Evaluation of precursor-based peptide quantification software tools within the Galaxy framework.

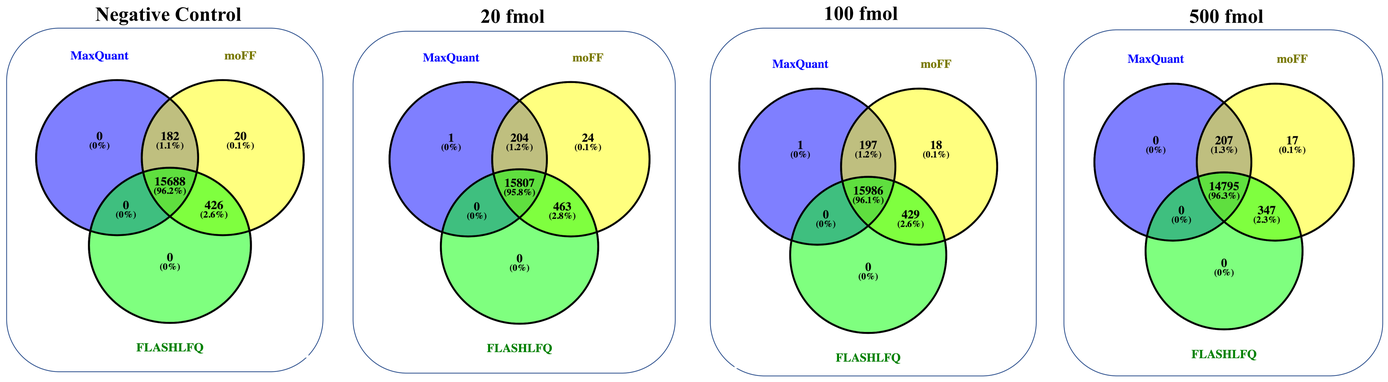
Subina Mehta[1], Caleb Easterly[1], James E. Johnson[2], Björn Grüning[3], Andrea Argentini[4-6], Robert J.Millikin[7], Michael Shortreed[7], Thomas McGowan[2], Praveen Kumar[9] , Lennart Martens[4-6], Lloyd Smith[7,8], Timothy J Griffin[1] and Pratik Jagtap[1]

[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]Department of Medical Protein Research, 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

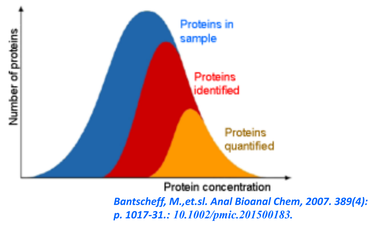
- Protein / peptide level quantification (either labeled or label-free) is routinely used in shotgun proteomics data analysis for determining the abundance of proteins in a given sample.
- Mass Spectrometry (MS) has been widely used for proteomics to quantify the absolute or relative protein expression levels from different biological conditions.
- Accurate and robust label-free quantitation is still a major challenge in the field of quantitative proteomics.
- Label-free quantification based on precursor peak intensities (MS1) reliable due to its accuracy, efficiency & ease.
- Primary goal is to evaluate open-source quantitation tools such as moFF (DOI:10.1038/nmeth.4075) and FlashLFQ (DOI:10.1021/acs.jproteome.7b00608) within and outside the Galaxy-P platform and comparing outputs with MaxQuant (DOI:10.1038/nprot.2016.136) .
- Secondary goal is to implement these tools within the Galaxy-P framework.
- Galaxy-P team has developed workflows for proteomics (identification of modified and unmodified peptides); proteogenomics (identification of variant peptides) and metaproteomics (identification of 'metapeptides' for taxonomy and functional assignment). In all of these studies, quantification at the peptide-level is critical.

COMPARISON OF ALL QUANTIFIED PEPTIDES



- Across all the samples, FlashLFQ and moFF quantifies approximately 400+ peptides more than MaxQuant.
- About 20+ peptides were uniquely quantified by moFF across all samples.

OBJECTIVES

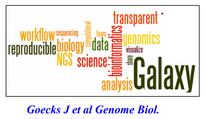


Bantscheff, M.,et.sl. Anal Bioanal Chem, 2007. 389(4): p. 1017-31.: 10.1002/pmic.201500183.

- Evaluate open-source label-free peptide quantitation tools.
- Testing tools inside and outside Galaxy-P.
- Integrating evaluated tools within Galaxy workflows for quantifying identified peptides.

GALAXY PLATFORM

- A web-based bioinformatics data analysis platform.
- Software accessibility and usability.
- Share-ability of tools, workflows and histories.
- Reproducibility and ability to test and compare results using multiple parameters.
- Ability to assimilate disparate software into integrated workflows.



Goecks J et al Genome Biol. 2010;11(8):R86.

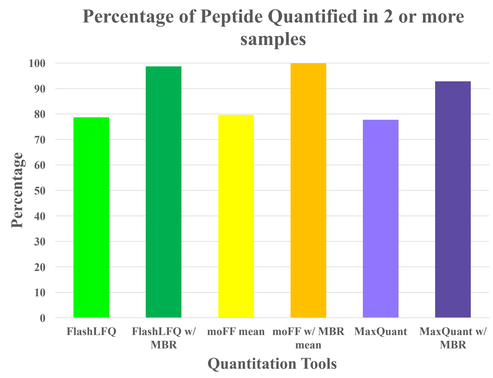
QUANTIFICATION TOOLS

FLASHLFQ

moFF

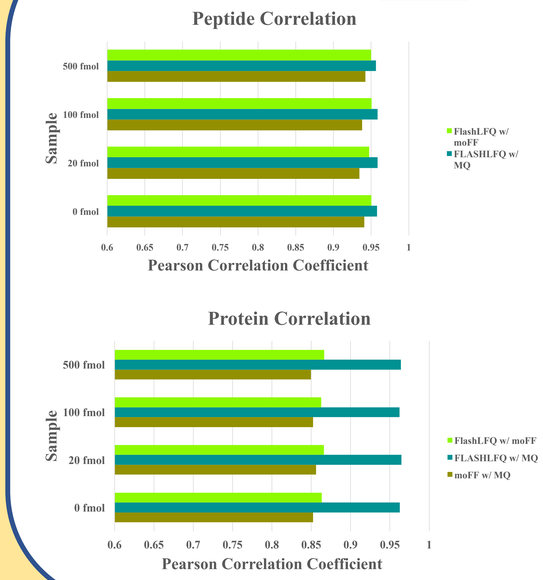
MQ

MATCH BETWEEN RUNS (MBR) VS No-MATCH BETWEEN RUNS



- Proportion of peptides that are quantified in 2+ samples increase when MBR feature is used.
- Using MBR, moFF presents the highest percentage of quantified peptides shared across samples.

INTENSITY CORRELATION BETWEEN TOOLS

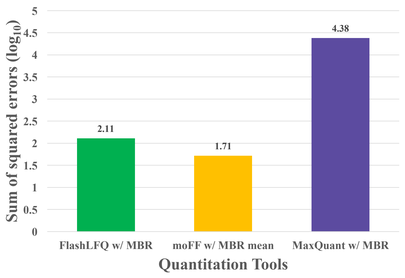


- For evaluating the tools, the raw intensities of the peptides and proteins were correlated.
- Pearson Correlation coefficient was used for the comparison of intensities at peptide level.
- FlashLFQ has higher correlation with MaxQuant than moFF.

METHODS

- Four human cell lysate samples spiked with four proteins at different concentrations (20, 100 and 500 fmol) were obtained. A sample without spiked-in proteins functioned as a negative control.
- These proteins were labeled in the protein search database as 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*).
- The database search was done using Andromeda search engine within MaxQuant. The tabular output (msms.txt) file was used as the peptide identification file for evaluating the quantification tools.
- The parameters (RT window and Tolerance) were kept similar for all the three tools for evaluation purpose.
- moFF and FlashLFQ is installed within the Galaxy-P Platform. Tests were conducted both, within and outside the Galaxy-P platform.
- Outputs generated from moFF and FlashLFQ were compared with the MaxQuant.

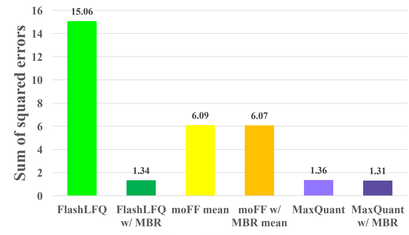
ACCURACY OF FOLD CHANGE ESTIMATION



$$SSE^{*}\ FOLD\ CHANGE = \sum_{S=1}^{3}\sum_{P=1}^{4}(\hat{R}_{PS} - R_S)^2$$

Where,
S = comparison (500 fmol:100 fmol, 500 fmol: 20 fmol, 100 fmol:20 fmol)
P = protein
$\hat{R}_{PS}$= Estimated ratio for protein P and comparison S
R_S = True ratio for comparison S

- For evaluating the accuracy of quantified results, we observed the fold change between the spiked-in proteins.
- We compared three sample ratios (500 fmol with 100 fmol, 500 fmol with 20 fmol and 100 fmol with 20 fmol).
- For evaluation purpose, the Sum of squared errors (SSE) was used for fold change estimation.
- The initial outputs comparing all ratios from this dataset estimates that MaxQuant with MBR gives more errors compared to FlashLFQ and moFF.
- For this dataset, the most accurate ratio was 500 fmol to 100 fmol across all tools. FlashLFQ without MBR displayed higher SSE, whereas MaxQuant performed better.



$$SSE^{*}\ FOLD\ CHANGE = \sum_{P}^{4}(\hat{R}_{P,500:100} - 5)^2$$

Where:
P = protein
$\hat{R}_{P,500:100}$= Estimated ratio for protein P and comparison 500 fmol to 100 fmol

*SSE=SUM OF SQUARED ERRORS

SPIKED-IN PROTEINS INTENSITY COMPARISON

	METHOD	PROTEIN	NEGATIVE CONTROL	20 fmol	100 fmol	500 fmol
FlashLFQ	NO-MBR	ABRF-1	-	-	7.009	7.902
		ABRF-2	-	5.251	6.285	7.142
		ABRF-3	-	-	6.545	7.355
		ABRF-4	-	5.542	6.578	7.309
	MBR	ABRF-1	5.938	6.159	7.131	7.893
		ABRF-2	5.487	5.449	6.513	7.143
		ABRF-3	5.659	6.043	6.645	7.348
		ABRF-4	5.065	5.958	6.649	7.308
moFF	NO-MBR	ABRF-1	-	-	5.762	6.430
		ABRF-2	-	5.281	5.774	6.444
		ABRF-3	-	-	5.587	6.121
		ABRF-4	-	5.554	5.787	6.622
	MBR	ABRF-1	5.223	5.311	5.702	6.433
		ABRF-2	5.330	5.444	5.807	6.451
		ABRF-3	4.532	5.024	5.473	6.122
		ABRF-4	4.553	5.411	5.762	6.625
MAXQUANT	NO-MBR	ABRF-1	-	-	8.248	8.970
		ABRF-2	-	6.424	7.373	8.062
		ABRF-3	-	-	7.647	8.372
		ABRF-4	-	6.718	7.608	8.393
	MBR	ABRF-1	-	6.936	8.306	8.972
		ABRF-2	-	6.441	7.387	8.074
		ABRF-3	-	5.642	7.631	8.379
		ABRF-4	-	7.176	7.795	8.408

The values listed here are the $\log_{10}$ intensity.

- The Match Between runs (MBR) is a FDR-controlled algorithm that facilitates MS/MS free identification of MS features and also increases the number of quantified protein in the sample.
- The MBR feature from moFF and FlashLFQ tools were evaluated against the no-MBR feature to determine the intensity of the spiked-in proteins in the samples.
- The initial results show (red) that MBR feature in moFF and FlashLFQ outputs values for the Sample 2 (negative control).
- The MBR feature is still under development for moFF and FlashLFQ.
- moFF performs peptide-level quantification, so we averaged peptide intensities to obtain protein-level quantification.

OBSERVATIONS AND CONCLUSIONS

- For testing moFF and FlashLFQ, tabular output from MaxQuant (msms.txt) was used, respectively, along with the RAW files.
- We compared the results of extracting MS1 intensity with and without "match between runs" (MBR) (sharing peptide identifications across files), in terms of number of peptides quantified and the total peptide intensities obtained.
- Our initial evaluation shows that moFF and FlashLFQ quantifies similar to MaxQuant but the MBR feature works well in MaxQuant.
- As tools are still in the development stage, the developers of these software tools were provided with the feedback after active testing, so that parameters and features in the software tools could be optimized.
- 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 tools within our previously developed workflows to provide better insights regarding the biological significance of our analysis.

ACKNOWLEDGEMENTS

- *This project is supported by National Science Foundation (NSF) grant "1458524" and National Institutes of Health (NIH) grant "U24CA199347".*
- *Data used for evaluation was generated through the collaborative work of the ABRF Proteomics Research Group.(https://abrf.org/research-group/proteomics-research-group-prg).*
- *Computational resources were provided by Minnesota Supercomputing Institute (MSI).*
- *The Galaxy-P team also uses Jetstream cloud instance (Indiana University) as part of a XSEDE Research Allocation.*