task for which generally techniques from machine learning are employed. Here, algorithms are trained on a training dataset where a class label for each sample is known, followed by an application of the trained algorithm on new independent test data. For performance evaluation, class labels of the test data also have to be known. In case that these independent test data are difficult to obtain, cross-validation may be used for performance evaluation. Here, the complete data set is iteratively split into training and test data. In case that additional parameters relevant for feature selection and the classification algorithm need to be optimized, nested cross-validation schemes are recommended [33] where these parameters are optimized in the inner loops of these schemes.

In NMR-based metabolomics approaches based on Partial Least Squares-Discriminant Analysis (PLS-DA) [34], often in the combination with orthogonal projection to latent structures (OPLS-DA) [35], are frequently employed. Other methods well established, for example, in gene expression analysis such as Random Forests (RF) [36] and Support Vector Machines (SVM) [37] are less frequently used. As a consequence, Hochrein et al., recently performed a systematic comparison of selected algorithms on typical NMR data [38].

It was found that RFs are particularly well suited for the analysis of high-dimensional NMR data. In short, a RF classifier is built from a set of tree predictors, where each tree is created from a different bootstrap sample of the training data. At each node of the tree the splitting in branches is based on a random selection of the input features. The final class label given to a sample is based on a majority vote over all trees. Additionally, RFs provide different measures of variable importance, which can be used for the identification of predictive features [39, 40].

Also SVMs showed good performance on typical NMR derived metabolic datasets [38]. Here a separating hyperplane is determined to maximize the distance between the individual classes of the training data. The hyperplane is built in a high-dimensional vector space defined by the individual feature levels.

Both classifiers are for example included in the MetaboAnalyst [25] software package and corresponding libraries are available for *R*. Classification approaches are often combined with data-driven variable selection approaches to improve classification performance and to reduce the risk of overfitting. This strategy is different compared to an a priori pre-selection of metabolites [41]. Feature selection approaches are normally grouped into three classes: wrapper, filter, and embedded methods, respectively [42]. The feature selection approach of the Elastic Net [43] classification algorithm, which can be viewed as a combination of classification by RIDGE [44] and LASSO [45] regression, is an example for the class of embedded feature selection methods. In contrast, the feature selection used by the Nearest Shrunken Centroids classification approach [47] is a representative of the wrapper methods. Among the large variety of available filter algorithms a *t*-score-based feature filtering performed well for high-dimensional NMR data [38]. However, it was recently demonstrated on two NMR-derived metabolic datasets that the set of features selected for classification critically depends on the choice of the data scaling approach [18]. Figure 3a and b illustrate the effect of different normalization methods on classification results for urinary NMR fingerprints. Both the performance and the derived metabolic signatures for (a) a SVM in combination with *t*-test based feature filtering and (b) a LASSO classification strongly depend on the a priori chosen normalization strategy. As a consequence reproducibility of metabolic studies based on common normalization and classification approaches is possible only if the exact same scaling protocols are used. This greatly limits the applicability of metabolic signatures derived by standard statistical analysis approaches.

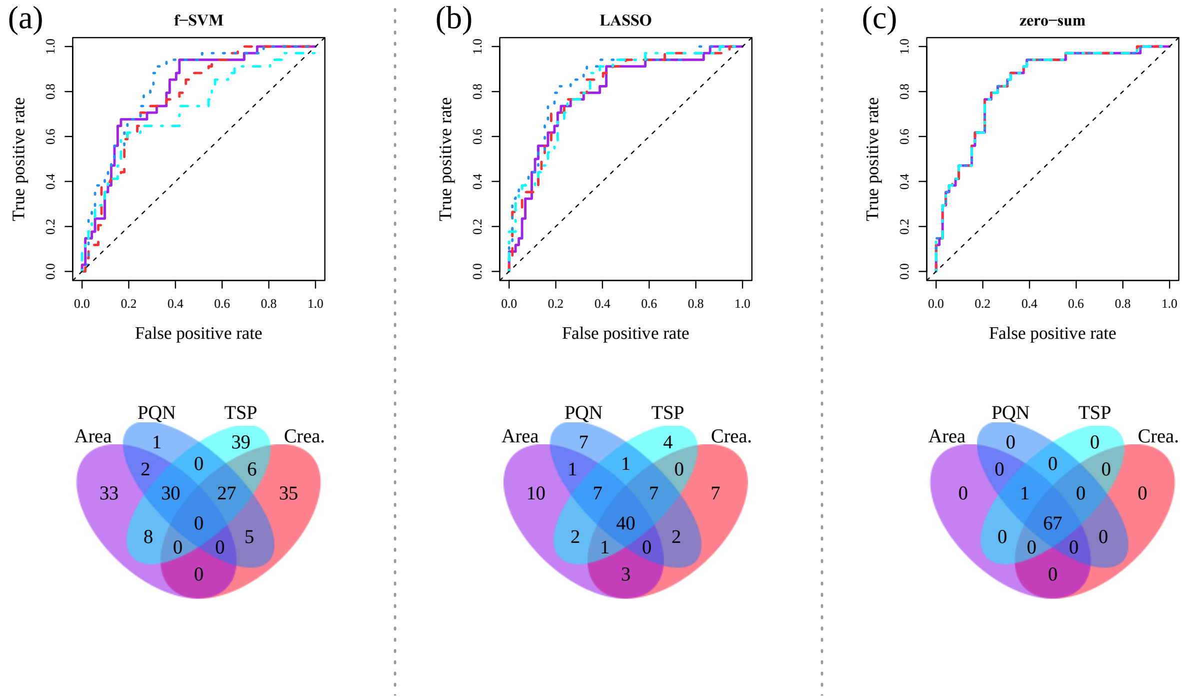

Fig. 3 Receiver operating characteristic (ROC) curves as well as Venn diagrams of selected classification features for the discrimination of AKI from non-AKI patients based on urinary 1D ^{1}H NMR fingerprints. Four different normalization strategies were employed: scaling to total spectral area (violet solid line), scaling to creatinine (red dashed line), probabilistic quotient normalization (PQN) (blue dotted line), and scaling to the internal reference TSP (cyan dashed-dotted line). Common classification approaches such as (a) SVM in combination with *t*-test based feature filtering, and (b) LASSO regression show a clear dependence on the chosen normalization strategy, whereas (c) zero-sum regression is completely independent thereof. Figure adapted from (Zacharias et al. 2017).

To overcome these issues zero-sum regression [48, 49], which has recently been demonstrated to be invariant under any rescaling of data [49], has been extended to logistic zero-sum regression [18]. In contrast to commonly used approaches, logistic zero-sum regression always employs the same set of biomarkers for sample classification regardless of the chosen scaling method. Therefore, prior data normalization may be omitted completely. Logistic as well as linear zero-sum regression are available as an *R* package and as a high-performance computing software at https://github.com/rehbergT/zeroSum.

In brief, it is based on the following concept: We start with the binned fingerprinting data $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, where x_{ij} is the logarithm of the intensity of bin $j \in \{1, \ldots, p\}$ in spectrum $i \in \{1, \ldots, N\}$ and y_i is the corresponding (clinical) response of patient *i*. In standard regression analysis, prior data scaling to a common unit like the total spectral area is required. As the data are on a logarithmic scale, scaling to a common unit becomes a shifting of the binned value x_i by some spectrum-specific value γ_i.

Therefore, in case of normalized data the regression equation reads

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \left(x_{ij} + \gamma_i\right) + \epsilon_i. \quad (1)$$

Eq. 1 becomes independent of the normalization factor γ_i if and only if the regression coefficients β_j sum up to zero, i.e.

$$\sum_{j=1}^{p} \beta_j = 0. \quad (2)$$

As a consequence, the additional constraint that all regression coefficients have to sum up to zero is set in zero-sum regression. Zacharias et al. 2017 showed for two metabolomic datasets that the obtained biomarker signatures were indeed independent of any prior data scaling. Figure 3c illustrates these results for a urinary 1D ^{1}H NMR metabolic dataset.

3. Conclusions

In this review we mainly concentrated on data analysis of NMR-derived metabolic fingerprints. Special emphasis was given to the issue of data normalization and in this context to the novel logistic zero-sum regression method that is independent of prior data normalization and therefore, has the potential to greatly enhance the reproducibility of biomarker studies.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Klein, M.S.; Buttchereit, N.; Miemczyk, S.P.; Immervoll, A.K.; Louis, C.; Wiedemann, S.; Junge, W.; Thaller, G.; Oefner, P.J.; Gronwald, W. NMR metabolomic analysis of dairy cows reveals milk glycerophosphocholine to phosphocholine ratio as prognostic biomarker for risk of ketosis. *J.Proteome Res.* **2012**, *11*, 1373–1381.
2. Zacharias, H.U.; Hochrein, J.; Vogl, F.C.; Schley, G.; Mayer, F.; Jeleazcov, C.; Eckardt, K.-U.; Willam, C.; Oefner, P.J.; Gronwald, W. Identification of Plasma Metabolites Prognostic of Acute Kidney Injury after Cardiac Surgery with Cardiopulmonary Bypass. *J. Proteome Res* **2015**, *14*, 2897–2905.
3. Beckonert, O.; Keun, H.C.; Ebbels, T.M.D.; Bundy, J.; Holmes, E.; Lindon, J.C.; Nicholson, J.K. Metabolic Profiling, Metabolomic and Metabonomic Procedures for NMR Spectroscopy of Urine, Plasma, Serum and Tissue Extracts. *Nat. Protocols* **2007**, *2*, 2692–2702.
4. Gronwald, W.; Klein, M.S.; Kaspar, H.; Fagerer, S.; Nürnberger, N.; Dettmer, K.; Bertsch, T.; Oefner, P.J. Urinary Metabolite Quantification Employing 2D NMR Spectroscopy. *Anal. Chem.* **2008**, *80*, 9288–9297.
5. Schlippenbach, T.v.; Oefner, P.J.; Gronwald, W. Systematic Evaluation of Non-Uniform Sampling Parameters in the Targeted Analysis of Urine Metabolites by ^{1}H,^{1}H 2D NMR Spectroscopy. *Sci. Rep.* **2018**, *8*, 4249.
6. Klein, M.S.; Almstetter, M.; Schlamberger, G.; Nürnberger, N.; Dettmer, K.; Oefner, P.J.; Meyer, H.H.D.; Wiedemann, S.; Gronwald, W. Nuclear Magnetic and Mass Spectrometry-based Milk Metabolomics in Dairy Cows During Early and Late Lactation. *J. Dairy Sci.* **2010**, *93*, 1539–1550.

7. Wallmeier, J.; Samol, C.; Ellmann, L.; Zacharias, H.U.; Vogl, F.C.; Garcia, M.; Dettmer, K.; Oefner, P.J.; Gronwald, W. Quantification of Metabolites by NMR Spectroscopy in the Presence of Protein. *J. Proteome Res* **2017**, *16*, 1784-1796.
8. Meiboom, S.; Gill, D. Modified Spin Echo Method for Measuring Nuclear Relaxation Times. *Rev. Sci. Instr.* **1958**, *29*, 688–691.
9. Savorani, F.; Tomasi, G.; Engelsen, S.B. Icoshift: A versatile Tool for the Rapid Alignment of 1D NMR Spectra. *J. Magn. Reson.* **2010**, *202*, 190–202.
10. Wishart, D.S. Computational Approaches to Metabolomics. *Methods Mol.Biol.* **2010**, *593*, 283–313.
11. Zacharias, H.U.; Hochrein, J.; Klein, M.S.; Samol, C.; Oefner, P.J.; Gronwald, W. Current Experimental, Bioinformatic and Statistical Methods used in NMR Based Metabolomics. *Curr. Metabol.* **2013**, *1*, 253–268.
12. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application to ^{1}H NMR Metabolomics. *Anal. Chem.* **2006**, *78*, 4281–4290.
13. Kohl, S.M.; Klein, M.S.; Hochrein, J.; Oefner, P.J.; Spang, R.; Gronwald, W. State-of-the Art Data Normalization Methods Improve NMR-Based Metabolomic Analysis. *Metabolomics* **2012**, *8*, 146–160.
14. Waikar, S.S.; Sabbisetti, V.S.; Bonventre, J.V. Normalization of Urinary Biomarkers to Creatinine During Changes in Glomerular Filtration Rate. *Kidney Int.* **2010**, *78*, 486–494.
15. Curhan, G. Cystatin C: A Marker for Renal Function of Something More? *Clin. Chem.*, **2005**, *51*, 293–294.
16. Stevens, L.A.; Levey, A.S. Measured GFR as a confirmatory test for estimated GFR. *J. Am. Soc. Nephrol.* **2009**, *20*, 2305–2313.
17. Hochrein, J.; Zacharias, H.U.; Taruttis, F.; Samol, C.; Engelmann, J.C.; Spang, R.; Oefner, P.J.; Gronwald, W. Data Normalization of ^{1}H NMR Metabolite Fingerprinting Data Sets in the Presence of Unbalanced Metabolite Regulation. *J. Proteome Res.* **2015**, *14*, 3217–3228.
18. Zacharias, H.U.; Rehberg, T.; Mehrl, S.; Richtmann, D.; Wettig, T.; Oefner, P.J.; Spang, R.; Gronwald, W.; Altenbuchinger, M. Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints. *J. Proteome Res.* **2017**, *16*, 3596-3605.
19. Gromski, P.S.; Xu, Y.; Hollywood, K.A.; Turner, M.L.; Goodacre, R. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics* **2015**, *11*, 684–695.
20. Jauhiainen, A.; Madhu, B.; Narita, M.; Narita, M.; Griffiths, J.; Tavaré, S. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics* **2014**, *30*, 2155–2161.
21. Saccenti, E. Correlation Patterns in Experimental Data Are Affected by Normalization Procedures: Consequences for Data Analysis and Network Inference. *J. Proteome Res.* **2017**, *16*, 619–634.
22. Huber, W.; Heydebreck, A.V.; Sültmann, H.; Poustka, A.; Vingron, M. Variance Stabilisation Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics* **2002**, *18*, S96-S104.
23. Parsons, H.M.; Ludwig, C.; Günther, U.L.; Viant, M.R. Improved Classification Accuracy in 1- and 2-Dimensional NMR Metabolomics Data Using the Variance Stabilising Generalised Logarithm Transformation. *BMC-Bioinformatics* **2007**, *8*, 234.
24. Development Core Team, R. *R: A Language and Environment for Statistical Computing;* R Foundation for Statistical Computing: Vienna, Austria, 2009.
25. Xia, J.; Psychogios, N.; Young, N.; Wishart, D.S. MetaboAnalyst: a Web Server for Metabolomic Data Analysis and Interpretation. *Nucl. Acids Res.* **2009**, *37*, W652-W660.
26. Draisma, H.H.; Reijmers, T.H.; Meulman, J.J.; van der, G.J.; Hankemeier, T.; Boomsma, D.I. Hierarchical clustering analysis of blood plasma lipidomics profiles from mono- and dizygotic twin families. *Eur. J. Hum. Genet.* **2013**, *21*, 95–101.
27. Hartigan, J. *Clustering Algorithms;* John Wiley: New York, 1975.
28. Frey, B.J.; Dueck, D. Clustering by passing messages between data points. *Science* **2007**, *315*, 972–976.
29. Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite Fingerprinting: Detecting Biological Features by Independent Component Analysis. *Bioinformatics* **2004**, *20*, 2447–2454.
30. Klein, M.S.; Dorn, C.; Saugspier, M.; Hellerbrand, C.; Oefner, P.J.; Gronwald, W. Discrimination of Steatosis and NASH in Mice Using Nuclear Magnetic Resonance Spectroscopy. *Metabolomics* **2011**, *7*, 237–246.
31. Dow, L.K.; Sandeep, K.; Dow, E.R. Self-organizing Maps for the Analysis of NMR Spectra. *Biosilico* **2004**, *2*, 157–163.
32. Student. The Probable Error of a Mean. *Biometrika* **1908**, *6*, 1–25.

33. Varma, S.; Simon, R. Bias in Error Estimation when Using Cross-Validation for Model Selection. *BMC-Bioinformatics* **2006**, *7*, 91.
34. Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemometrics* **2003**, *17*, 166–173.
35. Trygg, J.; Wold, S. Orthogonal Projections to Latent Structures. *J. Chemometrics* **2002**, *16*, 119–128.
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
37. Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data. Min. Knowl. Disc.* **1998**, *2*, 121–167.
38. Hochrein, J.; Klein, M.S.; Zacharias, H.U.; Li, J.; Wijffels, G.; Schirra, H.J.; Spang, R.; Oefner, P.J.; Gronwald, W. Performance Evaluation of Algorithms for the Classification of Metabolic ^{1}H-NMR Fingerprints. *J. Proteome Res.* **2012**, *11*, 6242–6251.
39. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC-Bioinformatics* **2009**, *10*, 213.
40. Bryan, K.; Brennan, L.; Cunningham, P. MetaFIND: a feature analysis tool for metabolomics data. *BMC-Bioinformatics* **2008**, *9*, 470.
41. Eisner, R.; Stretch, C.; Eastman, T.; Xia, J.; Hau, D.; Damaraju, S.; Greiner, R.; Wishart, D.S.; Baracos, V.E. Learning to Predict Cancer-Associated Skeletal Muscle Wasting from ^{1}H-NMR Profiles of Urinary Metabolites. *Metabolomics* **2011**, *7*, 25–34.
42. Haury, A.C.; Gestraud, P.; Vert, J.P. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE* **2011**, *6*, e28210.
43. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. B* **2005**, *67*, 301–320.
44. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.
45. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. Roy. Stat. Soc. B* **1996**, *58*, 267–288.
46. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
47. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *P. Natl. Acad. Sci USA.* **2002**, *99*, 6567–6572.
48. Lin, W.; Shi, P.; Feng, R.; Li, H. Variable selection in regression with compositional covariates. *Biometrika* **2014**, *101*, 785–797.
49. Altenbuchinger, M.; Rehberg, T.; Zacharias, H.U.; Stämmler, F.; Dettmer, K.; Weber, D.; Hiergeist, A.; Gessner, A.; Holler, E.; Oefner, P.J.; *et al.* Reference point insensitive molecular data analysis. *Bioinformatics* **2017**, *33*, 219–226.