proteomes

MDPI

Technical Note

Precursor Intensity-Based Label-Free Quantification Software Tools for Proteomic and Multi-Omic Analysis within the Galaxy Platform

Subina Mehta [1,*], Caleb W. Easterly [1], Ray Sajulga [1], Robert J. Millikin [2], Andrea Argentini [3], Ignacio Eguinoa [3], Lennart Martens [3], Michael R. Shortreed [2], Lloyd M. Smith [2], Thomas McGowan [4], Praveen Kumar [1], James E. Johnson [4], Timothy J. Griffin [1] and Pratik D. Jagtap [1,*]

[1] Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN 55455, USA; caleb_easterly@med.unc.edu (C.W.E.); rsajulga@umn.edu (R.S.); kumar207@umn.edu (P.K.); tgriffin@umn.edu (T.J.G.)

[2] Department of Chemistry, University of Wisconsin, Madison, WI 53706, USA; rmillikin@wisc.edu (R.J.M.); mshort@chem.wisc.edu (M.R.S.); smith@chem.wisc.edu (L.M.S.)

[3] VIB-UGent Center for Medical Biotechnology, VIB, Ghent University, 9000 Ghent, Belgium; a.argentini@gmail.com (A.A.); ignacio.eguinoa@psb.vib-ugent.be (I.E.); lennart.martens@vib-ugent.be (L.M.)

[4] Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN 55455, USA; jj@umn.edu (J.E.J.); mcgo0092@umn.edu (T.M.)

* Correspondence: smehta@umn.edu (S.M.); pjagtap@umn.edu (P.D.J.)

Received: 12 June 2020; Accepted: 7 July 2020; Published: 8 July 2020

Abstract: For mass spectrometry-based peptide and protein quantification, label-free quantification (LFQ) based on precursor mass peak (MS1) intensities is considered reliable due to its dynamic range, reproducibility, and accuracy. LFQ enables peptide-level quantitation, which is useful in proteomics (analyzing peptides carrying post-translational modifications) and multi-omics studies such as metaproteomics (analyzing taxon-specific microbial peptides) and proteogenomics (analyzing non-canonical sequences). Bioinformatics workflows accessible via the Galaxy platform have proven useful for analysis of such complex multi-omic studies. However, workflows within the Galaxy platform have lacked well-tested LFQ tools. In this study, we have evaluated moFF and FlashLFQ, two open-source LFQ tools, and implemented them within the Galaxy platform to offer access and use via established workflows. Through rigorous testing and communication with the tool developers, we have optimized the performance of each tool. Software features evaluated include: (a) match-between-runs (MBR); (b) using multiple file-formats as input for improved quantification; (c) use of containers and/or conda packages; (d) parameters needed for analyzing large datasets; and (e) optimization and validation of software performance. This work establishes a process for software implementation, optimization, and validation, and offers access to two robust software tools for LFQ-based analysis within the Galaxy platform.

Keywords: proteomics; label-free quantification; galaxy framework; workflows

1. Introduction

Peptide- and protein-level quantification (either labeled or label-free) is routinely used in mass spectrometry (MS)-based shotgun proteomics data analysis workflows to determine the relative abundance of peptides or proteins in a given sample [1], including post-translationally modified peptides [2] and amino acid sequence variants identified by proteogenomics [3,4]. In the field of metaproteomics, where protein samples obtained from environmental microbiomes are studied, the quantification of microbial peptides or "metapeptides" (peptides obtained from shotgun sequencing of microbial communities) is essential to perform taxonomic and functional quantification of proteins expressed from the microbiome [5].

In the case of the label-free quantification (LFQ) methods, the peak intensity or area under the curve of a detected peptide ion allows the relative quantification of peptides across different samples. LFQ [6,7] is a useful method for quantification when the introduction of stable isotopes is impractical (for example, in human or animal model studies) or for applications such as proteogenomics or metaproteomics, which rely on peptide-level quantification. Currently, there are several software packages available for LFQ analysis [8]. LFQ analysis can be performed by public domain software suites such as MaxQuant [9] and Skyline [10], or by commercial software such as PEAKS [11] and Progenesis [12]. Although commercial and actively-supported software offers reliability and ease of use, its usage comes with a cost and usually includes canned features that are used for most standard datasets. Open-source software, on the other hand, has the benefit of being amenable to testing and optimization for emerging disciplines to offer economical options for data analysis.

In this study, through a rigorous testing and evaluation process, we have incorporated and optimized two established, open-source tools, moFF [13] and FlashLFQ [14] in the Galaxy platform. In order to achieve this, we worked with the software developers of these tools and tested features using two benchmark datasets, the ABRF Proteomics Research Group (PRG) 2015 dataset [15] and a Universal Proteomics Standard (UPS) dataset [1], and compared the outputs with results from MaxQuant, a highly used standalone software platform capable of LFQ analysis. Based on feedback, the tool developers of moFF and FlashLFQ made changes to the software's capabilities, which included (a) using match-between-runs (MBR); (b) ability to process and analyze large input datasets; (c) compatibility with a variety of input file formats. After this rigorous evaluation and optimization, these tools were implemented in the accessible and reproducible Galaxy [16] platform. Galaxy tools are maintained and developed by an international community of developers (https://galaxyproject.org/iuc/) so as to facilitate ease of usage and maintain its contemporary status for any emerging software tools or applications. An additional advantage of having these tools available via the Galaxy platform is the ability to process the data in workflows, wherein multiple tools can be used in a sequential manner to generate processed outputs from the input data. The Galaxy for proteomics (Galaxy-P) team has developed workflows related to MS-based multi-omic studies such as, proteogenomics [17,18] and metaproteomics [19,20]. The addition of these precursor intensity-based LFQ tools to the existing workflows will facilitate peptide level quantification for multi-omics research studies, as well as more standard proteomics applications.

As a result of this study, we made two quantitative software tools available to researchers via the Galaxy platform. These software are available via the Galaxy Tool Shed [21,22], GitHub, and on Galaxy public instances.

2. Methods

We used two datasets, (a) an ABRF dataset [15] and (b) a spiked-in benchmark UPS dataset [1] to determine the accuracy of each tool with regards to their calculated protein fold-changes. We obtained the MS (raw) data from publicly available repositories and converted them to MGF (Mascot generic format) files using MSConvert (vendor support) (Galaxy Version 3.0.19052.0) [23] to make it compatible with search algorithms within the Galaxy Platform. moFF and FlashLFQ processing were performed within the Galaxy platform (Version 19.09).

2.1. (A) ABRF Dataset

The spiked-in dataset from the ABRF PRG 2015 study was used to determine the accuracy of each software tool. This dataset, generated through the collaborative work of the ABRF Proteomics Research Group (https://abrf.org/research-group/proteomics-research-group-prg) contains four proteins added to human cell lysate samples: 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*) [15]. Each sample contained the four proteins at the same concentration, while the concentrations varied across the four samples: 0 (blank/negative control), 20, 100, and 500 fmol. The peptide raw data for the ABRF dataset was acquired on the LTQ orbitrap Velos with EASY-nLC.

2.2. (B) Spiked-In UPS Benchmark Dataset

To evaluate these tools, we downloaded publicly available data [1] (PRIDE #5412; ProteomeXchange repository PXD000279), wherein UPS1 and UPS2 standards (Sigma-Aldrich, St. Louis, MO, USA) were spiked into *E. coli* K12 strain samples. Based on the dynamic benchmark dataset protocol, the UPS and *E. coli* peptides we quantified using nanodrop spectrophotometer at 280 nm, and 2 µg of *E. coli* peptides were spiked with 0.15 µg of UPS1 or UPS2 peptides. About 1.6 µg of the mix was analyzed on the Q Exactive (Thermo Fisher, Waltham, MA, USA) mass spectrometer [1]. The UPS1 and UPS2 standards contain 48 human proteins at either the same (5000 fmol, UPS1) or varying concentration (50,000 fmol to 0.5 fmol, UPS2), respectively.

2.3. Peptide Identification

For both datasets, we used SearchGUI (SG) [24] (version 3.3.3.0) and Peptide Shaker (PS) [25] (version 1.16.26) to search the MS/MS spectra against respective protein FASTA databases along with contaminants from cRAP database (https://www.thegpm.org/crap/). Although SearchGUI has the option to use as many as eight search algorithms, we used only four search algorithms (X!tandem, OMSSA, MSGF+, and Comet) for this evaluation study.

For the spiked-in ABRF PRG dataset, a protein FASTA file was generated by merging the UniProt human reference database with spiked-in proteins and contaminant proteins (73,737 protein sequences database generated on 6 February 2019). Search parameters used were trypsin enzyme for digestion, where two missed cleavages were allowed. Carbamidomethylation of cysteine was selected as a fixed modification and methionine oxidation was selected as a variable modification. The precursor mass tolerance was set to 10 ppm and the fragment mass tolerance to 0.5 Da, with minimum charge as 2 and maximum charge of 6. For Peptide Shaker, the false discovery rate (FDR) was set at 1% at the PSM, peptide, and protein level, along with filtering the peptide length ranging from 6–65 peptides.

For the spiked-in UPS dataset, the mass spectra were searched against a protein FASTA database provided by Cox. et al., 2014 [1], (4494 protein sequences database generated on 25 July 2019). The parameters for SearchGUI-Peptide Shaker analysis were as follows: precursor mass tolerance was set to 10 ppm and the fragment mass tolerance to 20 ppm with minimum and maximum charge as 2 and 6, respectively.

For MaxQuant analysis (version 1.6.7.0, Cox lab, Max Planck institute of Biochemistry, Martinsried, Germany), the built-in Andromeda search engine [26] was used. The parameters for MaxQuant were matched with the SearchGUI-PeptideShaker search. The fixed modification was set for carbamidomethylation of cysteine and oxidation of methionine as a variable modification. The FDR was set at 1% and the MS/MS tolerance was set at 10 ppm. The tabular output data from Peptide Shaker (PSM.tab) and Andromeda (msms.txt) were used for protein quantification.

2.4. Quantification Tools

moFF and FlashLFQ, were initially tested outside of the Galaxy platform. We tested various releases for moFF (versions 1.2.0 to 2.0.2) and FlashLFQ (versions 0.1.99 to 1.0.3) and provided developers with feedback to improve software stability and data quality. We then implemented these updated tools within Galaxy. The results from moFF (version 2.0.2) and FlashLFQ (version 1.0.3) were then compared with MaxQuant (version 1.6.0.16), a widely used LFQ quantification software suite. For testing, all the quantification tools were set at monoisotopic tolerance of 10 ppm and run with or without MBR, where indicated.

2.5. Normalization and Protein Quantification

After peptide-level precursor intensity values were generated, normalization was performed using limma [27], and peptides were summarized into protein-level abundances with protein expression control analysis (PECA) [28]. Specifically, the "normalizeBetweenArrays" limma function was used for most normalization methods (i.e., scale, cyclic loess, and quantile). For VSN (variance stabilizing normalization), the "normalizeVSN" limma function was used [27,29]. After normalization, PECA was used to combine the peptide-level measurements to protein-level values for the detection of differentially expressed proteins. These two tools were run via custom R scripts (version 1.3), which can be accessed via the Supplementary Document 2 (https://github.com/galaxyproteomics/quant-tools-analysis).

3. Results

Both moFF and FlashLFQ are established software tools and contain useful features such as amenability to Galaxy implementation, compatibility with existing Galaxy upstream and downstream tools, ability to read mzML and Thermo raw file formats, open-source code, MBR functionality, and results that can be easily evaluated with performance metrics.

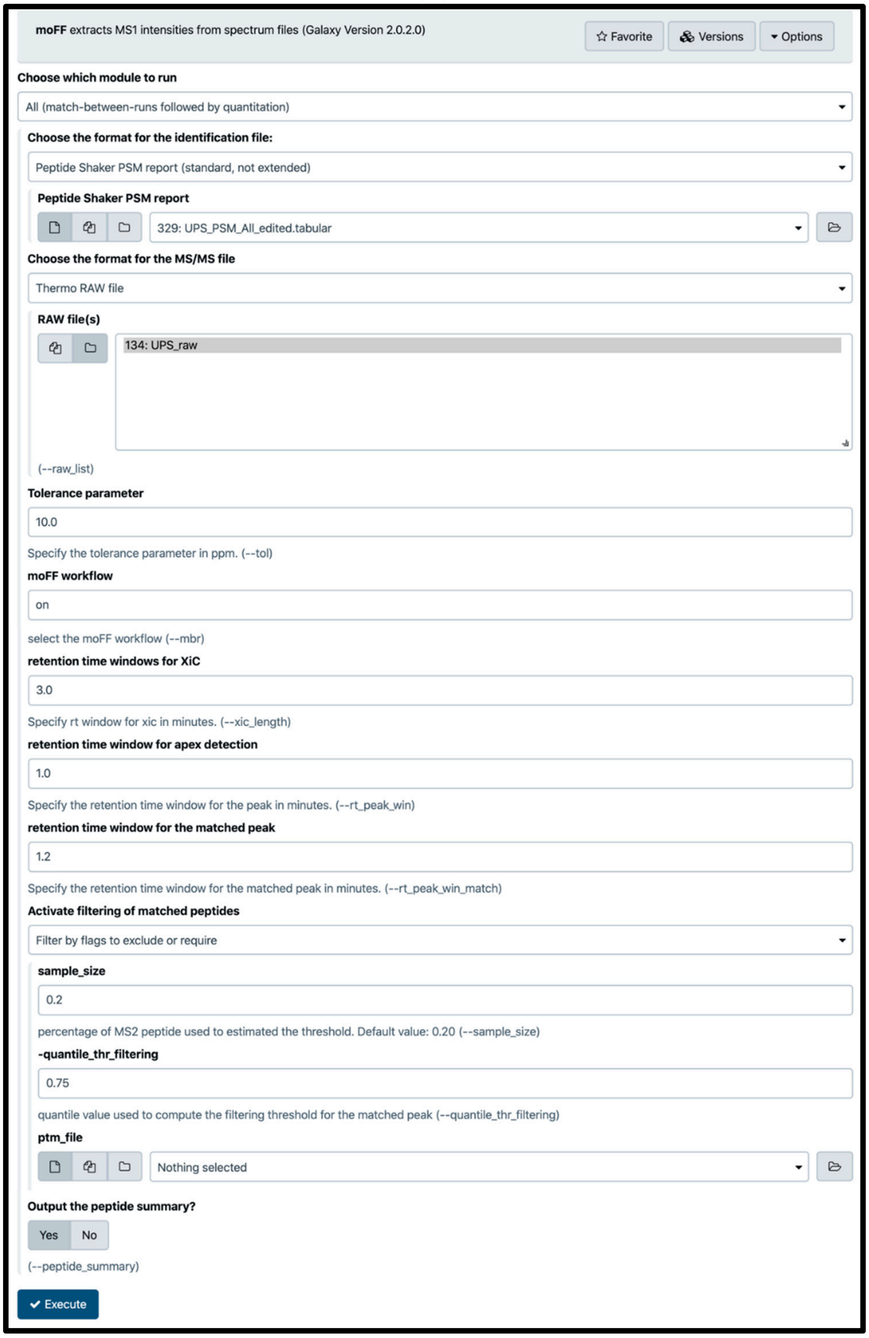
moFF is an extensible quantification tool amenable to any operating system. The input for moFF is peptide search engine output and Thermo raw files and/or mzML files; it performs both MS/MS as well as MBR quantification. moFF tool also has a novel filtering option for MBR peak intensities [30]. moFF has been wrapped in Galaxy (Figure 1A) using a Bioconda package [31]. The Galaxy version of moFF is available via Galaxy toolshed [21], GitHub [32] and Galaxy public instances (proteomics.usegalaxy.eu, usegalaxy.be and z.umn.edu/metaproteomicsgateway).

FlashLFQ is a peptide and protein LFQ algorithm developed for proteomics data analysis. It was developed to quantify peptides and proteins from any search tool, including MetaMorpheus, which also performs PTM identification from MS/MS data. It uses Bayesian statistics to estimate the difference in the abundance of inferred proteins between samples, though this feature was not evaluated here. FlashLFQ can normalize fractionated datasets by using a bounded Nelder–Mead optimizer [33] to find a normalization coefficient for each fraction, similar to MaxLFQ. FlashLFQ was implemented in Galaxy (Figure 1B) within a Singularity container [34] as FlashLFQ is a Windows application requiring the NET core framework for deployment in the Unix-based Galaxy environment. Singularity provides a secure means of running such tools in Galaxy. The Galaxy version of FlashLFQ is available via Galaxy toolshed [22] GitHub [35] and via Galaxy public instances (proteomics.usegalaxy.eu and z.umn.edu/metaproteomicsgateway).

Relevant features of moFF and FlashLFQ, as well as the design of the evaluation study, are summarized in Figure 2. An essential aspect of this study was both of these tools being in active development by groups amenable to collaboration, which greatly helped optimization-based tests and feedback from the Galaxy-P team members.

To generate peptide identification inputs for moFF and FlashLFQ, datasets were searched against appropriate protein databases using SearchGUI/PeptideShaker. The speed and accuracy of FlashLFQ and moFF were evaluated in comparison to MaxQuant, a popular software tool used for LFQ. For MaxQuant, searches were performed by MaxQuant's built-in Andromeda search algorithm. All three software programs have an MBR feature, where unidentified peaks are "matched" to identify peaks in other runs based on similar *m/z* and retention time. MaxLFQ, an algorithm within MaxQuant, normalizes raw intensities, and also aggregates them into protein groups [1]. For moFF and FlashLFQ, limma was used to normalize peptide intensities, and PECA was used to determine protein fold-changes and associated *p*-values. The limma tool within Galaxy implements different normalization techniques such as quantile, VSN, cyclic LOESS, and scale normalization. Users can choose between these normalization methods. FlashLFQ also has built-in normalization and protein quantification functions, which we have used in this study.

(**A**)



moFF extracts MS1 intensities from spectrum files (Galaxy Version 2.0.2.0)

Choose which module to run

All (match-between-runs followed by quantitation)

Choose the format for the identification file:

Peptide Shaker PSM report (standard, not extended)

Peptide Shaker PSM report

329: UPS_PSM_All_edited.tabular

Choose the format for the MS/MS file

Thermo RAW file

RAW file(s)

134: UPS_raw

(--raw_list)

Tolerance parameter

10.0

Specify the tolerance parameter in ppm. (--tol)

moFF workflow

on

select the moFF workflow (--mbr)

retention time windows for XiC

3.0

Specify rt window for xic in minutes. (--xic_length)

retention time window for apex detection

1.0

Specify the retention time window for the peak in minutes. (--rt_peak_win)

retention time window for the matched peak

1.2

Specify the retention time window for the matched peak in minutes. (--rt_peak_win_match)

Activate filtering of matched peptides

Filter by flags to exclude or require

sample_size

0.2

percentage of MS2 peptide used to estimated the threshold. Default value: 0.20 (--sample_size)

-quantile_thr_filtering

0.75

quantile value used to compute the filtering threshold for the matched peak (--quantile_thr_filtering)

ptm_file

Nothing selected

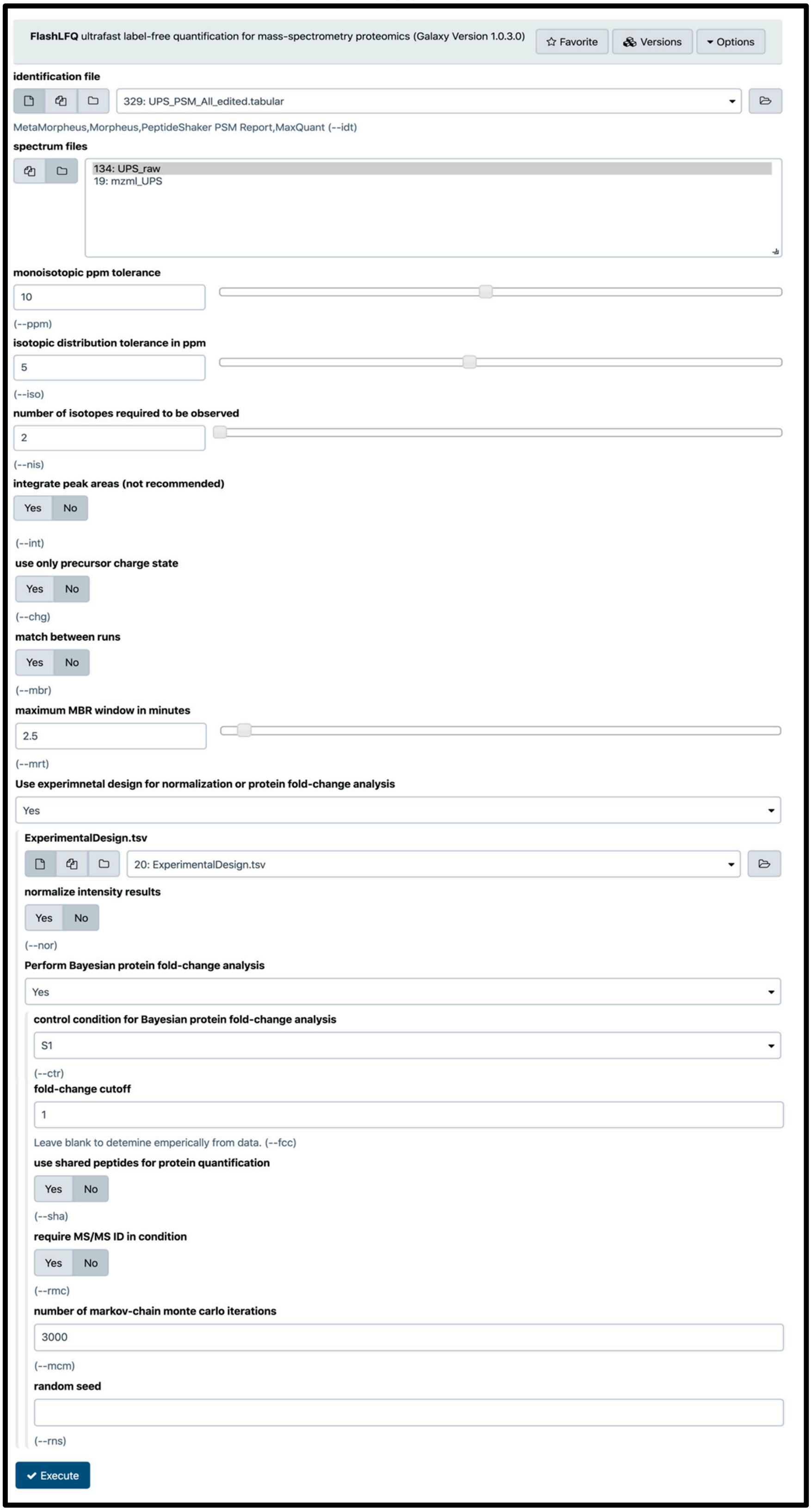
Output the peptide summary?

Yes

(--peptide_summary)

Figure 1. *Cont.*

(**B**)



FlashLFQ ultrafast label-free quantification for mass-spectrometry proteomics (Galaxy Version 1.0.3.0)

identification file
329: UPS_PSM_All_edited.tabular
MetaMorpheus,Morpheus,PeptideShaker PSM Report,MaxQuant (--idt)

spectrum files
134: UPS_raw
19: mzml_UPS

monoisotopic ppm tolerance
10
(--ppm)

isotopic distribution tolerance in ppm
5
(--iso)

number of isotopes required to be observed
2
(--nis)

integrate peak areas (not recommended)
Yes No
(--int)

use only precursor charge state
Yes No
(--chg)

match between runs
Yes No
(--mbr)

maximum MBR window in minutes
2.5
(--mrt)

Use experimnetal design for normalization or protein fold-change analysis
Yes

ExperimentalDesign.tsv
20: ExperimentalDesign.tsv

normalize intensity results
Yes No
(--nor)

Perform Bayesian protein fold-change analysis
Yes

control condition for Bayesian protein fold-change analysis
S1
(--ctr)

fold-change cutoff
1
Leave blank to detemine emperically from data. (--fcc)

use shared peptides for protein quantification
Yes No
(--sha)

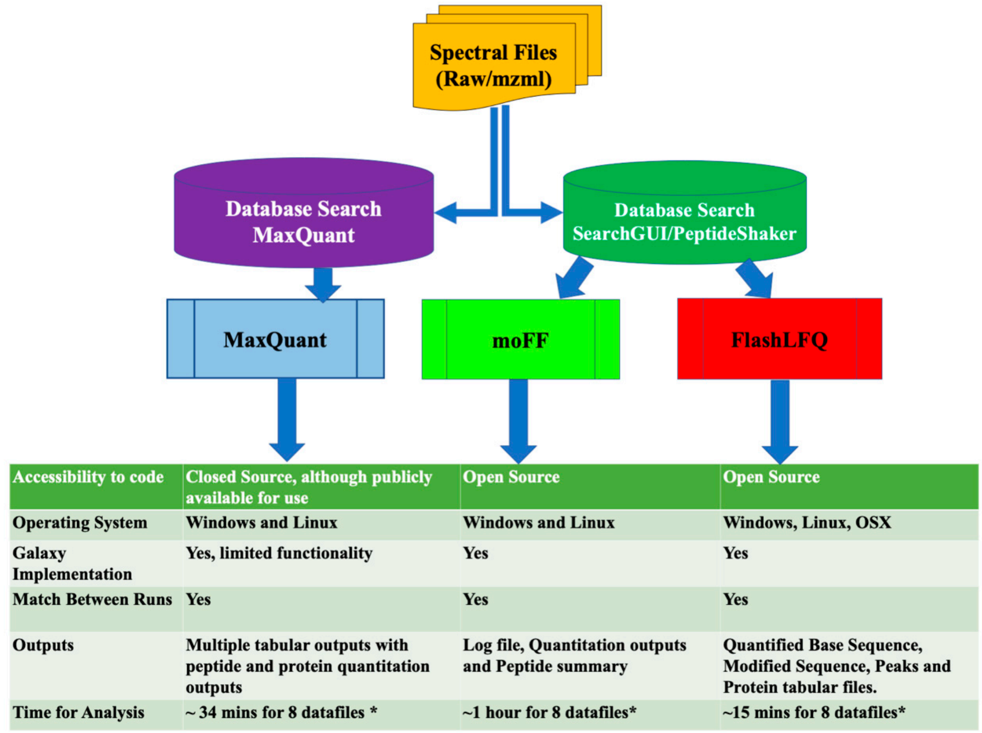
require MS/MS ID in condition
Yes No
(--rmc)

number of markov-chain monte carlo iterations
3000
(--mcm)

random seed

(--rns)

Figure 1. Galaxy interface of moFF and FlashLFQ: (**A**) Bioconductor package of moFF is wrapped within Galaxy and available via Galaxy toolshed [21] and Galaxy public instances (proteomics.usegalaxy.eu). (**B**) A docker/singularity container of FlashLFQ is wrapped within Galaxy and available via Galaxy toolshed [22] and Galaxy public instances (proteomics.usegalaxy.eu).


Spectral Files (Raw/mzml)
Database Search MaxQuant
Database Search SearchGUI/PeptideShaker
MaxQuant
moFF
FlashLFQ

Accessibility to code	Closed Source, although publicly available for use	Open Source	Open Source
Operating System	Windows and Linux	Windows and Linux	Windows, Linux, OSX
Galaxy Implementation	Yes, limited functionality	Yes	Yes
Match Between Runs	Yes	Yes	Yes
Outputs	Multiple tabular outputs with peptide and protein quantitation outputs	Log file, Quantitation outputs and Peptide summary	Quantified Base Sequence, Modified Sequence, Peaks and Protein tabular files.
Time for Analysis	~ 34 mins for 8 datafiles *	~1 hour for 8 datafiles*	~15 mins for 8 datafiles*

Figure 2. Experimental design of the evaluation study: spectra files are converted to MGF before mass spectra are matched with peptides using respective search engines. Each of the quantification tools use RAW files and the peptide identification tabular output as inputs. The figure also shows the features of each tool. The outputs from all of the tools were then compared against each other. The asterisk symbol (*) denotes that the files were run on same computing device.

After ascertaining that moFF and FlashLFQ results correlate well with MaxQuant results (Supplementary Figure S1), we set out to evaluate the MBR feature of the software tools. For this, we used the ABRF PRG datasets, with four spiked-in proteins at three different concentrations (20 fmol, 100 fmol, and 500 fmol) and a negative control (see methods). The spiked-in proteins should not be detected in the negative control, either with or without MBR. We observed that moFF and FlashLFQ outputs showed non-zero intensity values for the spiked-in proteins in blank control samples if MBR was enabled (Figure 3A, left). The match-between runs (MBR) module for the earlier version of both the software, moFF and FlashLFQ, simply predicted the retention time for the matched peptide in the target run by looking for intensity in the target *m/z* and RT window. If the target peptide was not present in that run, overlapping eluting peptides or noise would result in the assignment of spurious background signals. As a result, ABRF spiked in proteins were detected in the blank control sample whole using the MBR mode in the earlier versions of the software tools. We worked with the developers to improve their MBR algorithms so that intensity values for these proteins in the blank control samples were correctly reported as zero (Figure 3B, right). moFF's new version removes spurious matches by filtering scans for the similarity between the theoretical isotopic envelope of target peptide and corresponding envelope found in the target run. This method has been published in the moFF 2.0 version [30]. The FlashLFQ developers implemented an optional setting that requires a protein to have at least one peptide assigned to an MS/MS spectrum in a sample group so that a non-zero value can be assigned to the peptides in the sample group. Due to the above-mentioned changes in the algorithms and subsequent filtering steps, the newer versions of these tools do not detect spurious signals for target peptides in blank control samples.

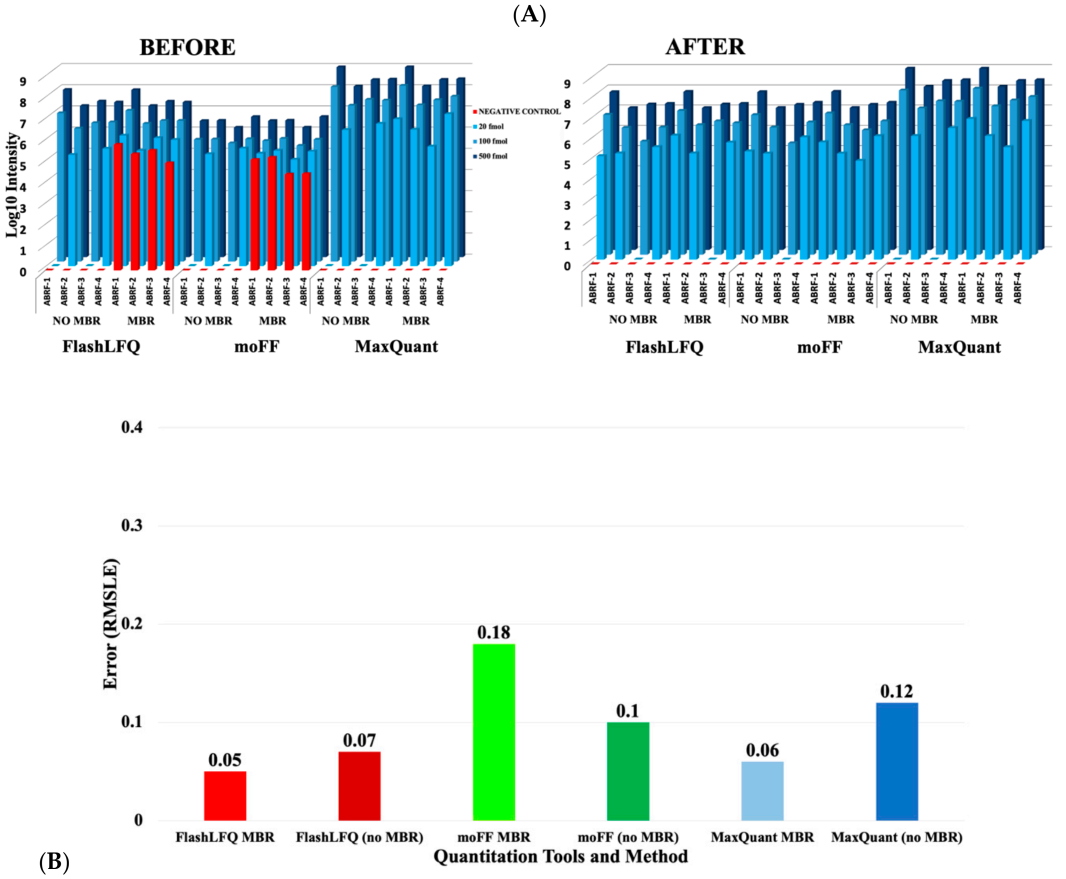


Figure 3. (**A**) Effect of MBR after software version updates: The log10 values of the intensities (blue bars) from each of the four ABRF spiked-in proteins (ABRF-1: beta Galactosidase from *E. coli*, ABRF-2: Lysozyme from *Gallus gallus*, ABRF-3: amylase from *Aspergillus*, ABRF-4: protein G *Streptococcus*) were plotted. The results from prior versions of moFF (v1.2.1) and FlashLFQ (v0.1.99) (before) shows that MBR detects ABRF proteins (shown in red) in the negative control sample in both software. The results from the current versions of moFF (v2.0.2) and FlashLFQ (v1.0.3.0) implemented in Galaxy (after), shows that the MBR feature does not detect ABRF proteins in the negative control. (**B**) Accuracy of fold-change estimation: for evaluating the accuracy of quantified results, we estimated the fold change of the spiked-in proteins in the 500 fmol sample as compared to 100 fmol sample. The root mean squared log error (RMSLE) was calculated for fold change estimation. For this dataset, moFF with MBR displayed significantly higher RMSLE value, whereas FlashLFQ's MBR performed similarly to MaxQuant's MBR.

The 500 fmol and 100 fmol datasets from the ABRF dataset were used to determine the fold-change accuracy (Figure 3B). In order to determine the accuracy of the fold-change, root mean squared log error (RMSLE) [36] was calculated,

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

where, r_i is the true ratio, $\hat{r}_i$ is the estimated ratio and N is the number of proteins identified via sequence database searching of the sample.

Root mean squared log error is a metric to evaluate the difference between predicted and observed values. In this case, the values being compared are the predicted (known) fold-changes and the observed fold-changes. RMSLE being an error based metric provides the true picture of prediction quality, however, deciding a suitable threshold value is challenging. The objective was to obtain an

RMSLE value closer to zero for all the tools. The RMSLE values for the three tools are shown in Figure 3B, with the MBR feature enabled. We observed that moFF with MBR had slightly higher error compared to the other tools. MaxQuant's MBR and FlashLFQ's MBR perform quite similarly, though all three tools show a low error when the MBR is enabled.

Although MaxLFQ and FlashLFQ have their own in-built methods for normalizing peptide abundance, for a more direct comparison of moFF and FlashLFQ performance, we normalized the peptide intensity levels using limma and obtained differentially expressed proteins through the PECA bioconductor package. Normalized peptide intensity values from moFF and FlashLFQ were input into PECA, wherein, the tool calculates the *p*-values of the peptide level data and then groups the values into protein level data. The PECA output was then compared with MaxQuant values using the UPS benchmark dataset. For this, quantitative information for the 48 proteins from the UPS dataset was extracted using an R-script to generate a tabular output with fold-change values.

The fold-change accuracy of all quantified UPS proteins after normalization was calculated by comparing the estimated protein fold-change with the true fold-change using the RMSLE (Figure 4A).

(**A**)

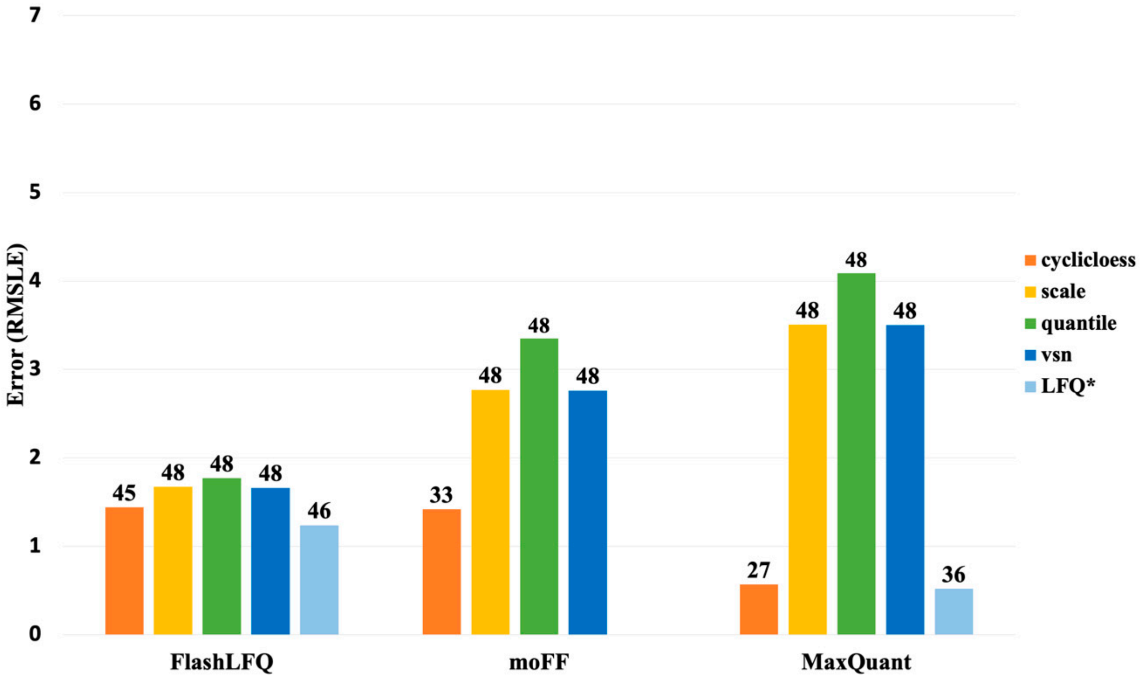


Figure 4. *Cont.*

(**B**)

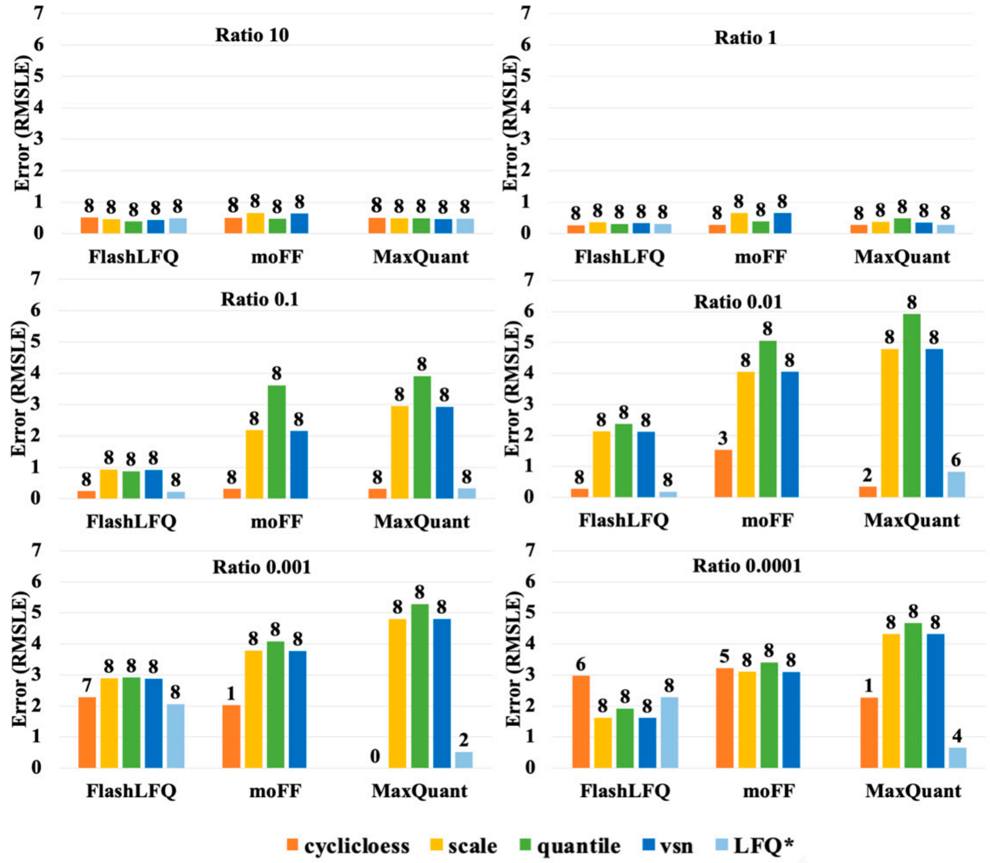


Figure 4. (**A**) Fold-change accuracy (MBR) of all proteins: after normalization, the estimated protein abundance ratios for all the identified UPS proteins were compared to the true abundance ratios, using the root mean squared log error (RMSLE). The plot represents the RMSLE values using different normalization methods. *LFQ denotes the LFQ values represent FlashLFQ's and MaxQuant's inbuilt normalization value. The value on the top of the bars denotes the number of proteins that were quantified. (**B**) Fold change accuracy (MBR) of proteins with similar estimated ratios: In total there are 48 UPS proteins, we classified the UPS proteins into different groups based on the UPS2/UPS1 ratio estimation, the true ratios run from 10 to 10^{-4}. The value on the top of the bars denotes the number of proteins that were quantified using each normalization method. The RMSLE of the intensity ratio was used to measure the accuracy of the estimated fold changes.

Figure 4A shows the comparison of different normalizations using the MBR values. Although the bar graph shows that MaxQuant's MaxLFQ performed the best compared to all, it did so at the cost of the number of proteins quantified. However, we noticed that MBR from FlashLFQ, with its in-built normalization (light blue bar in Figure 4A), performed better overall in terms of quantification and the number of proteins identified. Meanwhile, moFF and FlashLFQ provided higher numbers of quantified proteins while still maintaining low RMSLE values. We also performed comparison studies of MBR vs. no MBR, the results of which are shown in Supplementary Figure S3A. We also found that moFF and FlashLFQ quantified similar numbers of peptides across the ABRF and UPS datasets (Supplementary Figure S2).

After evaluating the RMSLE for all proteins, we estimated the accuracy of similarly abundant UPS proteins (Figure 4B). We categorized the UPS standards by their concentration ratio (UPS2/UPS1), which resulted in 6 different categories (i.e., ratios of 1 to 0.0001). The results showed that MaxQuant quantification works optimally for high and medium abundant proteins. However, for low abundance proteins, the fold-accuracy was lower, presumably because of missing intensity values. Another important observation was that MaxLFQ denotes a smaller error compared to the other tools in the low abundance samples, but quantifies fewer proteins. An evaluation study for MBR vs. noMBR was also performed and showed a similar trend represented in Supplementary Figure S3B.

After evaluation of the moFF and FlashLFQ tools, we worked with the developers of these tools to implement their optimized software in Galaxy, enabling integration into diverse MS-based proteomics workflows and promoting their usage by the Galaxy community. Our implementation will allow the users to choose their choice of tool and normalization (Supplementary Figure S4), which will benefit their research.

4. Discussion

Protein and peptide-level quantification has been used by proteomics researchers to determine how the proteome responds to biological perturbation [37]. In particular, precursor-intensity based LFQ has enabled researchers to perform quantitative proteogenomics analyses [38]. Quantitative changes in the proteome can also be correlated with transcript abundance changes [39] to get a more complete picture of how an organism responds to a stimulus. For example, in cancer proteogenomics studies, these abundance measurements help identify differential expression patterns of variant peptides that may have functional significance in cancer [40,41].

Peptide-level quantification also aids in functional studies of microbial communities and microbiomes using metaproteomics. For example, in metaproteomics studies, metapeptides or metaproteins detected from environmental [42] or host-derived samples [43] can be quantified to shed light on the dynamics of the taxa, biological function, and their abundance [44]. Our group has developed and optimized Galaxy-based tools and workflows for proteogenomics [45] and metaproteomics analyses [19,20]. Tools implemented through this study will extend these workflows to enable quantification of metapeptides and/or metaproteins.

In our analyses, the three LFQ tools, moFF, FlashLFQ, and MaxQuant, correlate well in their results according to our evaluation. In this study, we have added moFF and FlashLFQ to the Galaxy framework, which not only facilitates the dissemination of these tools but also enables automated data analysis by using them within workflows [16]. We also highlight the importance of the process of careful user evaluation, feedback to developers, and optimization of the tools and workflows. Preliminary testing was performed on the command line or GUI versions of the tool. These tools were then packaged into the Galaxy platform, where results were compared to the command-line/GUI versions and also optimized more usage in automated workflows.

Open-source software usage has faced challenges due to dependencies such as operating system (Windows, Linux, OSX), language or platforms (Python, C++, Java), lack of adhering to HUPO standards [46], and installation or usability issues [47]. To overcome these issues, the Galaxy-P project, as well as others in the Galaxy community, have sought collaborations with many research groups that have developed these tools, following a protocol which includes defining key input and output data types, establishing key operating parameters for the Galaxy tool, overcoming operating system compatibility issues (e.g., Singularity containers for Windows tools), along with rigorous testing and optimization. This collaborative and iterative process of development and optimization ensures the software performs accurately and efficiently within the Galaxy platform.

Ideally, software tools that are UNIX based, such as moFF, are easier for deployment within Galaxy. We also demonstrated here that tools such as FlashLFQ could be packaged within a singularity container to enable easy and secure implementation within Galaxy. MaxQuant, which is a popular, public-domain proteomics software package, is available in both Windows and Linux-compatible versions [9]. Although in the early development and testing phase, the LFQ module that uses MS1 precursor intensity data within MaxQuant (MaxLFQ) was made available within Galaxy toolshed [48]. Once fully tested and evaluated, accessibility to this software via the Galaxy platform will offer even more choices for precursor-intensity based quantification. Offering users a choice of multiple validated software tools also highlights the benefits of a workflow engine such as Galaxy, where users can easily develop parallel workflows using different combinations of tools to determine methods that provide optimal results based on user requirements.

In our tests, we observed that FlashLFQ has a faster runtime as compared to the other two tools. MaxQuant processing time is longer, presumably since it performs peptide identification and quantitation simultaneously. For example, on the same computing device, the UPS dataset was processed by FlashLFQ in approximately 15 min for quantification only, whereas MaxQuant and moFF took 34 min and 3 h, respectively. Our evaluation and availability of these tools within a unified platform such as Galaxy offers users a choice for their workflows where the speed of analysis can also be considered.

5. Conclusions

This study demonstrates a successful collaborative effort in software tool development and dissemination, which is a hallmark of the Galaxy community and the Galaxy-P project [19]. This community-driven approach brings together users and software developers who work together to validate and make the tool accessible and usable for other researchers across the world. The study described here provides a model of success for the process used to ultimately provide optimized, well-validated tools for community use. We did not seek a goal to determine the single best tool for LFQ use, but rather focused on offering users a choice of validated quantification tools amenable to customizable analytical workflows. In addition to our work here, others from the Galaxy community are also working on integrating tools within the MaxQuant suite [49], which will extend the choices for LFQ quantification available. As a result of this study, Galaxy users can now confidently use two rigorously validated LFQ software tools (moFF and FlashLFQ) for their quantitative proteomic studies. We are currently working on incorporating the quantitative capabilities of moFF and FlashLFQ within existing metaproteomics and proteogenomics workflows, so that they can be used by the research community in their quantitative multi-omics studies.

Supplementary Materials: The following are available online at http://www.mdpi.com/2227-7382/8/3/15/s1, Figure S1: Peptide Correlation: The raw intensities of the ABRF peptides were correlated using Pearson Correlation coefficient. Output from FlashLFQ and moFF correlated well with MaxQuant. Figure S2: Peptide Overlap across FlashLFQ and moFF: A Venn diagram of the quantified ABRF (a) and UPS (b) unique peptides are shown to display the coverage of the peptides across FlashLFQ and moFF. The input for both the tools was the PSM report from the Peptide shaker containing 18172(ABRF) and 33497 (UPS) unique peptides. Figure S3. (A) Fold-change accuracy (MBR vs no-MBR) of all proteins: After normalization, the estimated protein ratios for all the identified UPS proteins were compared to the true ratios, using the Root mean squared log error (RMSLE). The plot represents the comparison between MBR and no-MBR using normalization methods. Note that LFQ values represent MaxQuant's and FlashLFQ's normalized value. The value on the top of the bars denotes the number of proteins that were quantified. Although the bar graph shows that MaxQuant's MaxLFQ performs the best compared to all, we notice that the number of UPS proteins identified and used for the calculation were less compared to other tools. (B) Fold change accuracy (MBR vs no-MBR) of proteins with similar estimated ratios: In total there are 48 UPS proteins, we classified the UPS proteins into different groups based on the UPS2/UPS1 ratio estimation, the true ratios run from 10 to 10^{-4}. Value on the top of the bars denote the number of proteins that were quantified using each normalization method. This is a comparative study between MBR and no MBR. The RMSLE of the intensity ratio was used to measure the accuracy of the estimated fold change. The figure shows that moFF and FlashLFQ work well in comparison to MaxQuant for ratio estimation of low abundance proteins (<0.1). Figure S4: Fold Accuracy of Tools with inbuilt and external normalization and protein quantitation: This figure displays the errors in Fold change accuracy while using inbuilt and external normalization and protein level quantitation features of the evaluated tools. Supplementary document 2 (https://github.com/galaxyproteomics/quant-tools-analysis) is the GitHub repository of the Rscripts.

Author Contributions: S.M. compared the tools, communicated with software tool developers, performed data analysis, and wrote the manuscript. C.W.E. and R.S. developed the R-scripts and ran statistical data analysis. R.J.M., A.A., and I.E. helped update their quantitation tools according to the feedbacks provided. T.M. and J.E.J. helped package these tools within the Galaxy platform. P.K. provided inputs during manuscript writing. L.M., M.R.S., and L.M.S. provided scientific inputs and helped in manuscript writing. T.J.G. provided scientific directions and helped in manuscript writing. P.D.J. conceived of the project, led the project and helped in manuscript writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Cancer Institute-Informatics Technology for Cancer Research (NCI-ITCR) grant 1U24CA199347 and National Science Foundation (U.S.) grant 1458524 to T.G. We would also like to acknowledge the Extreme Science and Engineering Discovery Environment (XSEDE) research allocation BIO170096 to P.D.J. and use of the Jetstream cloud-based computing resource for scientific computing (https://jetstream-cloud.org/) maintained at Indiana University. The European Galaxy server that was used for some calculations is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)). Part of the work was performed by the Belgian ELIXIR node, also hosting the tools at the Belgian Galaxy instance, which is funded by the Research Foundation, Flanders (FWO) grant I002919N.

Acknowledgments: We would like to thank the European Galaxy team for the help in the support during Galaxy implementation. We would also like to thank Carlo Horro (from Barnes Group, University of Bergen, Norway) and Björn A. Grüning (University of Freiburg, Germany) for helping us during the quantification tools analysis. We thank Emma Leith for proofreading the manuscript. We acknowledge funding for this work from the grant We also acknowledge the support from the Minnesota Supercomputing Institute for the maintenance and update of the Galaxy instances. *Data Availability:* All the data files used for this study are uploaded onto a Zenodo repository at https://doi.org/10.5281/zenodo.3733904. We have also provided the input and output files of our data analysis. Supplementary Document 2 (https://github.com/galaxyproteomics/quant-tools-analysis) is the GitHub repository of the Rscripts. The original dataset for UPS study is available via ProteomeXchange identifier-PXD000279 (spiked-in Universal Proteomic Standard).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cox, J.; Hein, M.Y.; Luber, C.A.; Paron, I.; Nagaraj, N.; Mann, M. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteom.* **2014**, *13*, 2513–2526. [CrossRef] [PubMed]
2. Zhang, C.; Liu, Y. *Retrieving Quantitative Information of Histone PTMs by Mass Spectrometry, In Methods Enzymology*; Academic Press Inc.: Cambridge, MA, USA, 2017; pp. 165–191. [CrossRef]
3. Alfaro, J.A.; Ignatchenko, A.; Ignatchenko, V.; Sinha, A.; Boutros, P.C.; Kislinger, T. Detecting protein variants by mass spectrometry: A comprehensive study in cancer cell-lines. *Genome Med.* **2017**, *9*, 62. [CrossRef] [PubMed]
4. Yeom, J.; Kabir, M.H.; Lim, B.; Ahn, H.S.; Kim, S.Y.; Lee, C. A proteogenomic approach for protein-level evidence of genomic variants in cancer cells. *Sci. Rep.* **2016**, *6*. [CrossRef] [PubMed]
5. Easterly, C.W.; Sajulga, R.; Mehta, S.; Johnson, J.; Kumar, P.; Hubler, S.; Mesuere, B.; Rudney, J.; Griffin, T.J.; Jagtap, P.D. metaQuantome: An Integrated, Quantitative Metaproteomics Approach Reveals Connections Between Taxonomy and Protein Function in Complex Microbiomes. *Mol. Cell. Proteom.* **2019**, *18*, S82–S91. [CrossRef] [PubMed]
6. Krey, J.F.; Wilmarth, P.A.; Shin, J.B.; Klimek, J.; Sherman, N.E.; Jeffery, E.D.; Choi, D.; David, L.L.; Barr-Gillespie, P.G. Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. *J. Proteome Res.* **2014**, *13*, 1034–1044. [CrossRef] [PubMed]
7. Välikangas, T.; Suomi, T.; Elo, L.L. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief. Bioinform.* **2017**, *19*, 1344–1355. [CrossRef]
8. Chawade, A.; Sandin, M.; Teleman, J.; Malmström, J.; Levander, F. Data processing has major impact on the outcome of quantitative label-free LC-MS analysis. *J. Proteome Res.* **2015**, *14*, 676–687. [CrossRef]
9. Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **2016**, *11*, 2301–2319. [CrossRef] [PubMed]
10. Schilling, B.; Rardin, M.J.; MacLean, B.X.; Zawadzka, A.M.; Frewen, B.E.; Cusack, M.P.; Sorensen, D.J.; Bereman, M.S.; Jing, E.; Wu, C.C.; et al. Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: Application to protein acetylation and phosphorylation. *Mol. Cell. Proteom.* **2012**. [CrossRef]
11. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337–2342. [CrossRef]
12. Progenesis QI for Proteomics: Waters. Available online: https://www.waters.com/waters/en_US/Progenesis-QI-for-Proteomics/nav.htm?locale=en_US&cid=134790665 (accessed on 8 October 2019).

13. Argentini, A.; Goeminne, L.J.E.; Verheggen, K.; Hulstaert, N.; Staes, A.; Clement, L.; Martens, L. MoFF: A robust and automated approach to extract peptide ion intensities. *Nat. Methods* **2016**, *13*, 964–966. [CrossRef]
14. Millikin, R.J.; Solntsev, S.K.; Shortreed, M.R.; Smith, L.M. Ultrafast Peptide Label-Free Quantification with FlashLFQ. *J. Proteome Res.* **2018**, *17*, 386–391. [CrossRef] [PubMed]
15. Van Riper, S.; Chen, E.; Remmer, H.; Chien, A.; Stemmer, P.; Wang, Y.; Jagtap, P. Identification of Low Abundance Proteins in a Highly Complex Protein Mixture. 2016. Available online: https://zenodo.org/record/3563207#.XwRcuhJ5t2E (accessed on 7 July 2020). [CrossRef]
16. Afgan, E.; Baker, D.; van den Beek, M.; Blankenberg, D.; Bouvier, D.; Čech, M.; Chilton, J.; Clements, D.; Coraor, N.; Eberhard, C.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **2016**, *44*, W3–W10. [CrossRef] [PubMed]
17. Jagtap, P.D.; Johnson, J.E.; Onsongo, G.; Sadler, F.W.; Murray, K.; Wang, Y.; Shenykman, G.M.; Bandhakavi, S.; Smith, L.M.; Griffin, T.J. Flexible and accessible workflows for improved proteogenomic analysis using the galaxy framework. *J. Proteome Res.* **2014**, *13*, 5898–5908. [CrossRef]
18. Anderson, K.J.; Vermillion, K.L.; Jagtap, P.; Johnson, J.E.; Griffin, T.J.; Andrews, M.T. Proteogenomic Analysis of a Hibernating Mammal Indicates Contribution of Skeletal Muscle Physiology to the Hibernation Phenotype. *J. Proteome Res.* **2016**, *15*, 1253–1261. [CrossRef]
19. Blank, C.; Easterly, C.; Gruening, B.; Johnson, J.; Kolmeder, C.A.; Kumar, P.; May, D.; Mehta, S.; Mesuere, B.; Brown, Z.; et al. Disseminating metaproteomic informatics capabilities and knowledge using the galaxy-P framework. *Proteomes* **2018**, *6*, 7. [CrossRef] [PubMed]
20. Jagtap, P.D.; Blakely, A.; Murray, K.; Stewart, S.; Kooren, J.; Johnson, J.E.; Rhodus, N.L.; Rudney, J.; Griffin, T.J. Metaproteomic analysis using the Galaxy framework. *Proteomics* **2015**, *15*, 3553–3565. [CrossRef]
21. Proteomics moFF Galaxy Toolshed Repository. Available online: https://toolshed.g2.bx.psu.edu/view/galaxyp/proteomics_moff/7af419c90f5f (accessed on 7 July 2020).
22. Proteomics FlashLFQ Galaxy Toolshed Repository. Available online: https://toolshed.g2.bx.psu.edu/view/galaxyp/flashlfq/908ab13490dc (accessed on 7 July 2020).
23. Adusumilli, R.; Mallick, P. Data conversion with proteoWizard msConvert. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2017; pp. 339–368. [CrossRef]
24. Barsnes, H.; Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **2018**, *17*, 2552–2555. [CrossRef] [PubMed]
25. Vaudel, M.; Burkhart, J.M.; Zahedi, R.P.; Oveland, E.; Berven, F.S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets: To the editor. *Nat. Biotechnol.* **2015**, *33*, 22–24. [CrossRef]
26. Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R.A.; Olsen, J.V.; Mann, M. Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **2011**, *10*, 1794–1805. [CrossRef]
27. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Linear Models for Microarray and RNA-Seq Data. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]
28. Suomi, T.; Corthals, G.L.; Nevalainen, O.S.; Elo, L.L. Using peptide-level proteomics data for detecting differentially expressed proteins. *J. Proteome Res.* **2015**, *14*, 4564–4570. [CrossRef] [PubMed]
29. Välikangas, T.; Suomi, T.; Elo, L.L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief. Bioinform.* **2018**, *19*, 1–11. [CrossRef]
30. Argentini, A.; Staes, A.; Grüning, B.; Mehta, S.; Easterly, C.; Griffin, T.J.; Jagtap, P.; Impens, F.; Martens, L. Update on the moFF Algorithm for Label-Free Quantitative Proteomics. *J. Proteome Res.* **2019**, *18*, 728–731. [CrossRef]
31. Dale, R.; Grüning, B.; Sjödin, A.; Rowe, J.; Chapman, B.A.; Tomkins-Tinch, C.H.; Valieris, R.; Batut, B.; Caprez, A.; Cokelaer, T.; et al. Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **2018**, *15*, 475–476. [CrossRef]
32. Proteomics moFF GitHub Repository. Available online: https://github.com/compomics/moFF (accessed on 7 July 2020).
33. Luersen, M.A.; le Riche, R.; Guyon, F. A constrained, globalized, and bounded Nelder-Mead method for engineering optimization. *Struct. Multidiscip. Optim.* **2004**, *27*, 43–54. [CrossRef]
34. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **2017**, *12*, e0177459. [CrossRef] [PubMed]

35. Proteomics FlashLFQ GitHub Repository. Available online: https://github.com/smith-chem-wisc/FlashLFQ (accessed on 7 July 2020).
36. Walther, B.A.; Moore, J.L. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **2005**, *28*, 815–829. [CrossRef]
37. Bensimon, A.; Heck, A.J.R.; Aebersold, R. Mass Spectrometry–Based Proteomics and Network Biology. *Annu. Rev. Biochem.* **2012**, *81*, 379–405. [CrossRef]
38. Vermillion, K.L.; Jagtap, P.; Johnson, J.E.; Griffin, T.J.; Andrews, M.T. Characterizing cardiac molecular mechanisms of mammalian hibernation via quantitative proteogenomics. *J. Proteome Res.* **2015**, *14*, 4792–4804. [CrossRef]
39. Ori, A.; Toyama, B.H.; Harris, M.S.; Bock, T.; Iskar, M.; Bork, P.; Ingolia, N.T.; Hetzer, M.W.; Beck, M. Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats. *Cell Syst.* **2015**, *1*, 224–237. [CrossRef]
40. Roumeliotis, T.I.; Williams, S.P.; Gonçalves, E.; Alsinet, C.; Velasco-Herrera, M.D.; Aben, N.; Ghavidel, F.Z.; Michaut, M.; Schubert, M.; Price, S.; et al. Genomic Determinants of Protein Abundance Variation in Colorectal Cancer Cells. *Cell Rep.* **2017**, *20*, 2201–2214. [CrossRef]
41. Schlaffner, C.N.; Pirklbauer, G.J.; Bender, A.; Steen, J.A.J.; Choudhary, J.S. A fast and quantitative method for post-translational modification and variant enabled mapping of peptides to genomes. *J. Vis. Exp.* **2018**. [CrossRef]
42. Gans, J.; Wolinsky, M.; Dunbar, J. Microbiology: Computational improvements reveal great bacterial diversity and high toxicity in soil. *Science* **2005**, *309*, 1387–1390. [CrossRef] [PubMed]
43. Li, X.; LeBlanc, J.; Truong, A.; Vuthoori, R.; Chen, S.S.; Lustgarten, J.L.; Roth, B.; Allard, J.; Ippoliti, A.; Presley, L.L.; et al. A metaproteomic approach to study human-microbial ecosystems at the mucosal luminal interface. *PLoS ONE* **2011**, *6*. [CrossRef]
44. Cheng, K.; Ning, Z.; Zhang, X.; Li, L.; Liao, B.; Mayne, J.; Stintzi, A.; Figeys, D. MetaLab: An automated pipeline for metaproteomic data analysis. *Microbiome* **2017**, *5*, 157. [CrossRef] [PubMed]
45. Chambers, M.C.; Jagtap, P.D.; Johnson, J.E.; McGowan, T.; Kumar, P.; Onsongo, G.; Guerrero, C.R.; Barsnes, H.; Vaudel, M.; Martens, L.; et al. An Accessible Proteogenomics Informatics Resource for Cancer Researchers. *Cancer Res.* **2017**, *77*, e43–e46. [CrossRef]
46. Mayer, G.; Montecchi-Palazzi, L.; Ovelleiro, D.; Jones, A.R.; Binz, P.A.; Deutsch, E.W.; Chambers, M.; Kallhardt, M.; Levander, F.; Shofstahl, J.; et al. The HUPO proteomics standards initiativemass spectrometry controlled vocabulary. *Database* **2013**. [CrossRef] [PubMed]
47. Smith, R. Conversations with 100 Scientists in the Field Reveal a Bifurcated Perception of the State of Mass Spectrometry Software. *J. Proteome Res.* **2018**, *17*, 1335–1339. [CrossRef]
48. Proteomics MaxQuant Galaxy Toolshed Repository. Available online: https://toolshed.g2.bx.psu.edu/view/galaxyp/maxquant/175e062b6a17 (accessed on 7 July 2020).
49. Proteomics MaxQuant GitHub Repository. Available online: https://github.com/galaxyproteomics/tools-galaxyp/tree/master/tools/maxquant (accessed on 7 July 2020).