Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Proteomics

Phosphonormalizer: an R package for normalization of MS-based label-free phosphoproteomics

Sohrab Saraei[1,*], Tomi Suomi[1,2], Otto Kauko[1] and Laura L. Elo[1,*]

[1]Turku Centre for Biotechnology, University of Turku and Åbo Akademi, FI-20520 Turku, Finland

[2]Department of Future Technologies, University of Turku, FI-20014 Turku, Finland

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Global centering-based normalization is a commonly-used normalization approach in mass spectrometry-based label-free proteomics. It scales the peptide abundances to have the same median intensities, based on an assumption that the majority of abundances remain the same across the samples. However, especially in phosphoproteomics, this assumption can introduce bias, as the samples are enriched during sample preparation which can mask the underlying biological changes. To address this possible bias, phosphopeptides quantified in both enriched and non-enriched samples can be used to calculate factors that mitigate the bias.

Results: We present an R package phosphonormalizer for normalizing enriched samples in label-free mass spectrometry-based phosphoproteomics.

Availability: The phosphonormalizer package is freely-available under GPL (>=2) license from Bioconductor (https://bioconductor.org/packages/phosphonormalizer).

Contact: sohrab.saraei@utu.fi, laura.elo@utu.fi

1 Introduction

Protein phosphorylation is the most common post-translational modification of proteins and it plays an important role in signal transduction as well as in regulation of enzymatic activity, protein-protein interactions, and protein stability. Systematic characterization of phosphoproteome with mass spectrometry (MS) based techniques could improve the understanding of various biological processes and identify novel therapeutic and diagnostic targets (Pawson and Scott, 2005).

Normalization is an important step when analyzing phosphoproteomics data. Global-centering methods, such as median normalization, are used frequently in label-free MS-based proteomics and have shown reasonable performance (Välikangas *et al.*, 2016). However, they assume that most of the peptide abundances do not change between samples, which does not hold true in all phosphoproteomics experiments (Kauko *et al.*, 2015; Olsen *et al.*, 2010). Similar behavior has been seen also in other omics data (Lovén *et al.*, 2012). The fundamental problem with phosphoproteomics is that the enrichment of phosphopeptides during sample preparation can mask large unidirectional biological changes in the abundances. More specifically, applying global-centering normalization in situations, where majority of measured abundances are affected, risks introducing bias in distribution of fold changes of phosphopeptides across samples. To remove this bias, a novel approach called **pairwise normalization** has been suggested (Kauko *et al.*, 2015). With this method, the phosphopeptides that are identified and quantified in both enriched and non-enriched MS runs are used to calculate an additional normalization factor to be included in median normalization.

Here we introduce an R package called phosphonormalizer that implements the pairwise normalization to enable its wider use. To our knowledge, this is the first R package specifically focused on normalization of phosphoproteomics data. To evaluate the package, global phosphorylation changes in HeLa cells were investigated under three conditions in a manner similar to that carried out by Kauko et al (2015): the activation of protein phosphatase 2A (PP2A) by depleting cancerous inhibitor of PP2A (CIP2A), depleting RAS, or the inhibition of PP2A by

okadaic acid (OA) treatment. For the evaluation, abundances of five phosphopeptides were measured by means of western blotting and compared to the abundances quantified through MS with five different normalization methods: global centering, quantile centering, global pairwise, quantile pairwise and a method using spiked-in alpha-casein protein to normalize the data (Kauko *et al.*, 2015). The comparison confirmed that the pairwise method produced results that were most similar to the western blot validation in the data.

2 Implementation

We implemented the approach of Kauko et al. (2015) in the R programming environment. The package, named phosphonormalizer, is publicly available from Bioconductor. The phosphonormalizer package also supports integration into Bioconductor proteomics workflow by supporting the MSnSet data structure.

The normalization is performed in five steps as detailed below and illustrated in Figure 1A.

1. Median normalization is applied to both enriched and non-enriched data.
2. Overlap of phosphopeptides in enriched and non-enriched data is determined. Only identical peptides with the same phosphorylation site in enriched and non-enriched samples are considered. If there are no common phosphopeptides, normalization using this package is impossible. In this study, we estimated that pairwise normalization can have superior performance compared to other normalization methods, even when there are only fifteen phosphopeptides in the overlap (Supplementary Figures 1 and 2).
3. For each phosphopeptide in the overlap, its abundance in the non-enriched data is divided by its counterpart in the enriched data to calculate the peptide abundance ratios. Peptides, whose abundance ratios exhibited extreme sample to sample variation (maximum difference > 1.5 * IQR) are discarded. Disposing of these phosphopeptides ensures that the method is not sensitive to outliers.
4. In each sample, an additional normalization factor is calculated as the median of the log-transformed peptide abundance ratios.
5. Finally, these factors are applied to the corresponding enriched samples.

The fold changes reproduced from Kauko et al. (2015) using our phosphonormalizer package are represented as a heatmap in Figure 1B.

3 Conclusion

MS-based label-free phosphoproteomics makes possible the analysis of large-scale experiments in various treatment conditions. Despite this effective method, there is risk of introducing bias at different steps of the experiment. In phosphoproteomics, due to the dynamic nature of phosphorylation and the effects of enrichment during sample preparation, the bias can be a serious problem. Therefore, conclusions drawn from these experiments are highly dependent on accurate normalization that can deal with the bias. Here, we present an R package that incorporates non-enriched data as a reference for normalizing the enriched data and show its superior performance to other normalization methods tested. In theory, the package can also be applied to studies of other post-translational modifications than phosphorylation that are commonly studied using enrichment and that can also be detected in non-enriched samples with sensitive methods. However, further evaluation would be needed to assess its applicability in such studies.

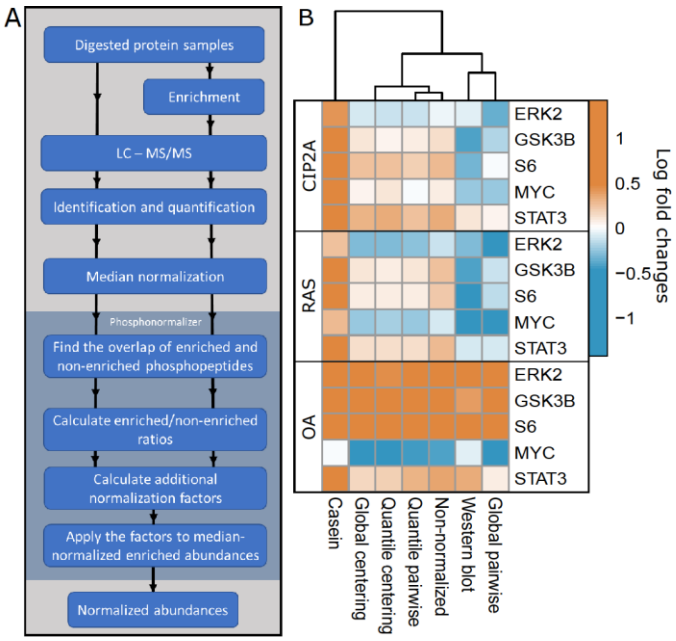

Figure 1. A) Schematic illustration of the phosphonormalizer workflow. **B)** Heatmap of fold changes of phosphoprotein abundances in different samples versus control (rows) after applying five different normalization methods (columns) quantified using LC-MS/MS data or western blotting.

Acknowledgements

The authors wish to thank Aidan J. McGlinchey for manuscript proofreading.

Funding

Dr. Elo reports grants from the European Research Council (ERC) (677943), European Union's Horizon 2020 research and innovation programme (675395), Academy of Finland (296801 and 304995), Juvenile Diabetes Research Foundation JDRF (2-2013-32), Tekes – the Finnish Funding Agency for Innovation (1877/31/2016) and Sigrid Juselius Foundation, during the conduct of the study.

References

Kauko,O. *et al.* (2015) Label-free quantitative phosphoproteomics with novel pairwise abundance normalization reveals synergistic RAS and CIP2A signaling. *Sci. Rep.*, **5**, 13099.

Lovén,J. *et al.* (2012) Revisiting global gene expression analysis. *Cell*, **151**, 476–482.

Olsen,J. V. *et al.* (2010) Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis. *Sci. Signal.*, **3**.

Pawson,T. and Scott,J.D. (2005) Protein phosphorylation in signaling – 50 years and counting. **30**, 286–290.

Välikangas,T. *et al.* (2016) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief. Bioinform.*, bbw095.