Journal of proteome research

Technical Note

pubs.acs.org/jpr

Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets

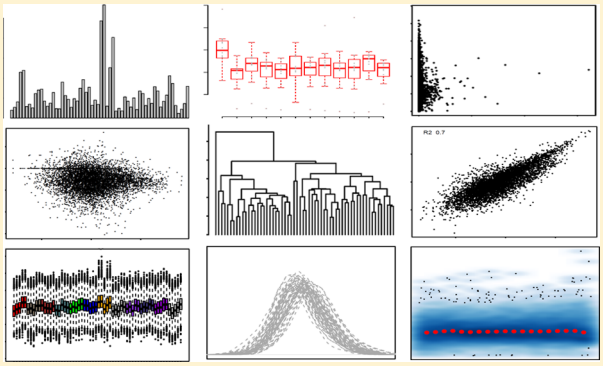
Aakash Chawade,[†] Erik Alexandersson,[‡] and Fredrik Levander*[,†]

[†]Department of Immunotechnology, Lund University, Medicon Village 406, SE 223 81 Lund, Sweden

[‡]Department of Plant Protection Biology, Swedish University of Agricultural Sciences, SE 230 53 Alnarp, Sweden

Supporting Information

ABSTRACT: High-throughput omics data often contain systematic biases introduced during various steps of sample processing and data generation. As the source of these biases is usually unknown, it is difficult to select an optimal normalization method for a given data set. To facilitate this process, we introduce the open-source tool "Normalyzer". It normalizes the data with 12 different normalization methods and generates a report with several quantitative and qualitative plots for comparative evaluation of different methods. The usefulness of Normalyzer is demonstrated with three different case studies from quantitative proteomics and transcriptomics. The results from these case studies show that the choice of normalization method strongly influences the outcome of downstream quantitative comparisons. Normalyzer is an R package and can be used locally or through the online implementation at http://quantitativeproteomics.org/normalyzer.



KEYWORDS: *normalization, preprocessing, label-free, mass spectrometry, microarray, proteomics, transcriptomics*

INTRODUCTION

High-throughput technologies such as DNA microarrays and mass spectrometry (MS) generate vast amount of information-rich transcriptomics, proteomics, and metabolomics data. These technologies have made significant progress in the past decade enabling detection and expression level quantification of thousands of genes, proteins, and metabolites in biological samples. Technical advancement of MS-based instruments in recent years has increased the detection accuracy and reduced the data generation time. This enables accurate detection and quantitative comparison of thousands of proteins from two to several samples at a time. However, high-throughput omics data often contain systematic biases introduced during various steps of sample processing and data generation. Failing to account for these biases could lead to misleading conclusions from quantitative analysis. Data normalization, if properly done, reduces systematic biases and is thus necessary prior to any downstream quantitative analysis. Different normalization methods address systematic biases in the data differently, and thus choosing an optimal normalization method for a given data set is critical. As the source of systematic bias in the data is usually unknown, an exhaustive comparative evaluation of both un-normalized data and the data normalized through different methods is required to select a suitable normalization method. For a detailed review on normalization of label-free proteomics data, refer to Karpievitch et al.[1]

Different normalization methods for omics data have been evaluated lately, and it is apparent that different methods produce considerably different results.[2−8] Callister et al. evaluated four different normalization methods for label-free proteomics data and concluded that methods based on linear regression were most optimal but suggested that further investigation is needed.[5] Kultima et al.[3] proposed a new normalization method, RegrRun, which performed best among 10 different methods on peptidomics data. Choe et al.[2] evaluated four different normalization methods for DNA-microarray data and concluded that the LOESS method is most optimal. Lyutvinskiy et al.[9] developed a normalization strategy for label-free proteomics data to account for fluctuations in the electrospray ionization in the time domain. Wang et al.[10] hypothesized that the missing values in the proteomics data set are non-random and proposed a two-step approach where data are first normalized by top 80 order statistics to estimate a scaling factor for each sample, followed by missing value imputation taking into account the scaling factor for each sample. Webb-Robertson et al.[8] proposed a statistical selection strategy called SPANS based on Rank Invariant Peptides and provided a tool to evaluate different peptide selection methods for normalization and subsequent normalization with emphasis on possible bias introduced by the normalization. It is thus apparent that suitability of a normalization method is dependent on the intrinsic characteristics of the data.

Received: December 18, 2013
Published: April 28, 2014

dx.doi.org/10.1021/pr401264n | J. Proteome Res. 2014, 13, 3114−3120

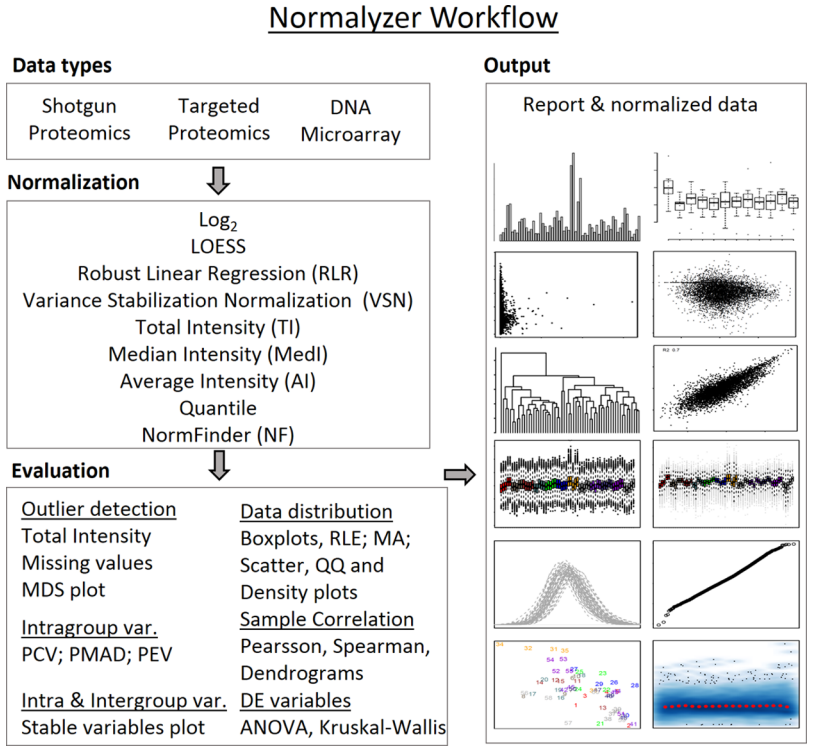


Figure 1. Normalyzer workflow highlighting types of input data, normalization, analysis methods, and final output.

Evaluation of data normalization can be done both quantitatively and qualitatively. Quantitative analysis is mainly based on the measure of dispersion around the mean within and between groups. The most common quantitative measures are standard deviation (SD), coefficient of variation (CV), median absolute deviation (MAD), and pooled estimate of variance (PEV). SD can be either positive or negative and is described relative to the sample mean, making it difficult to compare samples with differing mean. Measuring PEV could be an alternative for comparisons as it is always positive. CV measures variation as a percentage of mean and thus can be expressed independently of the mean, making it easier to compare variability between samples. However, CV is highly sensitive when the sample mean is close to zero as even low variation could produce high CV. Moreover, SD, PEV, and CV are sensitive to outliers. MAD measures the median of the absolute deviations around the sample median and thus is more robust and less sensitive to outliers. These methods were used previously for normalization evaluation of omics data.[3,5,7] Qualitative evaluation can be based on boxplots, MA plots, dendrograms, or correlation plots. Optimally, a normalization method for a given data set should be selected on the basis of both quantitative and qualitative evaluation measures and by further analysis of previously known housekeeping genes or proteins.

Here, we introduce Normalyzer, a new tool developed to evaluate the suitability of different normalization methods for a given data set based on commonly used quantitative and qualitative parameters. Normalyzer can be used for normalizing data from DNA microarrays, label-free proteomics, metabolomics, targeted mass spectrometry, or quantitative RT-PCR as long as the data are approximately normally distributed and are formatted as per the requirements. Normalyzer is fully automated and outputs normalized data from 12 different normalization methods along with an evaluation report. It is an open-source tool and can be run online with a user-friendly interface or can be installed locally as an R-package. Here, the usability of Normalyzer is demonstrated with three different case studies.

METHODS

Implementation

Normalyzer is implemented in R using Bioconductor[11] packages. The Normalyzer R-package can be downloaded from (http://quantitativeproteomics.org/normalyzer) and can be installed locally with R (version 3.0). Installation and usage instructions can be found at the above URL. An online service with a graphical user interface is also provided at the Web site.

Data Requirements

Normalyzer accepts data with raw intensities in a tab-separated format. The raw data should not be in logarithmic scale. Any number of rows and column annotations can be included if labeled accordingly. The data set should be relatively large, preferably at least a few hundred variables, and the observations need to contain replicate groupings to enable normalization evaluation. The data can be read in from a text file or as a data frame to facilitate inclusion of Normalyzer in existing pipelines.

A challenge with shotgun proteomics data is the occurrence of missing values due either to peptide quantities being below the detection limit or other technical issues. Imputation of missing values could in some cases lead to erroneous results and thus should be done with precaution.

Normalization Methods

Several popular normalization methods are included, such as total intensity (TI), median intensity (MedI), average intensity (AI), quantile (preprocessCore package),[12] NormFinder[13] (NF), Variance Stabilizing Normalization (VSN, vsn package),[14] Robust Linear Regression (RLR), and LOESS (limma

dx.doi.org/10.1021/pr401264n | J. Proteome Res. 2014, 13, 3114−3120

package).[15] These methods are implemented as global normalization methods (denoted by 'G'). Furthermore, VSN, LOESS, and RLR are also implemented as local methods (denoted by 'R') wherein the replicate groups are normalized separately. Due to computational reasons, NormFinder is automatically turned off for data sets where the number of variables with non-missing values is higher than 1000. Missing values (denoted NA) are tolerated differently by different normalization methods. Missing values are excluded during the $\log_2$ transformation, TI, MedI, AI, and RLR normalization; thus, NAs remain NAs even after normalization and only numerical data are normalized. For determining the control variables by NormFinder, only variables with no missing values are considered. For LOESS and VSN normalization, the data set is processed as-is and all warnings (if any) generated during LOESS and VSN normalization are saved to the warnings file. The data normalized by these methods are then evaluated both quantitatively and qualitatively.

Evaluation Measures

To aid in the selection of an optimal normalization method, different quantitative and qualitative statistical measures are considered. The results from these measures are plotted and saved to the report. Measures include total intensity, total missing values, Pooled intragroup Coefficient of Variation (PCV), Pooled intragroup Median Absolute Deviation (PMAD), Pooled intragroup estimate of variance (PEV), stable variables plot, CV-intensity plot, dendrograms, Pearson and Spearman correlation, MA-plots,[16] boxplots, density plots, Q-Q plots, Multidimensional scaling (MDS) plots, meanSD plot, and Relative Log Expression (RLE) plots as illustrated in Figure 1.

Case Studies

To evaluate the performance of Normalyzer, three different data sets with varying characteristics were selected. Case studies 1 and 2 contain benchmark data generated by spiked-in variables at varying concentrations, but with controlled background and negligible biological variation, whereas case study 3 contains experimental data with considerable biological variation.

Case Study 1: LC–MS/MS Proteomics Benchmark Data. A previously published shotgun proteomics data set[17] was used in this case study. The samples consist of 48 human proteins (UPS1, Sigma) spiked-in at five different known concentrations (0.25, 0.74, 2.2, 6.7, and 20 fmol/μL) in a standard yeast lysate. The raw data (OrbitrapO@65) were downloaded from the CPTAC data portal and were converted to mzML with MS Numpress compressed binaries (https://github.com/ms-numpress/ms-numpress) and MGF using Proteowizard.[18] The files were processed in the Proteios Software Environment (ProSE)[19] through a label-free quantitative workflow described previously.[20] MS/MS identification was performed in Mascot Server 2.4.1 (http://www.matrixscience.com) with a database consisting of *S. cerevisiae* proteins from SwissProt (downloaded 20 October 2009) and the protein sequences found in the Sigma UPS1 protein set, concatenated with an equal size decoy database. Match tolerances were 7 ppm for precursors and 0.5 Da for fragments. Carbamidomethylation of cysteine was used as fixed modification setting and oxidation of methionine as variable, and one missed cleavage was allowed. The resulting data set with raw intensities for 36,484 features was used for evaluation of normalization methods in Normalyzer.

Case Study 2: DNA Microarray Benchmark Data. A previously published benchmark data set with 3,860 spiked in cRNAs generated with Affymetrix GeneChips[2] was used to evaluate Normalyzer performance on array data. The samples consist of 1,309 cRNAs spiked in at differing concentrations between S and C samples and 2,551 cRNAs spiked in at identical relative concentrations. The S and C samples were hybridized in triplicate to Affymetrix GeneChips (six arrays). The raw data were downloaded and preprocessed by MAS5 in R/Bioconductor.[11,21] Filtering of probe sets to retain those with more than one present call in six samples resulted in a final data set with 4,156 probe sets that was used in Normalyzer.

Case Study 3: LC–MS/MS Proteomics Biological Data. Shotgun proteomics data generated from the secreted protein fraction of *P. infestans* infected leaves of three potato (*S. tuberosum*) cultivars from a previous study (Ali et al., submitted, ProteomeXchange DOI 10.6019/PXD000435) was used as the third case study. It consists of label-free quantitative mass spectrometry data with up to five replicates collected just before infection and at three different time points post-infection. Sample processing was conducted essentially as described previously,[22] and the data were processed as in Sandin et al.[20] with msInspect peptide feature detection.[23] The extracted and aligned features were used for the present study. Singly charged features and features with missing values in more than 40 samples were excluded. The data with raw intensities for 16,896 features from 60 samples were normalized in Normalyzer.

RESULTS AND DISCUSSION

The aim of Normalyzer is to aid in the selection of an optimal normalization method for a given data set based on quantitative and qualitative aspects of data variability. Normalyzer can be run both online with a Graphical User Interface or offline as an R package. Any type of omics data is supported as long as the basic data requirements are fulfilled. Normalyzer evaluates the suitability of 12 normalization methods for the uploaded data using quantitative and qualitative parameters (Figure 1). It should be noted that most normalization methods assume that the majority of variables are relatively stable between samples, and data that do not fulfill this requirement could be biased after global normalization. Therefore, methods are also implemented to normalize locally within replicate groups. These methods are denoted 'R' in the report, while methods denoted 'G' are global. However, for most data sets global normalization should be the first choice, since local normalization may skew group comparisons.

Output

The output from Normalyzer is a report with quantitative and qualitative evaluation measures of the normalization outcome. The total missing value plot and the total intensity plot summarize raw data characteristics and together with the MDS plot can be used to identify outlier samples due to sample degradation or other reasons. The PCV, PMAD, and PEV plots represent variability within replicates and help in the selection of normalization methods based on low intragroup variability. The variability within replicates suggests if the replicates are well correlated but fail to explore global alignment. In the stable variables plot, global variance of 5% of least DE variables are plotted against %PCV compared to $\log_2$. This plot helps in the exploration of both inter- and intragroup variance in the data, for detection of possible bias introduced during normalization, as normalization should not introduce variation in these

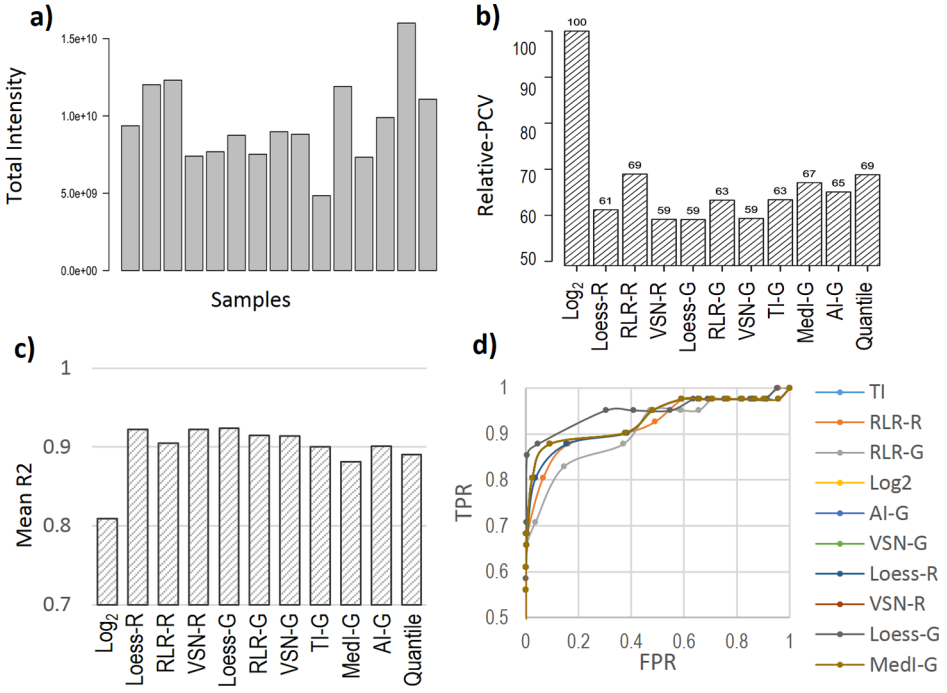


Figure 2. Case study 1. Benchmark data generated by shotgun proteomics. (a) Summed raw intensity from all peptides in each sample. (b) Relative pooled intragroup coefficient of variation (PCV). For percentage estimation, PCV in the un-normalized $\log_2$ transformed data is considered as 100%. (c) Mean R^2 values generated from observed and theoretical values for the UPS1 peptides in the dilution series. (d) Receiver operating characteristics (ROC) curves generated from the UPS1 proteins from differently normalized data sets with one-way ANOVA. UPS1 proteins were considered true positives, and the background proteins were considered true negatives.

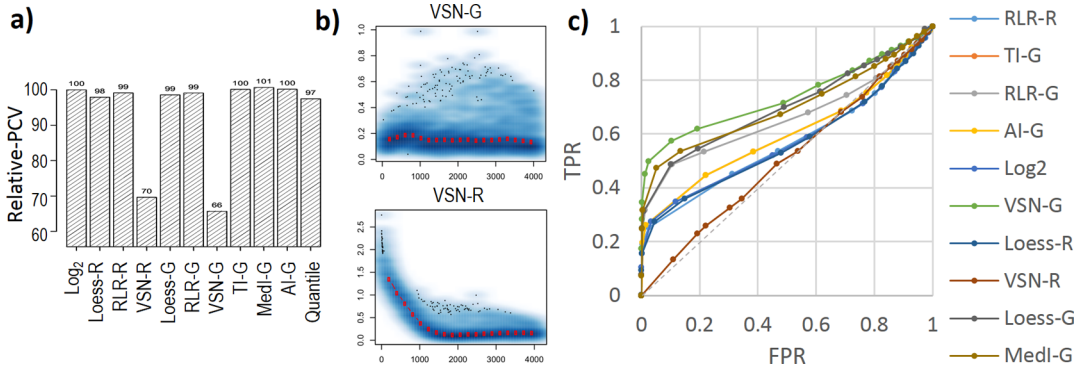


Figure 3. Case study 2. Benchmark data generated by Affymetrix microarray. (a) Percent PCV averaged over all groups. For percentage estimation, variability in un-normalized $\log_2$ transformed data is considered as 100%. (b) MeanSDplot of VSN-G and VSN-R normalized data. (c) ROC curves generated from the spiked-in probe sets from differently normalized data sets with one-way ANOVA.

variables.[8] Qualitative plots such as boxplots, MA plots, dendrograms, correlation plots, meanSD plot, MDS, and RLE plots explore data from all samples and guide in the method selection process. The data normalized by different methods are also exported together with the report and are ready for postnormalization analysis. Additional documentation including a flow chart for the decision-making process and a detailed explanation of various methods can be downloaded from the Normalyzer homepage.

Evaluation of Normalyzer

Features and analytical capabilities of Normalyzer were evaluated by three different case studies. Normalyzer reports from the three case studies are in the Supporting Information.

Case Study 1. The processed data set contains 36,484 features from a reconstituted yeast proteomics standard spiked-in with different levels of the Sigma UPS1 equimolar protein standard. From the Normalyzer report, it is apparent that there is an almost 3-fold difference in the total intensity between samples (Figure 2a). This suggests technical variation in the data set, and thus, normalization of the data set is necessary prior to quantitative analysis. The Normalyzer report showed that there was a decrease in PCV by 30−40% in the normalized data sets compared to un-normalized $\log_2$ transformed data (Figure 2b). Among the global normalization methods, relative-PCV was lowest (59%) in LOESS-G and VSN-G normalized data.

Out of 36,484 peptides, 304 peptides were from 43 Sigma UPS1 proteins. The UPS1 proteins were spiked in known absolute concentration in a dilution series (0.25, 0.74, 2.2, 6.7, and 20 fmol/μL). Thus, the spike-in protein set can be used for estimating the observed and theoretical correlation of $\log_2$ transformed peptide intensities. The mean coefficient of

dx.doi.org/10.1021/pr401264n | J. Proteome Res. 2014, 13, 3114−3120

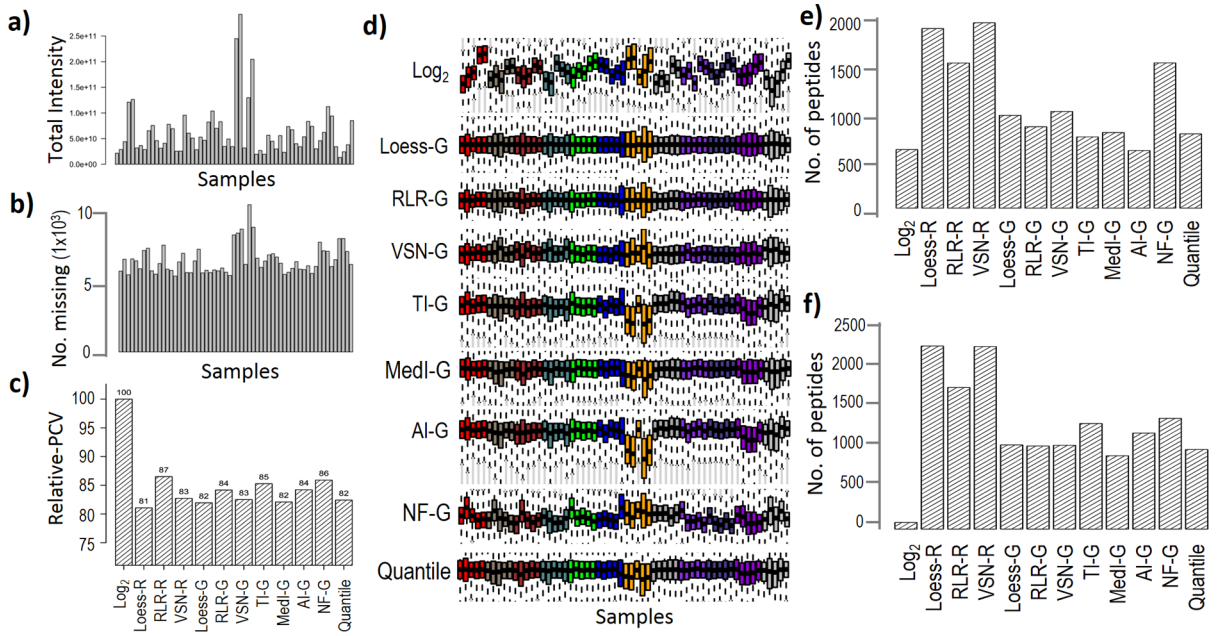


Figure 4. Case study 3. Biological data generated by shotgun proteomics from *P. infestans* infected potato leaves. (a) Summed raw intensity from all peptides in each samples. (b) Summed missing values in samples. (c) Relative PCV. (d) RLE plots for selected data sets. (e) One-way ANOVA (FDR < 0.05) and (f) Kruskal–Wallis test for statistical significance (FDR < 0.05).

determination (R^2) estimated from the un-normalized $\log_2$ transformed data and the theoretical values was 0.81, while the mean R^2 of LOESS-G, RLR-G, and VSN-G was >0.9 (Figure 2c). LOESS-G normalized data had the highest mean R^2 of 0.92, which supports the results from Normalyzer. Receiver operating characteristic (ROC) curves (Figure 2d) generated from the detected UPS1 proteins corroborate the suitability of LOESS-G normalization for this data set. From the results it is clear that all normalization methods performed better than just $\log_2$ transformation, and also that the choice of normalization method was of importance for this data set.

Case Study 2. To test the applicability of Normalyzer on DNA microarray data, a previously published Affymetrix microarray data set with 4,156 probe sets[2] was analyzed using Normalyzer. Results based on PCV suggested VSN-G and VSN-R as the most optimal methods for normalizing this data set (Figure 3a). However, the meanSD plot in the report showed that VSN-R normalized data contained bias introduced during the normalization step, leaving VSN-G as the most optimal normalization method (Figure 3b). As the data set was generated from spiked-in transcript levels, it was possible to calculate the ROC curve as an orthogonal evaluation of the normalization outcome (Figure 3c). The results from Normalyzer strongly support VSN-G normalization for this data set, and this is well in line with the orthogonal ROC calculations. Interestingly, in this analysis, the LOESS normalization method used in the original paper was not ranked the best, and this highlights the benefit of evaluating different normalization methods for any given data set.

Case Study 3. Finally, Normalyzer was used to select a normalization method for a shotgun proteomics data set with large intragroup variation in protein content and signal. Among the three case studies, this data set shows the highest variation in the total intensity within replicates (Figure 4a) and missing values (Figure 4b), indicating a clear need for normalization. Overall, the replicate samples in the normalized data sets had reduced variance compared to the $\log_2$ transformed data, and among the global normalization methods, LOESS-G, MedI-G, and Quantile normalized data had the least relative-PCV (Figure 4c). Further analysis of the RLE plots from the Normalyzer report indicate that samples in LOESS-G normalized data are centered better than the MedI-G and Quantile normalized data set (Figure 4d). Thus, for this data set, LOESS-G normalization could be an optimal normalization method. As there was no *a priori* information regarding expected sample protein content we evaluated the data set using standard statistical methods for quantitative comparisons. Both one-way ANOVA (Figure 4e) and Kruskal–Wallis test (Figure 4f) showed that LOESS-G normalized data contained a higher number of significantly differentially expressed peptides compared to un-normalized $\log_2$ transformed data. Indeed, the number of peptides passing the statistical tests as significantly regulated at a constant false discovery rate varied considerably between the normalization strategies. This highlights the need for selection of an appropriate normalization strategy, as downstream processing will be significantly affected by the choice.

■ CONCLUSION

In conclusion, effectiveness of normalization methods is dependent on the data, and extensive evaluation of different methods is necessary before choosing a method. Normalyzer is developed to aid in this selection process. The Normalyzer report is designed to help users narrow down the normalization methods. As seen in case study 2, normalization methods could sometimes be prone to overfitting, introducing additional bias to the data. Thus, while evaluating normalization methods, equal importance should be given to quantitative and qualitative plots and also to the existing knowledge on housekeeping genes or protein expression levels. As the tool is open-source, new normalization methods can be added-in and can be modified further for compatibility with existing pipelines. It can also be run in parallel with the SPANS[8] tool to further evaluate peptide selection for normalization.

dx.doi.org/10.1021/pr401264n | J. Proteome Res. 2014, 13, 3114–3120

While the present version of Normalyzer incorporates normalization methods for log-normally distributed data, the framework can readily be extended with other normalization methods that are better suited for count data from RNaseq experiments. We thus believe that Normalyzer will guide researchers in selecting the most appropriate normalization method for their omics data sets.

■ ASSOCIATED CONTENT

Supporting Information

Normalyzer reports from case studies 1−3. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +46-46-222 3835. Fax: +46-46-222 4200. E-mail: fredrik.levander@immun.lth.se.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Data used in case study 1 in this publication were created by the Clinical Proteomics Tumor Analysis Consortium (NCI/NIH). This work was supported by the Mistra Biotech research program and the Swedish Foundation for Strategic Research (RBb08-0006). Support by BILS (Bioinformatics Infrastructure for Life Sciences) is gratefully acknowledged.

■ REFERENCES

(1) Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D. Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinf.* **2012**, *13* (Suppl 16), S5.

(2) Choe, S. E.; Boutros, M.; Michelson, A. M.; Church, G. M.; Halfon, M. S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.* **2005**, *6* (2), R16.

(3) Kultima, K.; Nilsson, A.; Scholz, B.; Rossbach, U. L.; Falth, M.; Andren, P. E. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol. Cell Proteomics* **2009**, *8* (10), 2285−95.

(4) Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.* **2006**, *78* (7), 2262−7.

(5) Callister, S. J.; Barry, R. C.; Adkins, J. N.; Johnson, E. T.; Qian, W. J.; Webb-Robertson, B. J.; Smith, R. D.; Lipton, M. S. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J. Proteome Res.* **2006**, *5* (2), 277−86.

(6) Karpievitch, Y. V.; Taverner, T.; Adkins, J. N.; Callister, S. J.; Anderson, G. A.; Smith, R. D.; Dabney, A. R. Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics* **2009**, *25* (19), 2573−80.

(7) Deo, A.; Carlsson, J.; LindlÖF, A. How to choose a normalization strategy for Mirna Quantitative Real-Time (Qpcr) arrays. *J. Bioinf. Comput. Biol.* **2011**, *09* (06), 795−812.

(8) Webb-Robertson, B. J.; Matzke, M. M.; Jacobs, J. M.; Pounds, J. G.; Waters, K. M. A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors. *Proteomics* **2011**, *11* (24), 4736−41.

(9) Lyutvinskiy, Y.; Yang, H.; Rutishauser, D.; Zubarev, R. A. In silico instrumental response correction improves precision of label-free proteomics and accuracy of proteomics-based predictive models. *Mol. Cell Proteomics* **2013**, *12* (8), 2324−2331.

(10) Wang, P.; Tang, H.; Zhang, H.; Whiteaker, J.; Paulovich, A. G.; McIntosh, M. Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pac. Symp. Biocomput.* **2006**, 315−26.

(11) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y.; Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5* (10), R80.

(12) Bolstad, B. *preprocessCore: A collection of pre-processing functions*, R package version 1.20.0; http://www.bioconductor.org/packages/release/bioc/html/preprocessCore.html.

(13) Andersen, C. L.; Jensen, J. L.; Orntoft, T. F. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* **2004**, *64* (15), 5245−50.

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

(15) Smyth, G. K., Limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*; Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., Dudoit, S., Eds.; Springer: New York, 2005.

(16) Dudoit, S.; Yang, Y. H.; Callow, M. J.; Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* **2002**, *12* (1), 111−139.

(17) Paulovich, A. G.; Billheimer, D.; Ham, A. J.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell Proteomics* **2010**, *9* (2), 242−54.

(18) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918−20.

(19) Häkkinen, J.; Vincic, G.; Månsson, O.; Wårell, K.; Levander, F. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.* **2009**, *8* (6), 3037−43.

(20) Sandin, M.; Ali, A.; Hansson, K.; Månsson, O.; Andreasson, E.; Resjö, S.; Levander, F. An adaptive alignment algorithm for quality-controlled label-free LC-MS. *Mol. Cell Proteomics* **2013**, *12* (5), 1407−1420.

(21) Gautier, L.; Cope, L.; Bolstad, B. M.; Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **2004**, *20* (3), 307−15.

(22) Ali, A.; Moushib, L. I.; Lenman, M.; Levander, F.; Olsson, K.; Carlson-Nilson, U.; Zoteyeva, N.; Liljeroth, E.; Andreasson, E. Paranoid potato: phytophthora-resistant genotype shows constitutively activated defense. *Plant Signaling Behav.* **2012**, *7* (3), 400−8.

(23) Bellew, M.; Coram, M.; Fitzgibbon, M.; Igra, M.; Randolph, T.; Wang, P.; May, D.; Eng, J.; Fang, R.; Lin, C.; Chen, J.; Goodlett, D.; Whiteaker, J.; Paulovich, A.; McIntosh, M. A suite of algorithms for the

comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **2006**, *22* (15), 1902−9.