Comparison of Standardization Approaches Applied to Metabolomics Data

Qingxia Yang
School of pharmaceutical Sciences, Chongqing University
Chongqing, 401331, P. R. China
+86-(0)23-65678468
yangqx@cqu.edu.cn

Bo Li
School of pharmaceutical Sciences, Chongqing University
Chongqing, 401331, P. R. China
+86-(0)23-65678468
libcell@cqu.edu.cn

Feng Zhu*
School of pharmaceutical Sciences, Chongqing University
Chongqing, 401331, P. R. China
+86-(0)23-65678468
zhufeng@cqu.edu.cn

ABSTRACT
Some factors such as unwanted variations might affect the identification of biomarkers in metabolomics and proteomics analysis, which needs preprocessing including normalization (also named as standardization) by the standardization approach prior to marker selection. Many standardization approaches were applied to analysis of the metabolomics, and even proteomics data. But there are rarely comprehensive comparison of the standardization performance based on the sample size and various methods. The current study performed an overall comparison aiming at these methods based on a metabolomics dataset. As a result, 15 standardization approaches were classified into four groups according to the standardization performances of different sample sizes. The Log Transformation and the VSN method were regarded as the Superior performance methods, but the Contrast method was performed consistently worst in all datasets of various sample size. This study could provide a useful guidance for the choice of befitting standardization approaches when carrying out the metabolomics and proteomics analysis based on LC/MS.

CCS Concepts
• Applied computing → Metabolomics / metabonomics.

Keywords
Metabolomics analysis; preprocessing; normalization methods; standardization approach; support vector machine (SVM).

1. INTRODUCTION
The metabolomics and proteomics based on mass spectrometry (MS) or Nuclear Magnetic Resonance (NMR) is able to monitor hundreds of thousands of metabolites or proteins in bio-fluid and tissue [1]. In particular, metabolomics and proteomics analysis are helpful for the identification of therapy [1], discovery of new drugs [2], and the choice of biomarkers for all kinds of diseases.

Some factors such as unwanted variations and technical errors might affect the identification of differential features in metabolomics and proteomics analysis. To remove unwanted variations, many measures are adopted including the signal drift correction, the batch effect removal and scaling using quality control samples or internal standards [3]. The data-driven strategies are better for untargeted metabolomics data [4]. The performance of 11 data-driven normalization approaches for processing NMR based metabolomics data were compared [5]. The Quantile and the Cubic Splines were identified as the good performed methods, while the Contrast and the Li-Wong could not reduce bias at all between samples. A research on the performances of 8 normalization approaches was conducted for gas chromatography mass spectrometry (GC/MS) based metabolomics, which discovered good performance of the Auto Scaling and the Range Scaling [6]. The LC/MS is widespread for metabolomics analysis like NMR and GC/MS, and it is very important to analyze the difference of standardization for those methods on LC/MS based metabolomics and proteomics data. Ejigu *et al*. have compared the performance of 6 methods based on "average metabolite specific coefficient of variation (CV)" for metabolomics based on LC/MS, but there is no evident difference for the CVs of those methods [4]. It's very important that the performance of those standardization methods popularly adopted in the metabolomics and proteomics analysis has not been exhaustively compared.

In this study, there was a comprehensive comparison on the performance of 15 standardization approaches. 10 sub-datasets of different sample size were used to evaluate the performance of 15 standardization approaches, which were classified into four groups (the superior performance group, the consistently good performance group, the moderate performance group and the poor performance group). Above all, this study would provide valuable guidance to the choice of fitted standardization approaches in analyzing the metabolomics data based on LC/MS. To certain degree, this research also provided some clues to evaluation of standardization performance LC/MS-based proteomics.

2. MATERIALS AND METHODS
The benchmark dataset analyzed in this study was the dataset MTBLS28 which was collected from the MetaboLights (http://www.ebi.ac.uk/metabolights/) [7]. The positive ionization mode (ESI+) of MTBLS28 provided LC/MS based metabolomics profiles of 1,005 samples (469 lung cancer patients and 536 healthy individuals) [8]. The samples were further divided into training group (900 samples including 400 lung cancer patients and 500 healthy individuals) and independent testing group (105 samples) by random sampling.

Moreover, the *k*-means clustering was used to generate 10 datasets of various sample size as the sub-datasets. In each sub-dataset, the number of samples including lung cancer patients and healthy individuals were from 90 to 900 representing from 10% to 100% of the training datasets, respectively. The detailed workflow of

ICBCB '17, January 06-08, 2017, Hong Kong, Hong Kong

DOI: http://dx.doi.org/10.1145/3035012.3035023

data sampling was illustrated in Figure 1. In this study, a widely adopted data pre-processing procedure was applied, including peak detection, retention time correction and peak alignment using *XCMS* package, signal correction by QC samples [9] and imputation of missing signals by the *KNN* algorithm.

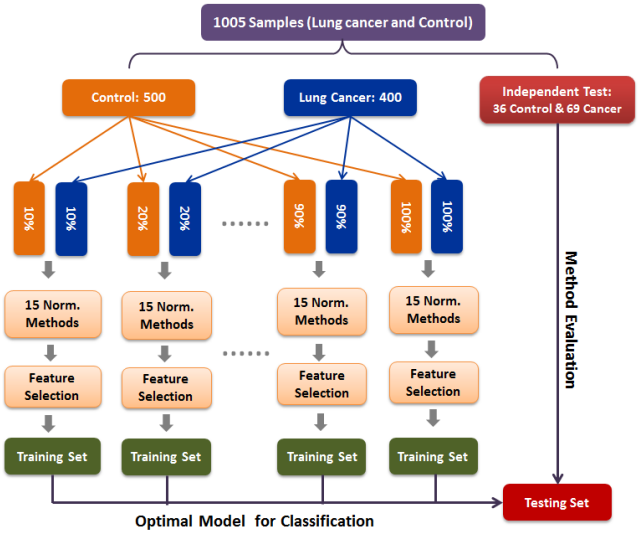

Figure 1. The overall research design in this study.

There were 15 methods included in this work and they are the Auto Scaling [10], the Contrast Normalization [11], the Cubic Splines [12], the Cyclic Loess [13], the Level Scaling [6], the Linear Baseline Scaling [11], the Log Transformation [14], the Non-Linear Baseline Normalization (Li-Wong) [15], the Pareto Scaling [16], the Power Scaling [17], the Probabilistic Quotient Normalization (PQN) [18], the Quantile Normalization [11], the Range Scaling [19], the Variance Stabilization Normalization (VSN) [20, 21] and the Vast Scaling [22].

The differential features were identified by VIP value (VIP > 1) of the partial least squares discriminant analysis (PLS-DA) in R package *ropls* [23] incorporating with p-value ($p < 0.05$) of Student *t*-test. And the SVM classification model in the R package *e1071* (http://cran.r-project.org/web/packages/e1071) was used to compare the standardization performance of 15 methods. The independent samples was used to assess the classification performance of the selected differential features by the area under the curve (AUC) of receiver operating characteristic (ROC) generated by R package *ROCR*. What's more, the hierarchical clustering was adopted to identify the similarity of performances for 15 methods among different sample sizes. All computational processing above was conducted in R (http://www.r-project.org) version 3.2.4.

3. RESULTS AND DISCUSSION

Cluster analysis of 15 standardization methods was conducted based on the benchmark datasets to assess the classification performance of the selected differential metabolites after standardization by different method. This procedure contained three detailed steps. Firstly, the AUC values of every standardization method among 10 sub-datasets of various sample size were used to generate a 10 dimensional vector. Secondly, the relationship of performance between any two methods was measured by the correlation of the 10 dimensional vectors of every method. Thirdly, hierarchical clustering was applied to investigate the relationship among methods. As an assessment of consistency between different distance metrics, two metrics including the *Euclidean* and the *Manhattan* and *Ward's* method were applied in the cluster analysis. The results of the clusters were shown in Figure 2.

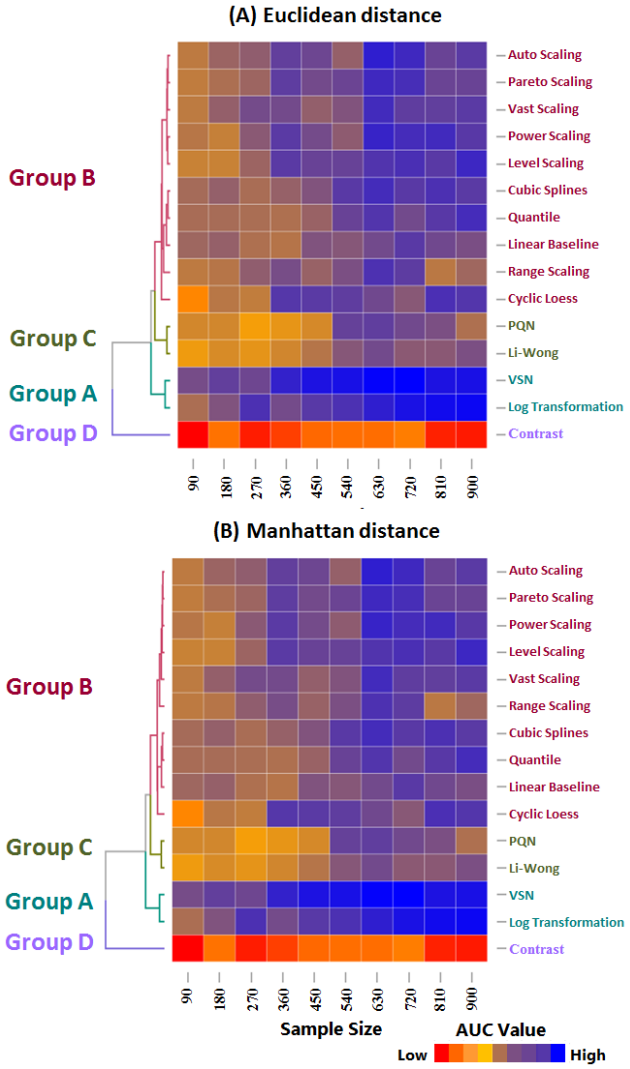

Figure 2. Cluster analysis of 15 methods based on their AUCs across 10 various sample sizes. (A) Hierarchical clustering with Euclidean metric. (B) Hierarchical clustering with Manhattan metric.

As illustrated by the clusters in Figure 2A and Figure 2B, the clustering results of both distance metrics were consistent with each other. The sample sizes of sub-datasets in the low AUC values with a yellow were 90, 180, and 270, and the sample sizes of sub-datasets in the high AUC values with a blue were ≥ 360. This result showed that the training datasets of large sample size (≥ 360) tended to have higher standardization performance. In particular, 15 standardization approaches were grouped by both metrics into four clusters (Group A, B, C and D) with slight variation in the dendrograms of clustering. Clearly, two methods (the VSN and the Log Transformation) were consistently performed best, while one method (the Contrast) always performed worst in the 10 sub-datasets.

As shown in Figure 3 (A), standardization approaches (the VSN and the Log Transformation) generated the best performance in comparison to the other 13 methods, which made group A (the Superior Performance Group). The VSN had been reported as a well-performed approach in metabolomics for the biological analysis [18, 24]. The Log Transformation was a powerful method for skewed distributions symmetric [6]. A lot of researches of metabolomics analysis adopted the standardization approaches to confirm above conclusion. For example, the Log Transformation was used in the pre-processed program for biomarker discovery of lung cancer [25]. And VSN is a hybrid between object-wise normalization and a scaling procedure to retain the consistency, for which performs both reduces the variation of samples and adjusts the variance of different metabolites [24]. The Contrast was the only one method in group D (the Poor Performance Group) in Figure 3 (D), the performance of which was consistently worst in 10 sub-datasets among all 15 methods. As reported by Kohl et al, the Contrast hardly reduced bias at all and could not improve comparability among samples [5].

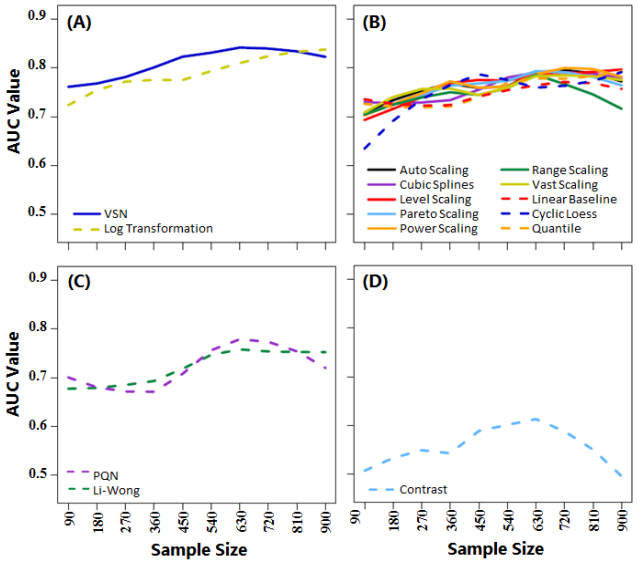

Figure 3. Four groups of methods identified in this study based on their normalization performances across various sample sizes. (A) Superior performance group; (B) Consistently good performance group; (C) Moderate performance group and (D) Poor performance group. All lines were generated by the LOESS regression.

The majority of methods (10 out of 15) were classified into Group B (Consistently Good Performance Group) in Figure 3 (B), which showed good standardization performances followed a similar fluctuation trends across 10 sub-datasets of various sample size. The 10 methods slightly underperformed comparing to Group A. And Group C consisted of two methods (the PQN and the Li-Wong) as shown in Figure 3 (C). Both methods demonstrated relatively good standardization performances with large variation, which made cluster C (the Moderate Performance Group). In the meantime, the Cyclic Loess was found to be very sensitive to the sample size of training dataset. Totally, this research could provide a useful suggestion to choose the suitable standardization approaches when researchers analyze the Omics data based on LC/MS. However, given the moderate differences observed between standardization approaches, the future work could apply a variety of datasets to aim for a more general recommendation.

4. CONCLUSION

This study conducted an overall comparison on 15 standardization approaches currently used for processing the LC/MS based metabolomics, and even proteomics data, and the dependence of their performance on the sample size of the training dataset was also evaluated. As a result, those 15 methods were classified into four groups according to their standardization performances in different sample sizes using the AUC values of ROC curves. The Log Transformation and the VSN were identified as methods of the superior performance group, while the performance of the Contrast method was consistently the worst across 10 sub-datasets among all 15 methods. In sum, this work could be regarded as a useful guidance to the choice of suitable standardization approaches in analyzing the LC/MS based Omics data, especially metabolomics data.

5. ACKNOWLEDGMENTS

This work was funded by the research support of National Natural Science Foundation of China (81202459 and 21505009); by the Fundamental Research Funds for the Central Universities (CDJZR14468801 and 2015CDJXY).

6. REFERENCES

[1] Weiss, R. H. and Kim, K. Metabolomics in the study of kidney diseases. *Nat. Rev. Nephrol.*, 8, 1 (Jan 2012), 22-33. DOI = http://dx.doi.org/10.1038/nrneph.2011.152.

[2] Kaddurah-Daouk, R. and Krishnan, K. R. Metabolomics: a global biochemical approach to the study of central nervous system diseases. *Neuropsychopharmacology*, 34, 1 (Jan 2009), 173-186. DOI = http://dx.doi.org/10.1038/npp.2008.174.

[3] De Livera, A. M., Sysi-Aho, M., Jacob, L., Gagnon-Bartsch, J. A., Castillo, S., Simpson, J. A. and Speed, T. P. Statistical methods for handling unwanted variation in metabolomics data. *Anal. Chem.*, 87, 7 (Apr 7 2015), 3606-3615. DOI = http://dx.doi.org/10.1021/ac502439y.

[4] Ejigu, B. A., Valkenborg, D., Baggerman, G., Vanaerschot, M., Witters, E., Dujardin, J. C., Burzykowski, T. and Berg, M. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *Omics : a journal of integrative biology*, 17, 9 (Sep 2013), 473-485. DOI = http://dx.doi.org/10.1089/omi.2013.0010.

[5] Kohl, S. M., 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,Suppl 1 (Jun 2012), 146-160. DOI = http://dx.doi.org/10.1007/s11306-011-0350-z.

[6] van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. and van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *Bmc Genomics*, 7(Jun 8 2006), 142. DOI = http://dx.doi.org/10.1186/1471-2164-7-142.

[7] Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., Gonzalez-Beltran, A., Sansone, S. A., Griffin, J. L. and Steinbeck, C. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.*, 41, Database issue (Jan 2013), D781-786. DOI = http://dx.doi.org/10.1093/nar/gks1004.

[8] Mathe, E. A., Patterson, A. D., Haznadar, M., Manna, S. K., Krausz, K. W., Bowman, E. D., Shields, P. G., Idle, J. R., Smith, P. B., Anami, K., Kazandjian, D. G., Hatzakis, E., Gonzalez, F. J. and Harris, C. C. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res.*, 74, 12 (Jun 15 2014), 3259-3270. DOI = http://dx.doi.org/10.1158/0008-5472.CAN-14-0109.

[9] Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J. D., Halsall, A., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Kell, D. B. and Goodacre, R. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.*, 6, 7 (Jun 30 2011), 1060-1083. DOI = http://dx.doi.org/10.1038/nprot.2011.335.

[10] Hu, C. and Xu, G. Mass-spectrometry-based metabolomics analysis for foodomics. *Trac-Trend Anal. Chem.*, 52, (Dec 2013), 36-46. DOI = http://dx.doi.org/10.1016/j.trac.2013.09.005.

[11] Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19, 2 (Jan 22 2003), 185-193. DOI = http://dx.doi.org/10.1093/bioinformatics/19.2.185.

[12] Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H. H., Nielsen, C., Brunak, S. and Knudsen, S. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.*, 3, 9 (Aug 30 2002), research0048.

[13] Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sinica*, 12, 1 (Jan 2002), 111-139.

[14] Purohit, P. V., Rocke, D. M., Viant, M. R. and Woodruff, D. L. Discrimination models using variance-stabilizing transformation of metabolomic NMR data. *Omics : a journal of integrative biology*, 8, 2 (Summer 2004), 118-130. DOI = http://dx.doi.org/10.1089/1536231041388348.

[15] Li, C. and Wong, W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 1 (Jan 2 2001), 31-36. DOI = http://dx.doi.org/10.1073/pnas.011404098.

[16] Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., Long, I., Lundstedt, T., Trygg, J. and Wold, S. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.*, 380, 3 (Oct 2004), 419-429. DOI = http://dx.doi.org/10.1007/s00216-004-2783-y.

[17] Brodsky, L., Moussaieff, A., Shahaf, N., Aharoni, A. and Rogachev, I. Evaluation of peak picking quality in LC−MS metabolomics data. *Anal. Chem.*, 82, 22 (Nov 15 2010), 9177-9187. DOI = http://dx.doi.org/10.1021/ac101216e.

[18] Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Anal. Chem.*, 78, 13 (Jul 1 2006), 4281-4290. DOI = http://dx.doi.org/10.1021/ac051632c.

[19] Smilde, A. K., van der Werf, M. J., Bijlsma, S., van der Werff-van der Vat, B. J. and Jellema, R. H. Fusion of mass spectrometry-based metabolomics data. *Anal. Chem.*, 77, 20 (Oct 15 2005), 6729-6736. DOI= http://dx.doi.org/10.1021/ac051080y.

[20] Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1 (2002), S96-104.

[21] Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S. V. and Lilley, K. S. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell Proteomics*, 9, 9 (Sep 2010), 1885-1897. DOI = http://dx.doi.org/10.1074/mcp.M900628-MCP200.

[22] Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E., Lindon, J. C. and Nicholson, J. K. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Anal Chim Acta*, 490, 1-2 (Aug 25 2003), 265-276. DOI = http://dx.doi.org/10.1016/S0003-2670(03)00094-1.

[23] Thévenot, E. A., Roux, A., Xu, Y., Ezan, E. and Junot, C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J. Proteome Res.*, 14, 8 (Aug 7 2015), 3322-3335. DOI = http://dx.doi.org/10.1021/acs.jproteome.5b00354.

[24] Hochrein, J., Zacharias, H. U., Taruttis, F., Samol, C., Engelmann, J. C., Spang, R., Oefner, P. J. and Gronwald, W. Data Normalization of (1)H NMR Metabolite Fingerprinting Data Sets in the Presence of Unbalanced Metabolite Regulation. *J. Proteome Res.*, 14, 8 (Aug 7 2015), 3217-3228.DOI = http://dx.doi.org/10.1021/acs.jproteome.5b00192.

[25] O'Shea, K., Cameron, S. J., Lewis, K. E., Lu, C. and Mur, L. A. Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study. Biochim Biophys Acta, 1860, 11 Pt B (Nov 2016), 2682-2687. DOI = http://dx.doi.org/10.1016/j.bbagen.2016.07.007.