Evaluation of moFF and FlashLFQ for label free peptide quantification in proteomic workflows within Galaxy-P Framework.

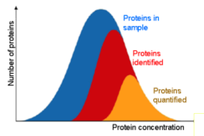
Subina Mehta[1], Caleb Easterly[1], James E.Johnson[2], Björn Grüning[3], Andrea Argentini[4-6], Robert J.Millikin[7], Michael R.Shortreed[7], Thomas McGowan[2], Praveen Kumar[9] , Lennart Martens[4-6], Lloyd M.Smith[7,8], Timothy J.Griffin[1] and Pratik Jagtap[1]

[1]Biochemistry, Molecular Biology, and Biophysics, University of Minnesota Twin Cities, Minneapolis; [2]Minnesota Supercomputing Institute, University of Minnesota Twin Cities, Minneapolis; [3]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Baden-Württemberg; [4]VIB-UGhent Center for Medical Biotechnology, VIB, Ghent, Belgium, Ghent, Belgium; [5]Bioinformatics Institute, Ghent, Belgium; [6]Department of Biochemistry, Ghent University, Belgium; [7]Department of Chemistry, University of Wisconsin, Madison, Wisconsin; [8]Genome Center of Wisconsin, University of Wisconsin, Madison, Wisconsin; [9]Bioinformatics and Computational Biology, University of Minnesota Twin Cities, Minneapolis

Introduction

- Mass Spectrometry (MS) is widely used in proteomics to estimate the absolute or relative protein expression levels within different biological conditions.
- Label-free quantification is attractive because it is relatively simple and economical compared to labeled quantification methods, but care must be take to ensure accuracy of estimates, including appropriate normalization.
- Label-free quantification based on precursor peak intensities is generally preferred over spectral counting methods because of its increased dynamic range.
- We evaluate open-source quantitation tools such as moFF and FlashLFQ inside and outside of the Galaxy platform and compare the outputs with MaxQuant, a widely used database search and quantification tool and to implement these tools within Galaxy.
- The Galaxy-P team has developed workflows for proteomics (identification of modified and unmodified peptides); proteogenomics (identification of variant peptides); and metaproteomics (proteomics applied to microbiota). In all of these studies, quantification is critical.

Objective



- ***Evaluate*** open-source label-free peptide quantitation tools.
- ***Compare*** normalization methods
- ***Integrate*** evaluated tools into Galaxy workflows to quantify identified peptides.

FLASHLFQ moFF MQ

The Galaxy Platform

A web-based bioinformatics data analysis platform.

Features:

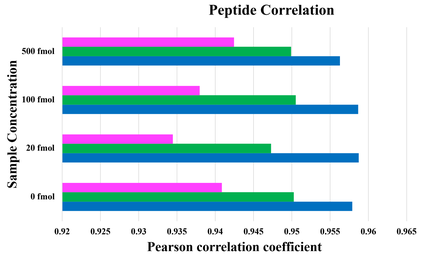
- Software accessibility and usability.
- Share-ability of tools, workflows and histories.
- Reproducibility and ability to test and compare results using multiple parameters.
- The ability to assimilate disparate software into integrated workflows.

Methods

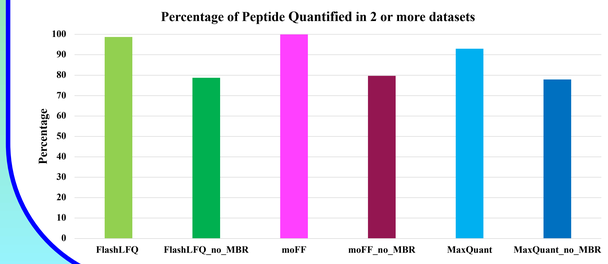
- ***Specificity of Match Between Runs:***
  - Four human cell lysate samples were spiked with four proteins: ABRF-1 (beta galactosidase from *Escherichia coli)*, ABRF-2 (lysozyme from *Gallus gallus*), ABRF-3 (amylase from *Aspergillus niger*) and ABRF-4 (protein G from *Streptococcus*). Each sample contained the four proteins at the same concentration, while the concentrations varied across samples (20, 100 and 500 fmol) with one negative control.
  - Database search was done using the Andromeda search engine within MaxQuant. The spectrum report file (msms.txt) was used as the peptide identification file for evaluating the quantification tools.
  - The outcome of interest was whether MBR identified absent proteins within the negative control.
- ***Fold Change Accuracy and Identification of Differentially Expressed Proteins:***
  - We obtained publicly available data (PRIDE #5412), in which Cox, *et al.* spiked *E. coli* K12 strain samples the UPS1 and UPS2 standards (SigmaAldrich). The UPS1 and UPS2 standards contain 48 human proteins at the same (5000 fmol) or varying concentration (50,000 fmol to 0.5 fmol), respectively. Database search was done using the Andromeda search engine within MaxQuant. The tabular output (msms.txt) file was used as the input for moFF and FlashLFQ.
  - For the FlashLFQ and moFF output, as well as peptide-level MaxQuant output (peptides.txt), we normalized the intensities using each of limma's (Ritchie *et al.*, 2015, *Nuc Acid Res.)* four different normalization methods, then aggregated the peptide intensities to protein level and estimated fold changes with PECA (Suomi.T *et al.*, 2015. *J Proteome Res*). MaxQuant's protein-level output (proteinGroups.txt) was also used. PECA performs protein quantitation as well as estimates fold changes.
  - For all tools and normalization methods, we compared the estimated fold changes (the UPS2-to-UPS1 ratio) with the true fold changes, and the ability of the tools to identify truly differentially expressed proteins as such.

Match Between Runs (MBR) Vs No-match Between Runs

- "Matching between runs" (MBR) is preferred in MaxQuant while perform LFQ quantitation.
- The MBR vs noMBR feature from moFF and FlashLFQ tools were evaluated with MaxQuant results.
- The initial results revealed that MBR feature in moFF and FlashLFQ outputs intensities for the negative control compared to MaxQuant.
- The MBR feature is still under development for moFF and FlashLFQ.

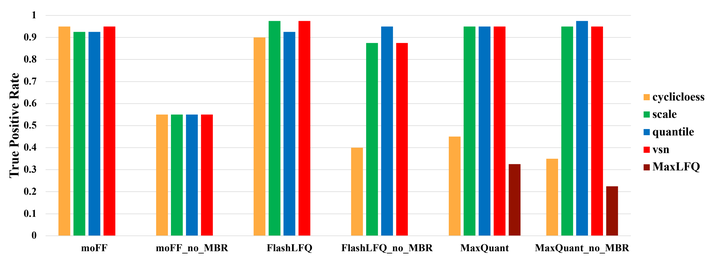


- For evaluating the tools, the raw intensities of the peptides and proteins were correlated.
- The Pearson correlation coefficient was used for the comparison of intensities at peptide level.
- Peptide intensities from all tools showed high correlations (> 93%)



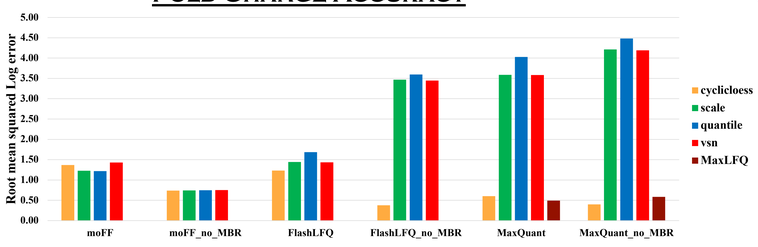
- The proportion of peptides that are quantified in 2 or more datasets increases when the MBR feature is used.
- Using MBR, moFF presents the highest percentage of quantified peptides shared across datasets.

Differential Expression



- The true positive rate (TPR) of differentially expressed proteins were compared between the normalization methods to evaluate the sensitivity of the tool. If a protein was not tested due to missing values, it was counted as a false negative.
- Notably, MaxQuant's built-in normalization method MaxLFQ led to the lowest TPR of all the tools.

Fold Change Accuracy

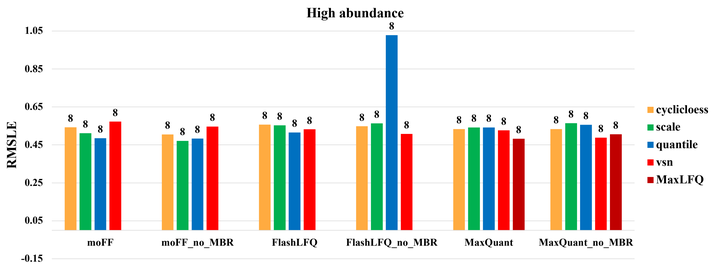


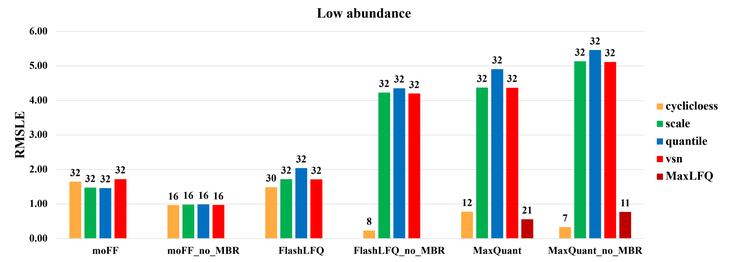
After normalization, the estimated protein ratios for the 48 UPS proteins were compared to the true ratios, using the Root mean squared log error (RMSLE).

$$\text{RMSLE} = \sqrt{\frac{\sum_{i=1}^{N}(\log_{10} r_i - \log_{10} \hat{r}_i)^2}{N}}$$

Where, r_i is the true ratio, $\hat{r}_i$ is the estimated ratio, and N is the number of proteins in the sample

High Abundance vs Low Abundance Fold Change accuracy

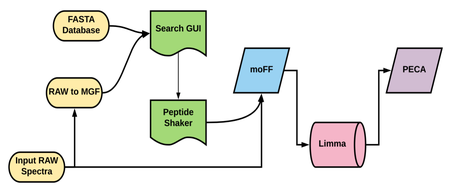




- Value on the top of the bars denote the number of proteins that were quantified.
- The UPS2/UPS1 concentration ratio of 10 were classified as high abundant and less than 1 were classified as low abundant proteins.
- The RMSLE of the intensity ratio was used to measure the accuracy of the estimated fold change

Observations And Conclusions

	MaxQuant	FlashLFQ	moFF
Open Source	No	Yes	Yes
Associated Software/ Search Algorithm	Andromeda Search Engine, doesn't take outputs from any other Search algorithms	Metamorpheus/Morpheus, MaxQuant results or Peptide Shaker output.	Peptide shaker or MaxQuant
Evaluated within Galaxy	No	Installed but testing required	Installed and tested
Quantitation Outputs	Tabular output (Peptide level and Protein level)	Tabular output (Peptide level and Protein level-(in development))	Tabular output (Peptide only)
Normalization	Yes (MaxLFQ)	No (in-development)	No
MBR Feature	Yes	Yes (in-development)	Yes (under testing)



Workflow

- For this evaluation study, we compared the accuracy of quantitation performed by moFF, FlashLFQ, and MaxQuant.
- In all three tools, MBR increased the number of peptides quantified in multiple samples. The MBR feature in FlashLFQ and moFF led to a higher sensitivity (more peptides quantified) that was however offset by a lower specificity (more false positives reported).
- Our initial evaluations suggests that all the normalization methods perform similarly, but cyclic LOESS normalization can force low abundance proteins to have an intensity of zero.
- Because the tools show similar performance, the researcher's choice of tool should depend largely on the desired analysis pipeline (see table above for tool features).
- Results from this evaluation study will be used as a benchmark for future proteomics and multi-omics quantification studies utilizing the Galaxy platform.

Future Directions

- Extensive testing and refinement of parameters to improve MBR outputs from these tools.
- Evaluation using fractionated datasets and making it compatible with Galaxy's search algorithm outputs (Peptide Shaker PSM report).
- Providing inputs and suggestions to the developers for improving the efficiency and accuracy of these tools.
- Integrating these quantitative tools within the existing multi-omic workflows to provide biological insights.

Acknowledgements

• This project is supported by National Science Foundation (NSF) grant "1458524" and National Institutes of Health (NIH) grant "U24CA199347".

• The ABRF Data for specificity of Match Between Runs was generated through the collaborative work of the ABRF Proteomics Research Group (https://abrf.org/research-group/proteomics-research-group-prg). Reference: Van Riper, S. et al. 'An ABRF-PRG study: Identification of low abundance proteins in a highly complex protein sample' at the 64th Annual Conference of American Society of Mass Spectrometry and Allied Topics" at San Antonio, TX'.

• Galaxy-P infrastructure is maintained by Minnesota Supercomputing Institute at the UMN (www.msi.umn.edu). The Galaxy-P team also uses Jetstream cloud instance (Indiana University) as part of a XSEDE Research Allocation. (https://jetstream-cloud.org/).