Normalization of large-scale mass spectrometry-based metabolic profiling experiments

Bedilu Ejigu [1,2], Dirk Valkenborg [1,2,3], Maya Berg [4], Jean-Claude Dujardin [4], Tomasz Burzykowski [1]

[1] I-BioStat, Hasselt University, Diepenbeek, Belgium, [2] Flemish Institute for Technological Research, VITO, Mol, Belgium, [3] Center for Proteomics, University of Antwerp, Antwerp, Belgium,
[4] Unit of Molecular Parasitology, Department of Biomedical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

1. Introduction

• To compare data from LC-MS experiments in a label-free quantitative setting, one needs to minimize non-biological differences that affect the measured intensity levels.

• Normalization is the process of removing undesirable systematic variations.

•In this study, we evaluate the performance of several normalization techniques that were developed for microarray data when applied to MS data.

2. Data

• First dataset (Figure 1, left panel) is composed of LC-MS runs of a standard sample containing 28 modified amino acids measured over three different time blocks, i.e., July, September, and October.

• Second dataset (Figure 1, right panel) is a sample of the Leishmania parasite BPK282/0 clone 4, which was repeatedly measured in two different time periods, i.e., July and September

• Clear running time (month) effect in both datasets.

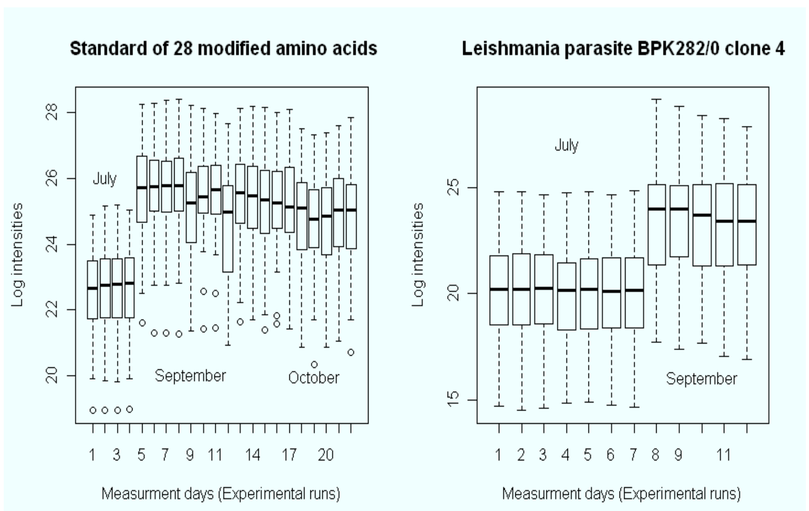

Figure 1. Box-whisker plots for the log-intensity before normalization: observed mean and variance are different across measurement blocks

3. Normalization Methods

Global normalization

• Uses a constant adjustment factor to remove the between-experiment intensity scale differences.

• Unsuitable if the differences are intensity-dependent.

Linear baseline normalization (Bolstad et al 2003):

• The baseline is constructed by calculating the median intensity for each amino acid/metabolite over all runs.

• The run-specific scaling factor is the ratio of the mean baseline intensity to the mean intensity.

Quantile normalization (Bolstad et al 2003):

• The main aim is to make the distribution of measured intensities in a set of runs the same.

Cubic splines (Workman et al. 2002, Kohl et al 2011):

• As in quantile normalization, the goal is to obtain a similar distribution across runs.

• Baseline run is built by computing the geometric mean of the intensities of each metabolites over all runs.

• For normalization, cubic splines regression is performed on the log(ratio) – average log(intensity) scatter plot between each run and a reference run.

Probabilistic quotient normalization (Dieterle et al 2006):

• The quotients of all metabolites in a run to the reference metabolite (median) are calculated.

•The scaling factor is the median of the quotients.

Cyclic loess normalization (Cleveland & Devlin 1988, Dudoit et al 2002):

• All pairs of runs are considered.

• Intensity-adjustment obtained by subtracting the normalization curve (loess) from the original values.

4. Results

• Evaluation of the normalization techniques based on descriptive statistics for the distribution of the original and transformed data of the two datasets.

• Successful normalization should reduce the between-run variability, as compared to the original data.

4.1 Standard amino acids

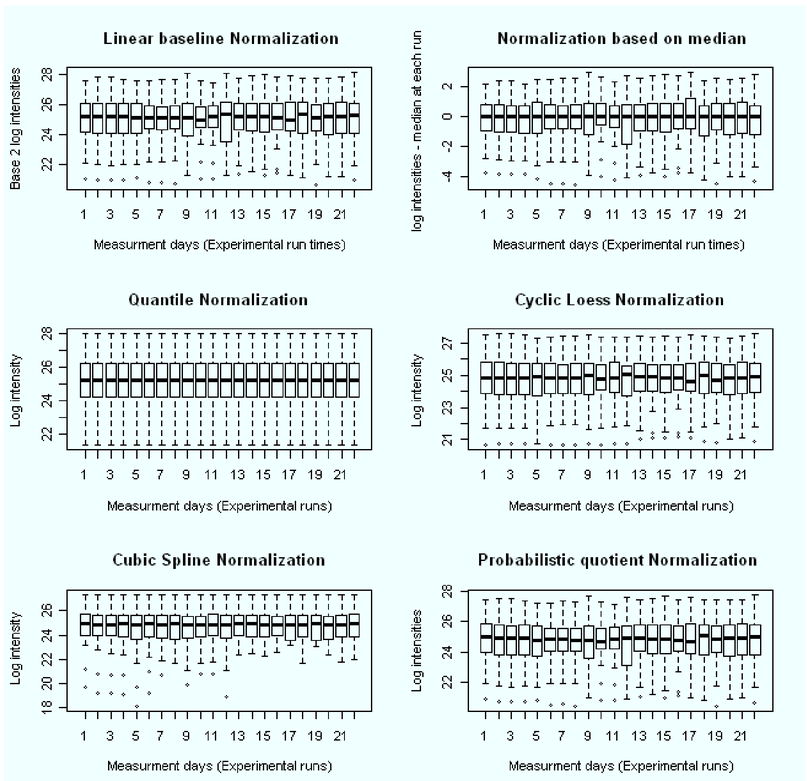

Figure 2. Box-whisker plots for the log-intensity after different normalizations

• Normalization (Figure 2) removes the month-effect seen in the original data (Figure 1, left panel),

• Mean intensity similar across different runs.

• For the quantile normalization, the distribution of the normalized intensity is identical across runs.

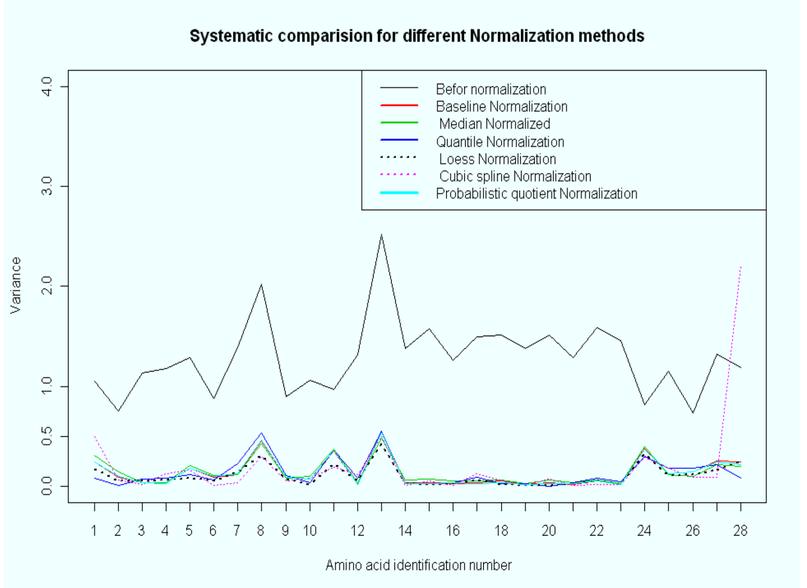

Figure 3. Line plot for the log-intensity variance for different amino acids across runs before and after normalization.

• All normalization methods reduce the variance of the intensities for all amino acids (Figure 3).

4.2 Leishmania sample

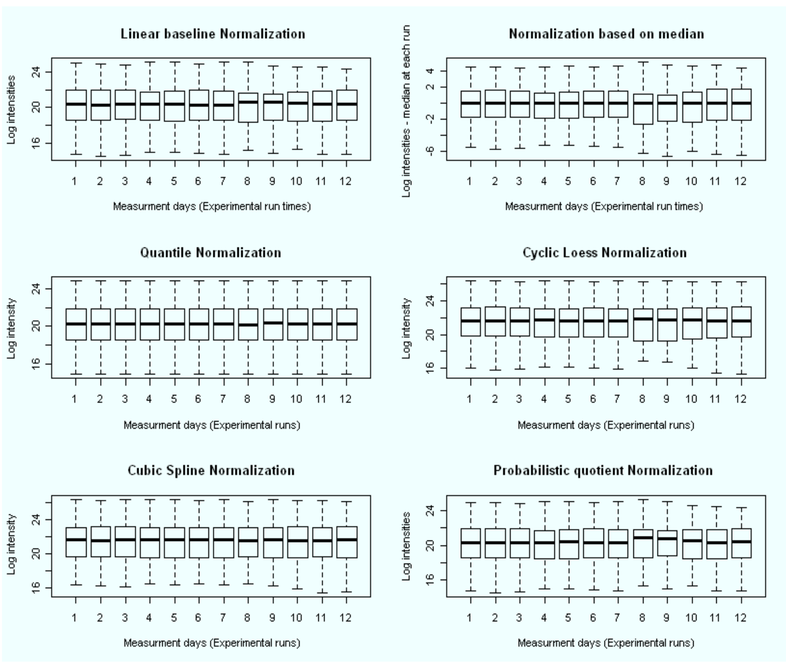

Figure 4. Box-whisker plots for the log-intensity after different normalizations

• Normalization (Figure 4) removes the month effect seen in the original data (Figure 1, right panel).

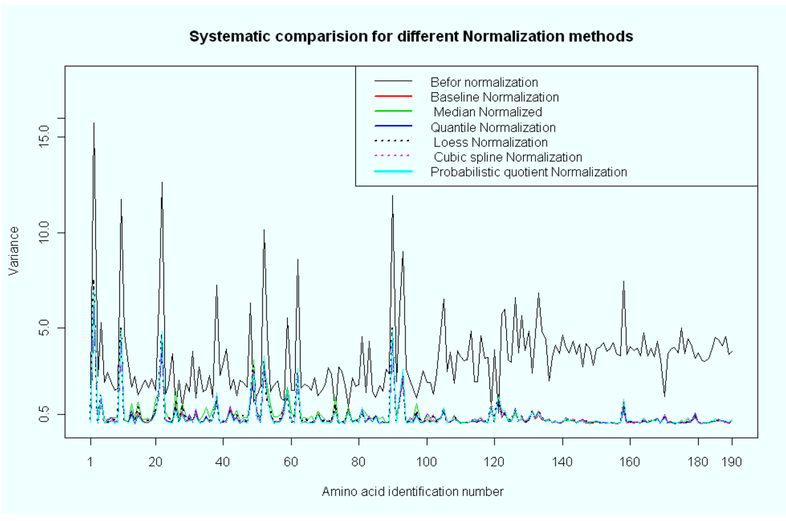

Figure 5. Line plot for the log-intensity variance for different amino acids across runs before and after normalization.

• All normalization methods reduce the variance of the intensities for all amino acids (Figure 5).

Conclusions

• Normalization reduces the between-run variability.

•The difference between different normalization methods is small. No single method performs uniformly best in both datasets.

•Different methods perform better in the different datasets.

References

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics,19, 185–193.

Cleveland, W. S., and Devlin, S. J. (1988). Locally weighted regression: An approach to regression-analysis by local fitting. Journal of the American Statistical Association,83, 596–610.

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dillution of complex biological mixtures. Application to 1H NMR metabolomics. Analytical Chemistry, 78, 4281–4290.

Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12, 111–139.

Kohl,S.M., Klein,M.S., Hochrein,J., Oefner,P.J, Spang,R. and Gronwald,W. (2011). State-of-the art data normalization methods improve NMR-based metabolomic analysis. Metabolomics,DOI 10.1007/s11306-011-0350-z.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., et al. (2002). A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Bioogy, 3, research0048.

VITO NV
Boeretang 200 – 2400 MOL – BELGIUM – Tel. + 32 14 33 55 11 – Fax + 32 14 33 55 99 – vito@vito.be – www.vito.be