Triqler for Data Independent Aquisition Data

Patrick Truong Matthew The Lukas Käll

August 20, 2021

Abstract

In this study we show that Triqler, a protein quantification and differential analysis tool based on probabilistical graphical models, has better performance than other protein quantification tools. To show this we compare different processing pipelines using different underlying concept for protein idenfication and quantification...

Introduction

Label-free quantification (LFQ) using Mass spectrometry (MS) based proteomics has been shown to be an effective methods for studying the relative concentration of proteins in complex mixtures. Compared to Data-dependent axquisition (DDA), Data-independent acquisition (DIA) mass spectrometry allows for a broader dynamical range and more reproducible peptide detection [**?** **?**].

Triqler is a novel software that uses a probabilitical graphical model for protein quantification and differential expression analysis, essentially eliminating the need for filtering, tresholding and imputational procedures required by many conventional methods. Triqler has been shown to distinguish more proteins for DDA data compared to other DDA protein quantification methods. [**?**].

Materials and methods

Data description

The data is a DIA dataset used in a previous benchmark of DIA protein quantification benchmarking study [LFQBenchPaper2016]. It is available from the ProteomeXchange Consortium with the dataset identifier PXD002952. The instrumentation used process the data was TTOF6600 system with 32 fixed windows. In the repository the data we use is referred to as the HYE124 hybrid proteome samples. It consists of tryptic peptides with the following ratios: Sample A composed of 65% w/w, 30% w/w yeast, and 5% w/w E. coli proteins. Sample B was composed of 65% w/w, 15% w/w yeast and 20% w/w E. coli proteins. Further details about mass spectrometric instrumentation and data acquisition is available in Navarro et al. [LFQBenchPape2016].

General workflow In this study we investigate two DIA protein quantification workflows 1) DDA-based spectral library protein quantification (DDA-SLPQ) and 2) Pseudo spectra-based spectral library protein quantification (PS-SLPQ). The DDA-SLPQ workflow uses MSFragger to perform a DDA-search and EasyPQP to contruct a spectra library from the DDA-search results. OpenSwath Workflow is then used with spectral library to perform an analysis on the the DIA-data. PyProphet is used for statistical validation and TRIC is used for feature alignment for before protein quantification using Top3, MSstats and msqrobsum (TRIC is not used for triqler protein quantification). The PS-SLPQ workflow is conducted using fragpipe software. It uses DIA-Umpire signal extraction to extract pseudo-spectra from the DIA-data. The pesudo-spectra is searched using conventional DDA-search using MSFragger and spectral library is build using easyPQP. The DIA-data is quantified with DIA-NN using the peudo-spectra based spectral libraries. The DIA-NN uses a mProphet-based algorithm for statistical validation [**?**] [**?**].

(add a third later encyclopedia)

(add workflow schematics)

Data preparation and spectral library generation The .wiff files are converted to .mzML files in centroided format using msconvert (using windows OS msconver version X.X) with the following options: [check options].

Two approaches was used for spectra library generation: DDA acquisition based spectral library generation and Prosit-based spectral library generation using only .fasta file [cite prosit paper].

DDA acquisitions of samples from each specie (human, yeast, E. coli) was provided in triplicates for spectral library generation. Uniprot fasta files with one protein seqeunce per gene was concatenated for each specie (UP000005640, UP000000625 and UP000002311, acquired on 2021-06-16).To control for the effect of different protein inference strategies (protein group, parsimony etc.) a modified .fasta file, without shared peptides, was used for database search. The unfiltered fasta files contained 20 590 human proteins, 6 046 yeast proteins and 4 373 E. coli proteins. After filtering the fasta file contained 20 302 proteins (-288 human proteins), 5 848 yeast proteins (198 yeast proteins) and 4 306 E. Coli proteins (-67 E. Coli proteins). Each sequence with length >7 amino acids mapping only to one protein. The fasta file contained reverse sequences as decoys for target-decoy analysis. MSFragger with parameters: [check parameters] was used for DDA-search, statistical validation was performed by peptide prophet and protein prophet, and EasyPQP with parameters: [check parameters] was used for spectral library building. OpenSwathDecoyGenerator was used with default setting to generate decoys for the resulting spectral libraries.

For Prosit-based spectral library generation, the fasta file was converted to prosit input format using encyclopeDIA converter. Prosit_2020_intensity_cid model was used as intensity prediction model and Prosit_2019_irt was used as iRT predicition model.

Protein inference problem and reduced database

In bottom-up proteomics, redundency in the matching of peptides to proteins hits is a challenge [?]. The Parsimony principle reports the minimum set of proteins which include all the observed peptides (Occam's razor principle), thereby resolving shared and ambiguous evidence. This method typically reduces the size of the protein list where peptides could belong to several proteins, as the cost of losing potential protein idenfications. In addition it may also lower the repoducibility of the identified proteins[?]. While the protein grouping approach bunch together proteins based on different schemas. For example it is possible to group together proteins that map to identical peptides, or when proteins contain overlapping subsets of peptides. The protein groupings are often assigned post peptide identification, which contrary to good statistical practices to formulate the hypothesis before data are observed [?]. An example of parsimony and protein grouping is shown in **table 1**.

	Peptides					Parsimony reported	Grouping reported
	1	2	3	4	5		
Protein A	x	x				Yes	Yes
Protein B		x	x			No	Yes
Protein C			x	x		Yes	Yes

Table 1: Example of Parsimony reported and protein grouping reported proteins.

In this benchmarking stude we remove proteins with shared peptides. Using different methods for protein inference yields different amount of protein matches given the same amount of observed peptides. Ideally, the number of proteins compared between different protein quantitation methods should be the same when benchmarking. By removing proteins with shared peptides from the FASTA-database the protein inferences procedures will yield the same protein identifications. This should give a fair comparison regardless of protein inference method used in the different protein quantitation pipelines.

DIA methods Write about why different methods where used.

OpenSwath Analysis Version (version) of OpenSwath was used. The spectral library generated above is converted to .pqp format using TargetedFileConverter. Data analysis was conducted using OpenSwathWorkflow with parameters (-Scoring:TransitionGroupPicker:background_subtraction original -Scoring:stop_report_after_feature -1, -min_upper_edge_dist 1, -tr_irt hroest_DIA_iRT.TraML, -extra_rt_extraction_window 100, -min_rsq 0.95, -min_coverage 0.6, -Scoring:Scores:use_dia_scores true, -rt_extraction_window 600, -mz_extraction_window 30, -threads 10, -Scoring:DIAScoring:dia_extraction_unit ppm). After data extraction the data .osw output was merged using pyprophet merge option and pyprophet was used for statistical validation. Pyprophet export was used without FDR filtering (–max_global_protein_qvalue 1.0, –max_global_peptide_qvalue 1.0, –max_rs_peakgroup_qvalue 1.0, –max_transition_pep 1.0) to give a complete list of peptide quantifications for further down-stream analysis.

DIAUmpire and DIA-NN analysis DIAUmpire signal extraction (SE) was used through Fragpipe GUI (v15.0). Default parameters was set (MS1 PPM: 10, MS2 PPM: 20, Max Missed Scans:1, Mass Defect Filter On, RP max: 25, RF max: 500, Corr Treshold:0, Delta Apex: 0.2, RT Overlap 0.3, Mass Defect Offset 0.1, Isotope Pattern: 0.3, MS1 SN: 1.1, MS2 SN 1.1, Adjust fragment intensity On). MSFragger was used on the resulting .mzML files from DIAUmpire SE with default parameters for Peak Matching (PPM: [-20, 20], Fragment mass tolerance PPM: 20, Calibration and Optimization: Mass Calibration, Parameter optimization, Isotop error: 0/1, Data type: DDA,), protein digestion (Load rules: stricttrypsin, Enzyme name: stricttrypsin, Cut after: KR, Cleavage: ENZYMATIC, Missed cleavages: 2, Clip N-term M: On, Peptide length 7-50, Peptide mass range: 500-5000, Split database: 1) and Modification (Variable modifications: M, [^, Fixed modification: ”all selected”).

EncyclopeDIA and PECAN analysis Prosit was used to contruct spectra libraries from the modified fasta files. The fragmentation model used was ”Prosit - Model - Fragmentation” and the iRT model ”Prosit - Model - iRT” (available from https://figshare.com/projects/Prosit/35582).

...

Protein quantification methods

Triqler

Triqler is based on probabilistic graphical models that allow error information to propagate through all steps from MS1 feature detection to protein quantification. Unconventionally, it produces posterior probabilities for fold changes rather than point estimates [**?**]. The robustness of results when faulty imputation of missing data is conducted is a common problem in missing data analysis [**?**]. These posteriors incorporate information about uncertainty and thesefore should result in robust protein quantification.

Triqler was used with –fold_change_eval between 0-2 with 0.04 increments. It computes the two-sided differential probability treshold between the two samples given a fold change evaluation limit.

Top3 The precursors are filtered by q-value ¿ 1% and the average of the three largest peptide intensities are taken for each protein. Protein with only one detected peptides (single hit proteins) and proteins detected only in two injections are discarded.

MSstat MSstats customizes a linear mixed model to the specific experiment and to each protein choi2014msstats. SWATH2Stats was used to convert the osw output files to MSStats input format. dataProcess function in the MSstats package is used for data pre-processing and quality control of the MS runs of the data into quantitative data for model fitting and group comparison (Default parameters was used (specifically the peptide-level data is filtered by an m-score ¡ 0.01 to reduce the memory consumption before running MSstats). quantification function in the MSstats package is applied on the pre-processed data to generate the quantification results for each protein.

Msqrobsum Msqrobsum

(Note: write something about difference FDR computation method and conservativeness, and why it does not make sense to compare between processing method, but it is ok to compare between post-processing methods)

Benchmarking methods

Variance Structure - why the DDA methods works and makes sense The field of protein quantification is relatively new. Many protein quantification methods have been developed specifically for DDA data. Many of these methods should be generalizable for DIA data. Triqler (probabilistic graphical model), MsStats (linear mixed model) and MSqRob (Ridge Regression) all require homoskedastic variance. This is investigated using statistical visualisation of the data.

Protein level results - scatter plot

Differential abundance - log2distribution and number of differentially expressed proteins The differential expressed proteins are proteins with a log2-fold change of proteins with a q-value lower than a certain treshold for q-value based protein summarization methods (Top3, MSstats and msqrob), while triqler computes a posterior distribution based log2-fold change.

(Write formula for triqler) (Write formula for other methods)

FDR control (calibration plot) A FDR control of the quantification methods is required for benchmarking. We use the False Discovery Proportion (FDP) and True Positive Rate (TPR). FDP is calculated using and calculated using

$$FDP = \frac{\text{false positives}}{\text{true positive} + \text{false positive}}$$

and,

$$TPR = \frac{\text{true positives}}{\text{all positives}}$$

The FDP is calculated using the fraction of human proteins discovered as DE and the TP is the fraction of E.Coli and Yeast discovered.

Results

Variance structure (Write why different protein quantification methods and the why triqler should be better in theory.)

In fig. it is show that the variance-to-mean intensity ratio does not increase for larger means. As the variance does not increase with mean heteroskedasticity does not exist in the DIA-data. This implied that methods based on linear regression, linear mixed models and other models requiring stable variance structure could be used for DIA protein quantification.

(Write something about implication of this) (I COULD TRY TO SHOW THIS WITH QQ PLOT INSTEAD).

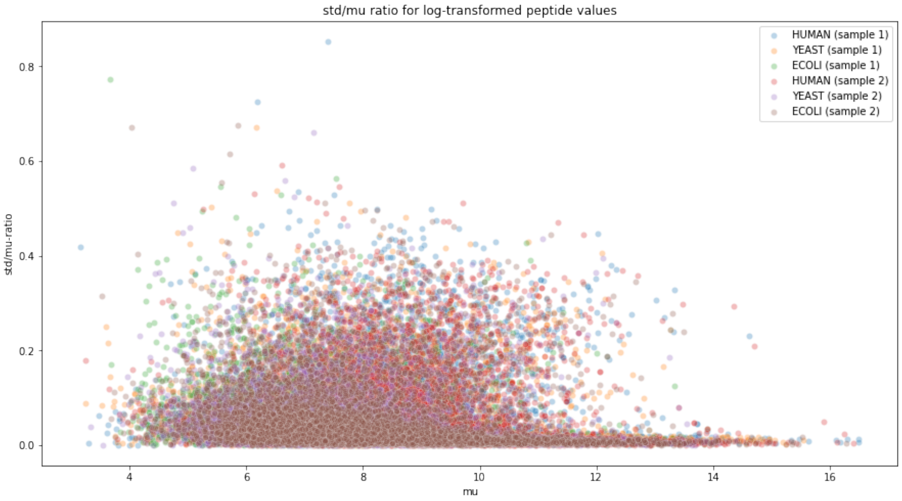

Figure 1: mean vs. variance-to-mean ratios. The variance-to-mean ratios stays similar across the means.

Protein quantification

Log-transformed ratios (log2(A//B)) of proteins (humans proteins in orange, yeast proteins in green and ecoli proteins in blue) were plotted for each benchmarked software over the log-transformed intensity of sample A. The dashed horizontal line represent the expected log-fold change, while the fitted dashed line represent a linear regression for each species. Fig. A shows the results for OpenSwath and fig. B show the results for DIA-NN.

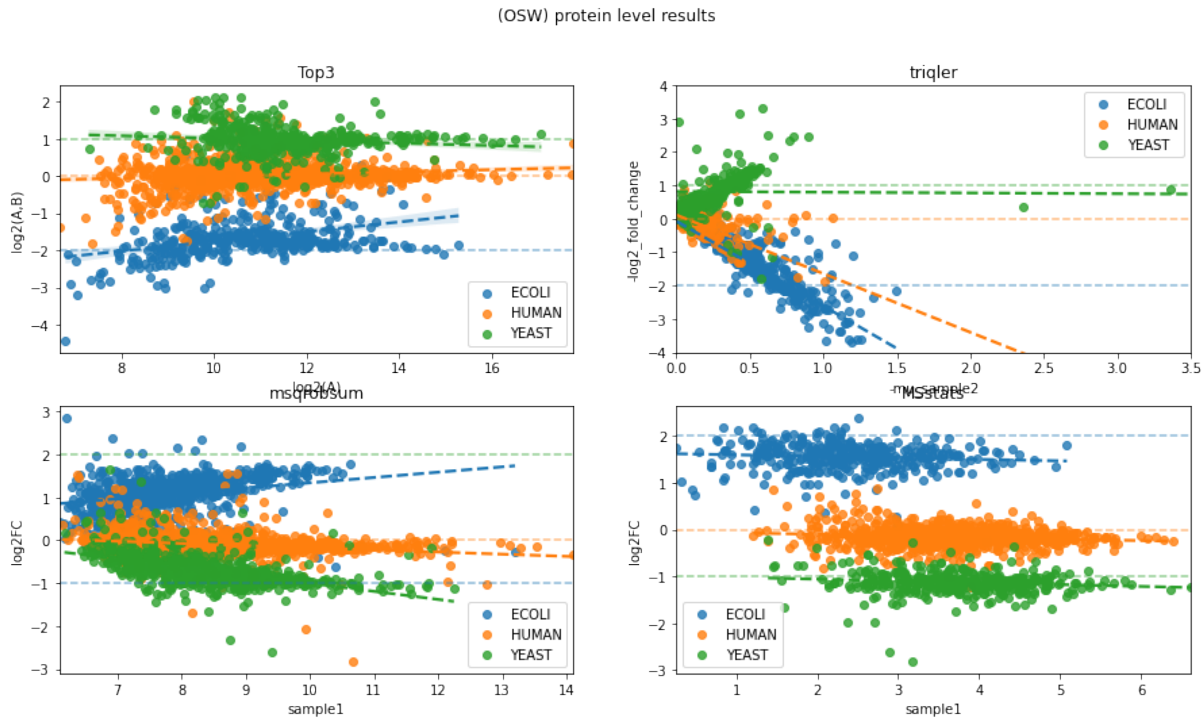

Figure 2: OSW Protein level results from Triqler, Top3, MSstats and MSqRobSum.

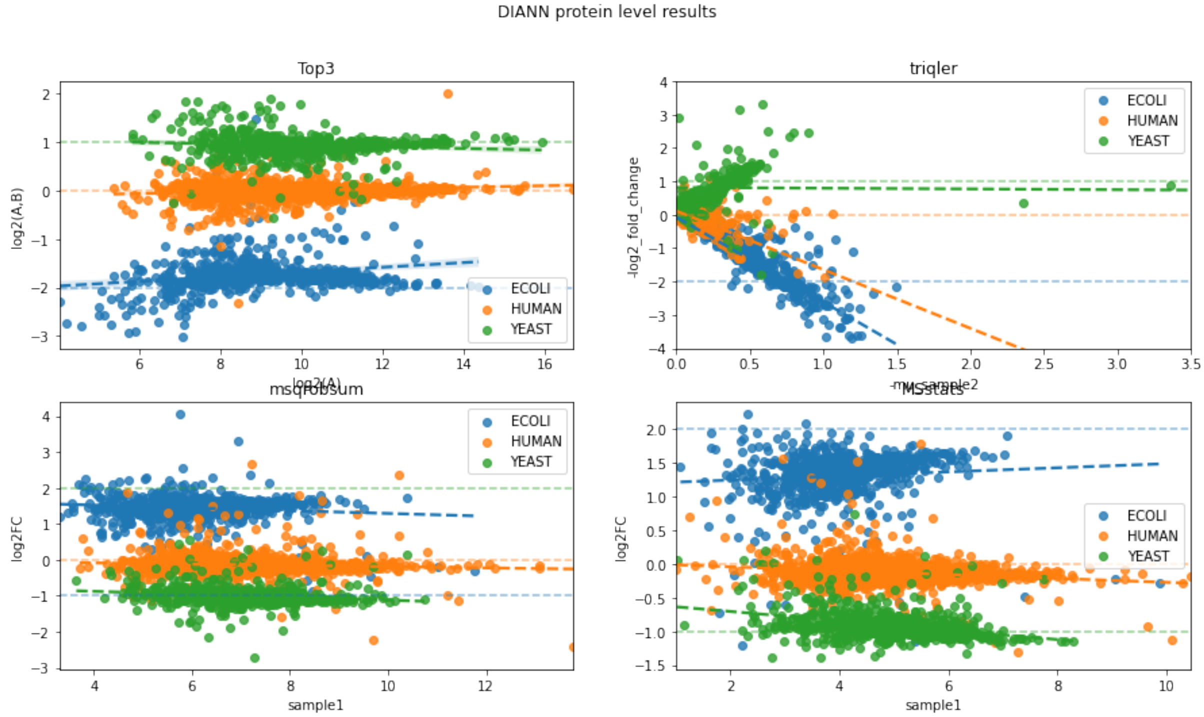

Figure 3: DIANN Protein level results from Triqler, Top3, MSstats and MSqRobSum.

Fig. X show the protein quantity distributions with the different species; human (orange), ecoli (blue) and yeast (green) highlighted in different colours. The human distribution apex is not centered at 0 in MsStats and MSqRob, while it is centered for triqler and top3. Likewise the distribution for e.coli and yeast is centered towards the true log2 fold change ratios for triqler and top3, while the apexes are skewed towards 0 for MsStats and MSqRob.

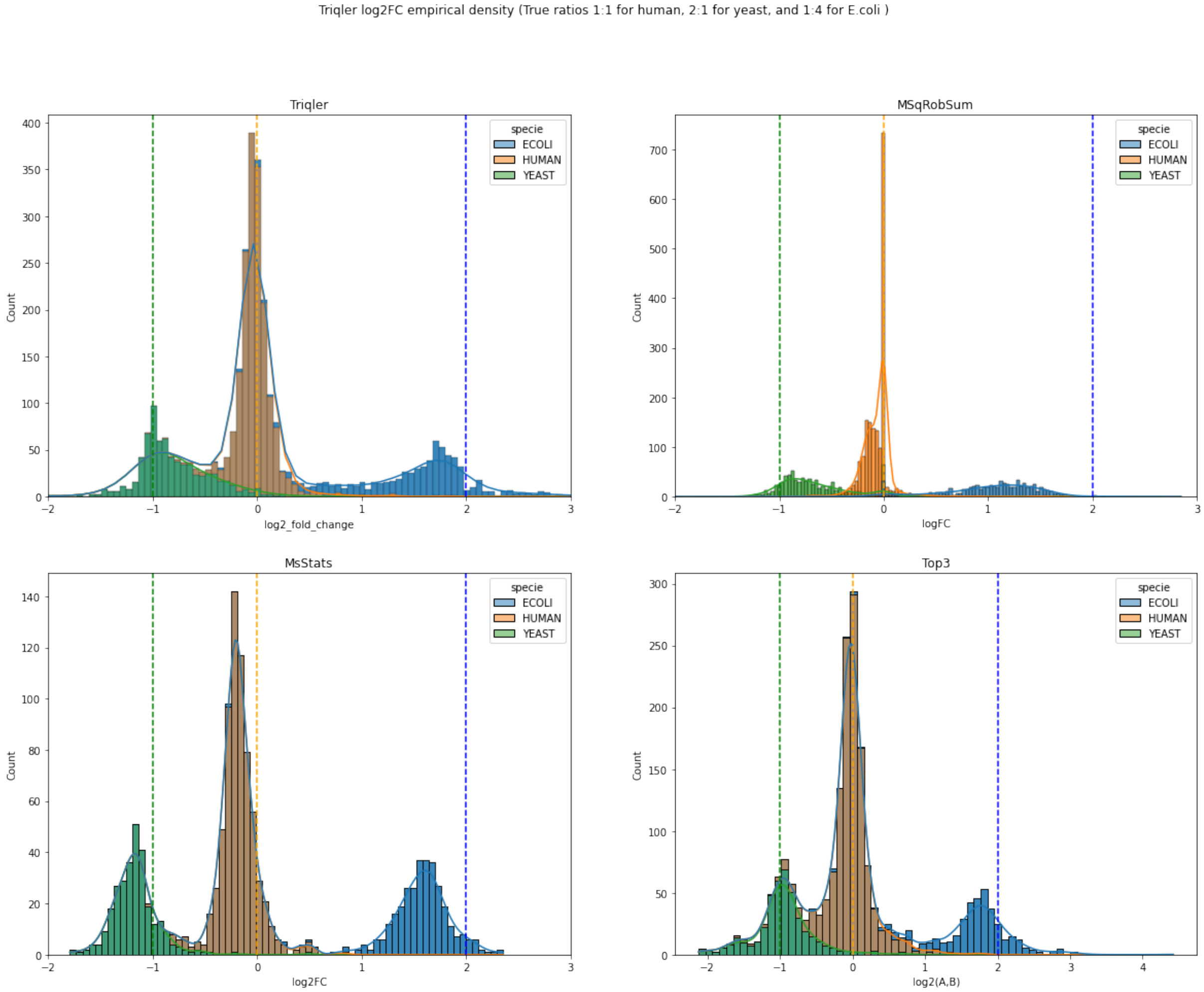

Figure 4: OSW Protein quantity distributions from Triqler, Top3, MSstats and MSqRobSum.

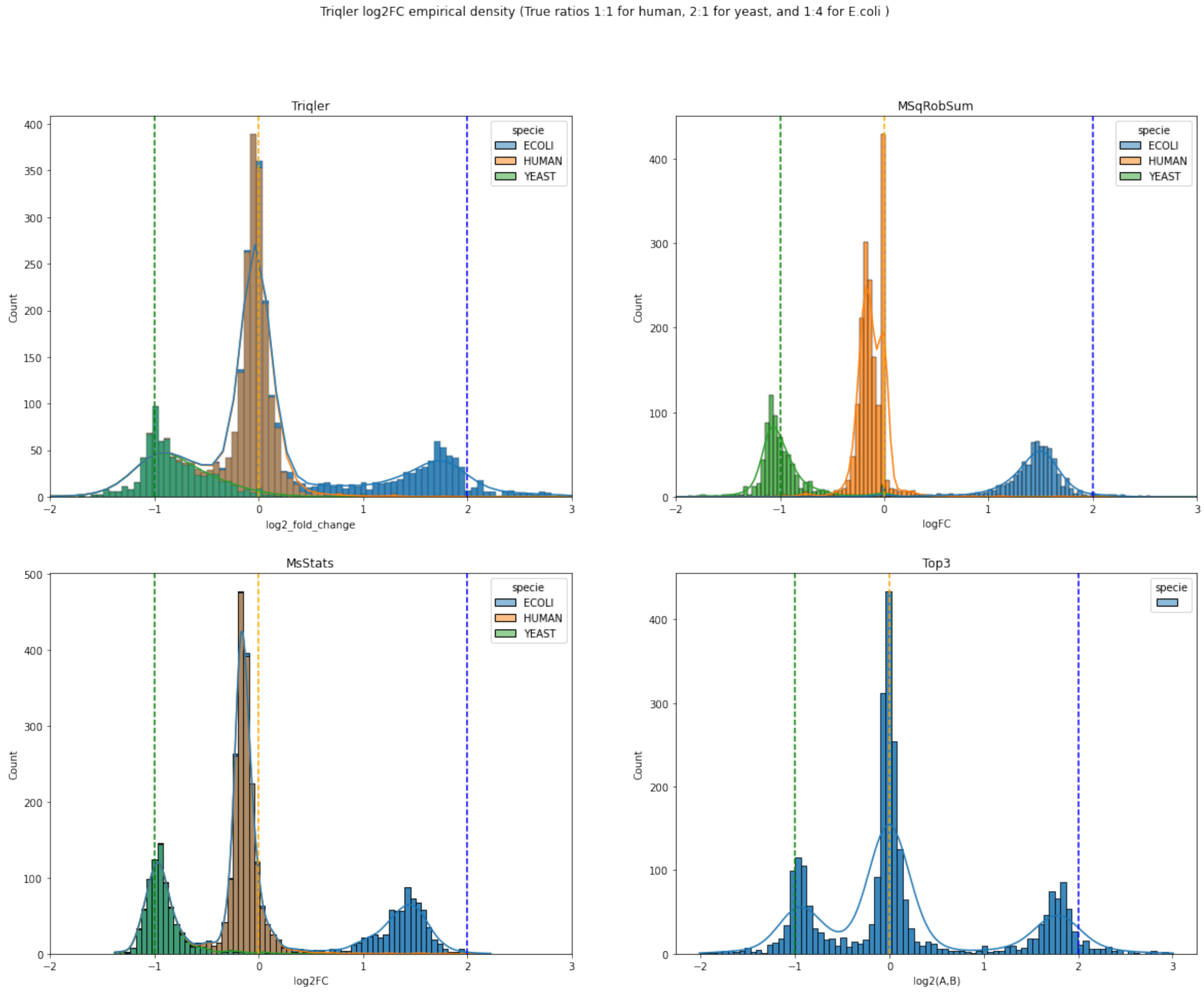

Figure 5: DIANN Protein quantity distributions from Triqler, Top3, MSstats and MSqRobSum.

Fig. Y. show that the number of differentially expressed proteins are significantly higher for Triqler and MSqRobSum at different false discovery rates. MSqRobSum has slightly higher number of differentially expressed proteins than triqler.

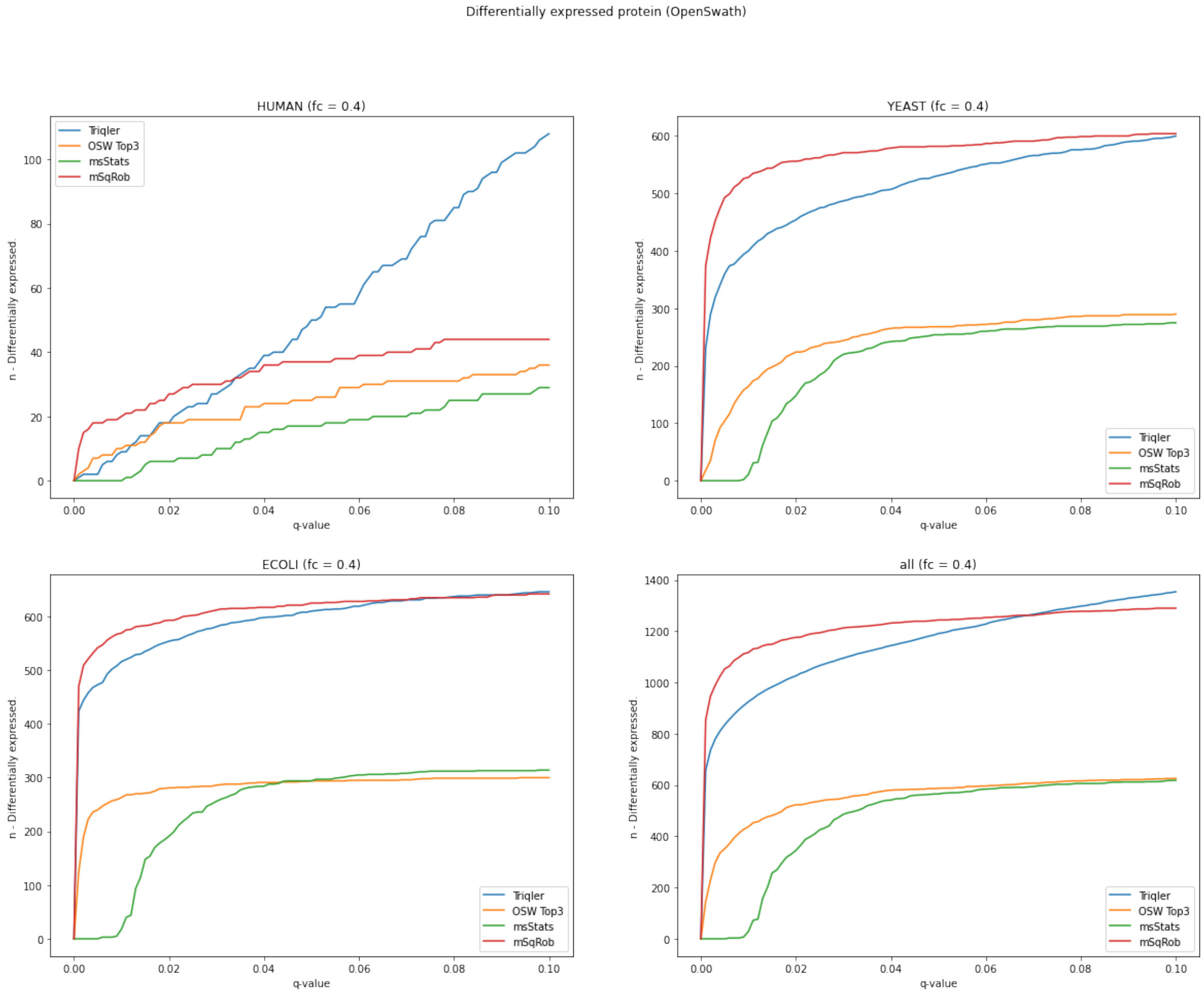

Figure 6: OSW Number of differentially expressed proteins for an OpenSwath analysis with triqler, top3, MsStats and MSqRobSum protein summarization.

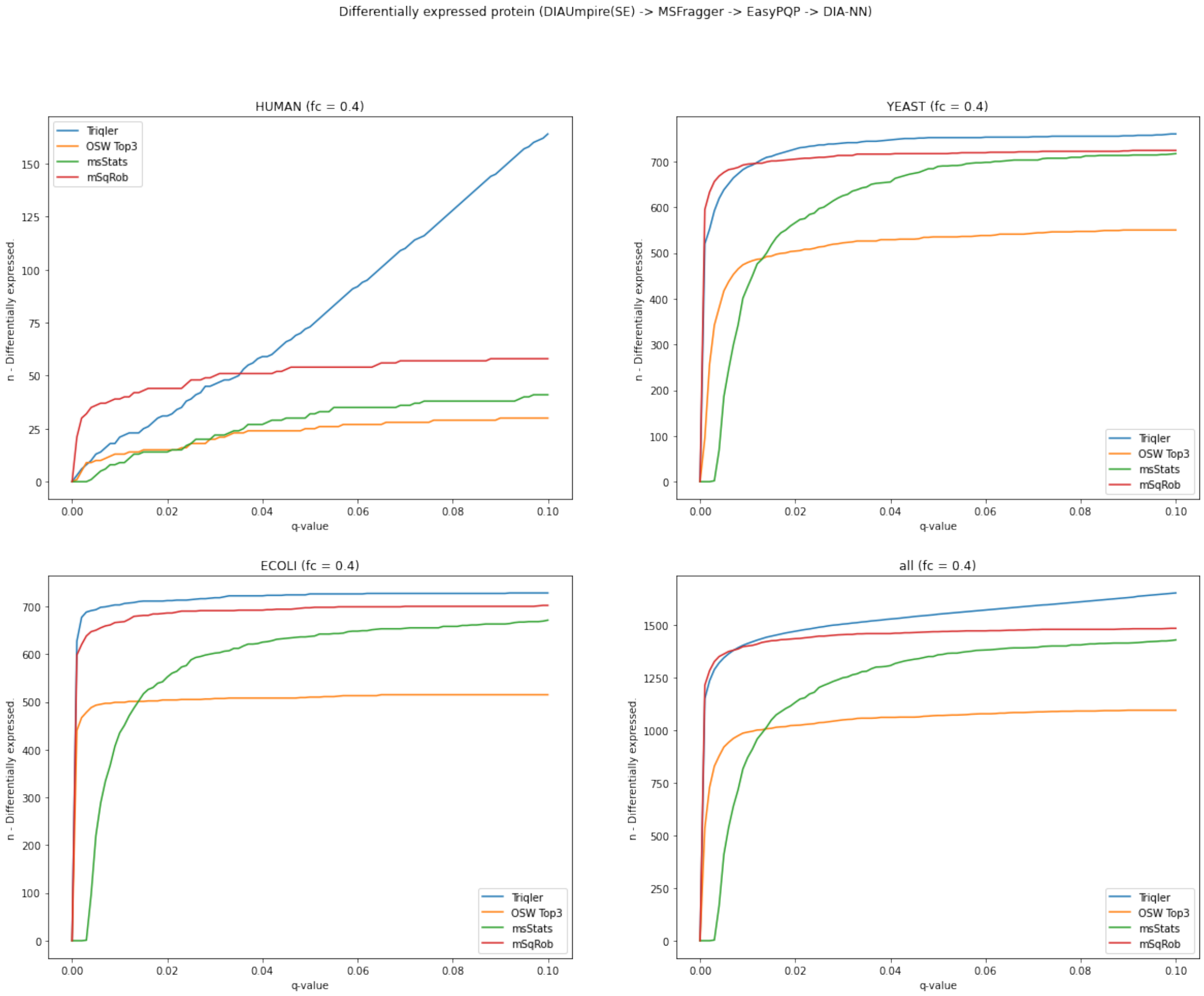

Figure 7: DIANN Number of differentially expressed proteins for an OpenSwath analysis with triqler, top3, MsStats and MSqRobSum protein summarization.

However, fig Y shows that the ratio of false positives (human proteins) to number of differentially expressed proteins for a given q-value level is more linear for triqler. At log2FC = 2 all methods are conservative at low q-values. Triqler is better calibrated for low q-values and gets more convervative as the q-value treshold is increased. Top3, MsStats and MSqRob are more conservative at low q-values and gets less convervative as the q-value treshold increases.

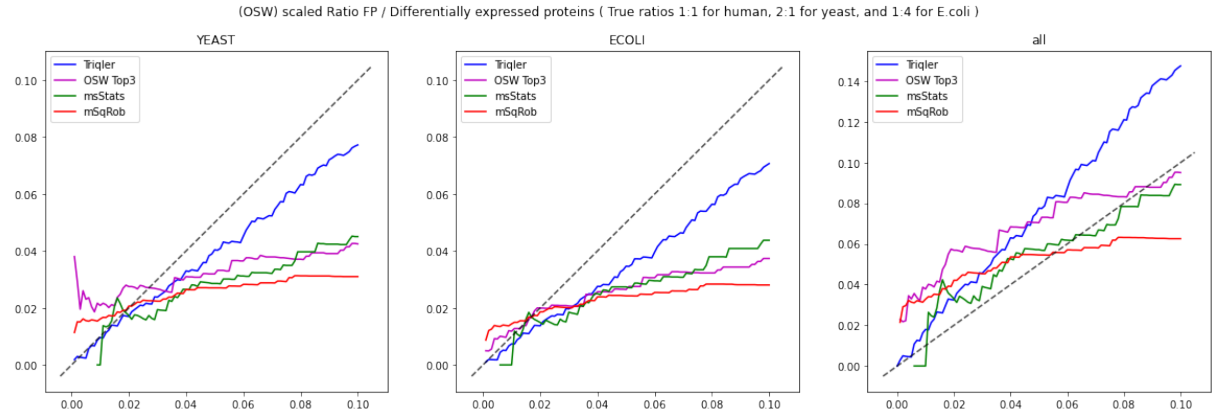

Figure 8: OSW Q-value (x-axis) and false positive / number differentially expressed protein (y-axis) ratio.

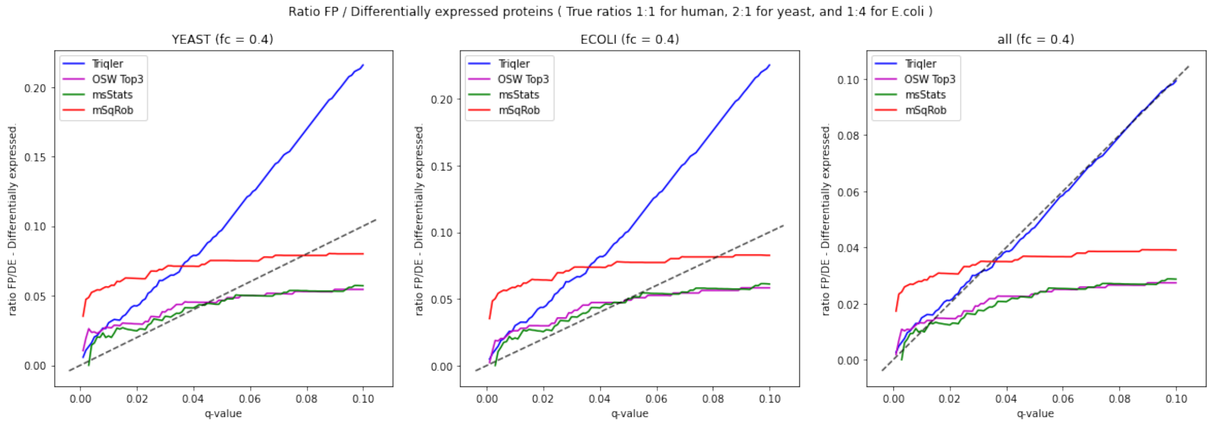

Figure 9: DIANN Q-value (x-axis) and false positive / number differentially expressed protein (y-axis) ratio.

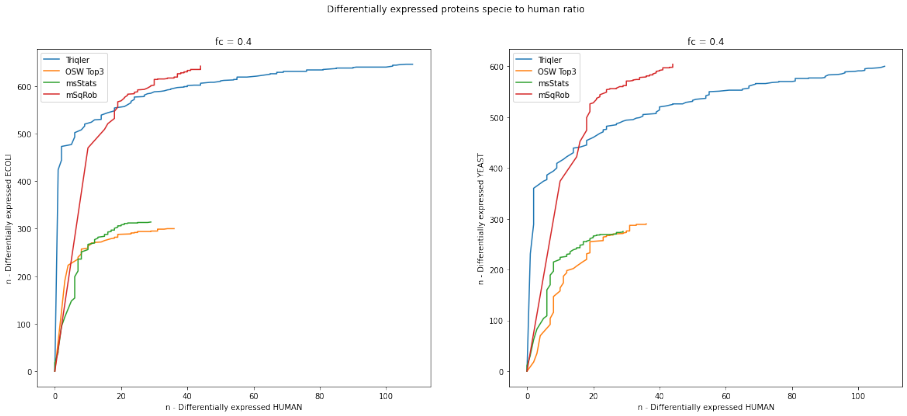

Figure 10: OSW Number of differentially expressed humans (x-axis) and differentially expressed e.coli and yeast (y-axis).

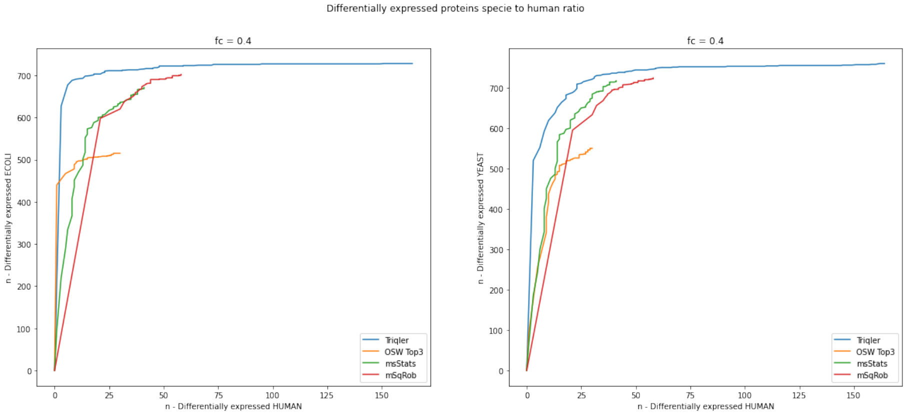

Figure 11: DIANN Number of differentially expressed humans (x-axis) and differentially expressed e.coli and yeast (y-axis).

Discussion

Acknowledgements

Funding

This work has been supported by a grant from the Swedish Foundation for Strategic Research (BD15-0043).

Supporting information