thermo scientific

POSTER NOTE

No. 64857

New Method for Label-Free Quantification in the Proteome Discoverer Framework

David M Horn,[1] Torsten Ueckert,[2] Kai Fritzemeier,[2] Katja Tham,[2] Carmen Paschke,[2] Frank Berg,[2] Hans Pfaff,[2] Xiaoyue Jiang,[1] Shijun Li,[1] and Dani Lopez-Ferrer,[1]
[1]Thermo Fisher Scientific, San Jose, CA, USA; [2]Thermo Fisher Scientific, Bremen, Germany

ABSTRACT

A new label-free quantification method based on the Minora algorithm is presented and compared to pre-existing label free quantification methods in the Thermo Scientific™ Proteome Discoverer™ software framework. The results of the new algorithm were significantly more accurate across a wide dynamic range compared to spectral counting and "Top N" quantification. The new algorithm was also run on a subset of the Akhilesh Pandey human proteome dataset to identify proteins specific to specific tissue types.

INTRODUCTION

Proteome Discoverer software is a node-based workflow engine and study management platform for analysis of mass spectrometry-based proteomics datasets. The latest released version 2.1 fully supports isotopically-labeled quantitative workflows, such as TMT™ reporter ion-based quantification and SILAC precursor ion quantification, but the supported label-free quantification methods are significantly less sophisticated. Currently, spectral counting is possible but not recommended when quantitative accuracy is required. The only supported label-free quantification workflow produces an average abundance of the top "N" most abundant peptides and this has been shown to be accurate for even highly complex datasets. However, "Top N" quantification results cannot be used to create ratios, scaled abundance values, or to be used as replicates to generate standard errors. Here we present a new workflow for untargeted label-free quantification using a new feature detection approach that provides the full suite of quantitative capabilities previously only available for isotopically-labeled quantification. The workflow will be compared to the two aforementioned label-free quantification workflows available within Proteome Discoverer 2.1 software.

MATERIALS AND METHODS

A standard dataset of Arabidopsis proteasome proteins spiked into a background of *E. coli* proteins (PXD003002) was downloaded from the PRoteomics IDEntifications (PRIDE) repository. This dataset was originally used to evaluate a spectral counting algorithm and is described in reference 1. The Pandey human proteome dataset[2] was also downloaded from PRIDE and a portions of the dataset to demonstrate untargeted label free quantification of data with a multi-dimensional separation.

For quantification using spectral counts, each of the datasets with the different levels of *Arabidopsis* proteasome proteins was run separately in batch mode using a standard Sequest™ HT-Percolator workflow and a basic consensus workflow. Subsequently, all Processing results were reprocessed using a single Consensus workflow with the "Merge Mode" parameter in the MSF files node set to Do Not Merge. With this setting, the number of unique peptides and PSMs for each of the datasets will be represented as a separate column. The Sequest HT search was performed against the entire *Arabidopsis thaliana* and *Escherichia coli* databases. The table with PSM values for each sample was exported to Microsoft Excel format and ratios were calculated manually.

For the "Top N" quantification workflow, a Precursor Ion Area Detector node was incorporated in the Processing workflow used for spectral counting above. The default "CWF_Comprehensive_Enhanced_Annotation_Quan" template was used for the Consensus workflow. In the Peptide and Protein Quantifier node, the "Top N Peptides Used for Quantification" parameter was set to 3. Like for spectral counting, the table with the reported Top N protein abundances was exported to Excel and ratios were calculated manually.

New Method for Feature Detection

The new feature detection algorithm is an extension of the Minora algorithm, which had already been used for precursor ion quantification since the release of Proteome Discoverer 1.2 software. Minora had always detected all isotopic peaks in a given data set, but up to now only those LC/MS peaks associated with peptide spectral matches (PSMs), and their associated isotopic forms in the case of SILAC, were used for quantification. In this pre-release version of Proteome Discoverer 2.2 software, the Minora algorithm has been modified to detect and quantify isotopic clusters regardless of whether or not they are associated with a PSM.

A typical Processing workflow for Minora feature detection is shown in **Figure 1**. The new label-free quantification workflow can be invoked by simply attaching the "Minora Feature Detector" to the Spectrum Files node. This new feature detector will also be used for the isotopically-labeled precursor quantification method such as SILAC.

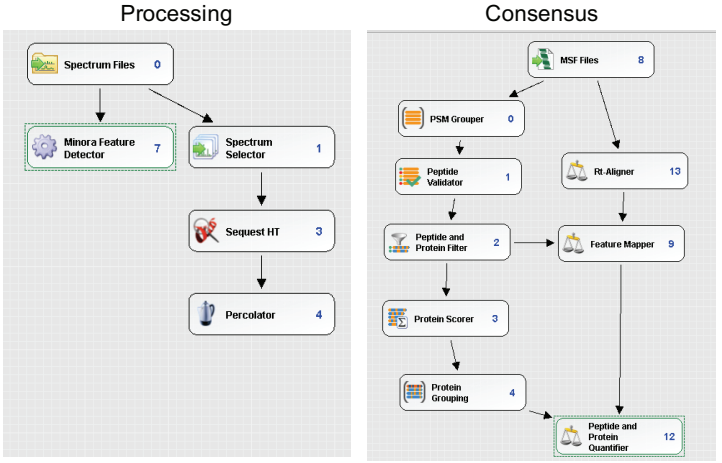


Figure 1. Typical Processing and Consensus workflows for untargeted label-free quantification. The Minora Feature Detector, Rt-Aligner and Feature Mapper are new nodes created for the untargeted label-free quantification workflow. The Minora Feature Detector will also replace the old Precursor Ions Quantifier node used for SILAC and other precursor ion quantification workflows.

ThermoFisher SCIENTIFIC

A typical Consensus workflow for label free quantification is also shown in **Figure 1**. There are two new nodes added to this workflow that perform retention time alignment and feature mapping. The feature mapper groups features detected from the Processing runs into "Consensus Features" that are mapped and quantified across all raw files and performs gap filling to find features that were not initially detected in the processing workflows. The Peptide and Protein Quantifier node works as previously, with improvements to scaling and normalization that benefit all quantification workflows.

There are three new tabs for feature detection results in the consensus report: Consensus Features, LCMS Features, and LCMS Peaks. The LCMS features are isotopic clusters grouped together for a given raw dataset and consist of multiple LCMS Peaks. Ultimately, the release may not include the LCMS Peaks list given that as much as 10's of millions peaks could be detected in complex datasets. The consensus features link directly to the associated peptide group as well as the list of LCMS features detected from each data files (**Figure 2**). Also, when a consensus feature is selected, the traces for each of the features are shown in the chromatogram traces view. When a single LCMS Feature is selected, the chromatographic profile for only that individual feature is displayed.

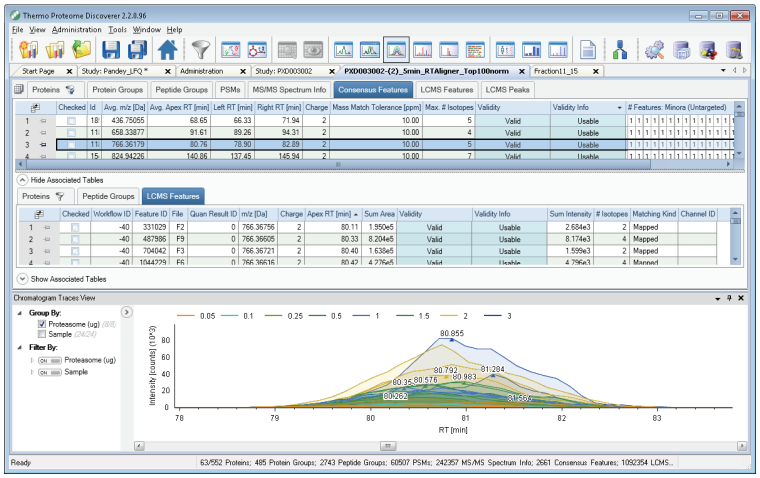


Figure 2. The Consensus Features table is linked to the collection of LCMS Features from each raw file. The chromatographic profiles for each LCMS Feature are shown in the Chromatogram Traces View at the bottom.

Like the other quantification workflows in Proteome Discoverer 2.1 software, the peptide group abundances from the new label-free quantification method are calculated as the sum of the abundances of the individual PSMs for a given study factor that pass a quality threshold. The protein abundance is calculated as the sum of the peptide group abundances associated with that protein.

RESULTS AND DISCUSSION

The database searches produced a total of 55 *Arabidopsis* proteins and 423 *E. coli* proteins. This is less than would be expected given the relatively high protein concentration and long gradient length, but the chromatography used for these data analyses was suboptimal with peaks up to 5 minutes wide (**Figure 3**). Also, as the amount of the *Arabidopsis proteins* added to the sample increased past 1 µg, the peptides from these proteasome proteins dominate the chromatogram and thus the number of *E. coli* proteins decreases dramatically with increasing *Arabidopsis* protein concentration (data not shown).

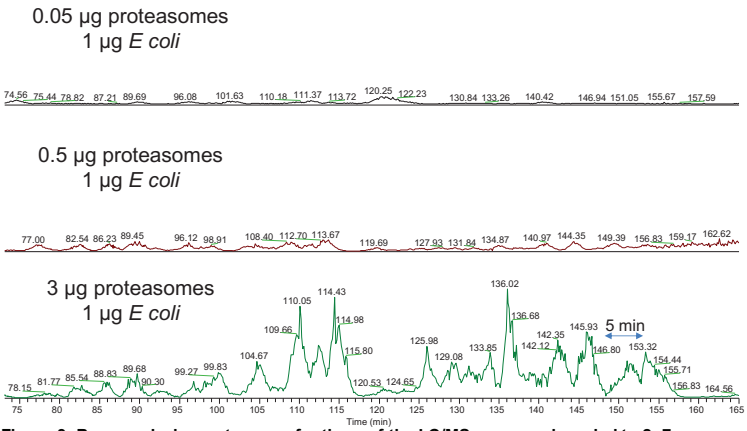


Figure 3. Base peak chromatograms for three of the LC/MS runs, each scaled to 2e7 intensity. The dataset at the bottom is dominated by *Arabidopsis* peptides, leading to significant suppression the *E coli* peptides. Also, the typical chromatographic peak in this chromatogram can be up to 5 minutes wide, also decreasing the number of peptides and proteins that can be identified.

The spectral counts-based quantification results correctly indicate the direction of expression for the *Arabidopsis* proteasome proteins, but the ratios are inaccurate for the more extreme ratios. The response is also relatively non-linear, with the average ratio for the 0.1 µg/1 µg samples showing a lower value than the 0.05 µg/1 µg samples and the 3 µg/1 µg ratio measuring lower than the 2 µg/1 µg ratio. These results are not a surprise given that it is widely known that this type of spectral counting is not expected to produce accurate quantification results. Normalized spectral counting algorithms are a significant improvement over the basic spectral counting method shown here and reference 1 from which these data were obtained describes a such a method. Implementation of such a method using emPAI values is planned for the individual study factors is being considered for a future Proteome Discoverer software release. However, all spectral counting-based quantification methods usually provide poorer sensitivity and dynamic range than other quantitative techniques due to the requirement for multiple PSMs for any given protein. As can be seen in Table 1, less than half of the *Arabidopsis* proteins could be quantified across the full dynamic range due to lack of PSMs in the samples with lowest protein abundance.

The "Top N" protein quantification results are shown in the second set of columns in Table 1. The accuracy of the ratios is noticeably improved compared to spectral counting, producing a response that is closer to linear. However, there are fewer quantified proteins in the "Top N" method than for spectral counting, primarily due to the requirement that the same three peptides need to be identified across all of the datasets. This is in effect even more stringent than the spectral counting method above and as a result even fewer proteins are quantified across the samples. Also, while the accuracy of the ratios is improved, the precision of the measurements are not much improved over spectral counting.

For feature detection-based quantification, the calculated ratios were significantly closer to the theoretically expected values at the lowest *Arabidopsis* concentrations. The precision of the ratios was also significantly improved in almost all cases for the feature detection results. The use of feature mapping led to a significantly increased number of quantified proteins given that only a single PSM is required for a given peptide across all raw files. The accuracy and precision of this method also benefits from the use normalization based on the *E. coli* proteins, which are known to be equally abundant across all samples.

A screen shot of the *Arabidopsis* protein identification results with untargeted label-free quantification is shown in **Figure 4.** The ratios and the scaled abundances for the identified proteins and peptide groups are color coded based on the level of expression. Scaled abundances were originally introduced in Proteome Discoverer 2.1 software primarily for the TMT quantification workflow and are now available in the preliminary version of Proteome Discoverer 2.2 software for feature detection-based label free quantification. The samples can be sorted by scaled abundance for any given sample type, as seen in **Figure 4** for the highlighted 0.05 sample group. It can be easily seen that each of the proteasome proteins exhibit a similar trend by simply looking at the pattern of blue and red boxes. Also, since the scaled abundances exhibit the same profile as the ratios, the need for the calculation of ratios is somewhat obviated.

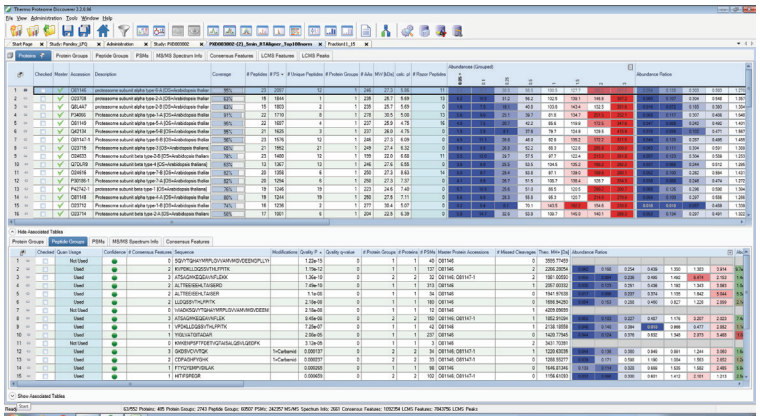
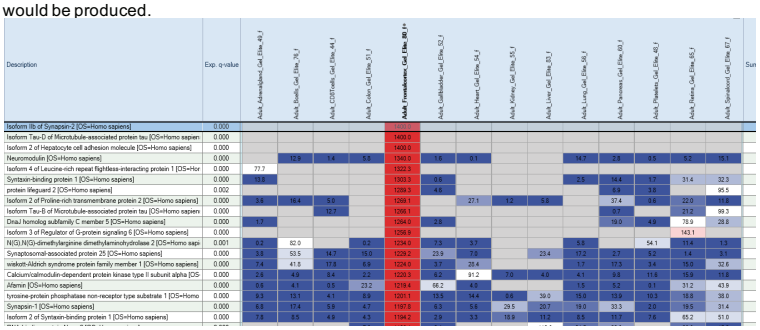


Figure 4. Untargeted label-free quantification results within the Proteome Discoverer software framework. Both the ratios and the scaled abundance values are color-coded to display significantly under- or over-expressed proteins.

Multidimensional LC profiling

The new untargeted label-free quantification algorithm also supports multi-dimensional label-free data. The processing step works as described previously for such data with the feature mapping and retention alignment steps only applied to the same fraction from other datasets. Fractions 11-15 for 14 of the samples from the Pandey group human proteome data were run using the same workflows as shown in the previous section. For these data, a total of 5116 proteins and 60616 unique peptides were identified. Unlike the previous version of Proteome Discoverer software where these data were run using "Top 3" protein quantification results, the pre-release Proteome Discoverer 2.2 software enables scaled abundance visualization of the various samples. **Figure 5** shows the proteins most overrepresented in the frontal cortex relative to the other samples by sorting by decreasing scaled abundance. Many of the most overrepresented proteins are all known to be neural proteins, including synapsin-1, synapsin-2, synapsin-3, neuromodulin, and microtubule-associated protein tau. Also, some of these neuronal proteins show no signals for any of the other samples in **Figure 5** by showing gray boxes indicating that there were no quantification values for these proteins. This is an indication that the Feature Mapping is actually working correctly by not associating random features from the other datasets. Also, this shows the value of the scaled abundance compared to ratio calculations. If any of the other samples were used as the denominator for the ratio calculations, it might be missed that the selected protein is found only in the frontal cortex sample due to the undefined ratios that would be produced.



Description	Exp. q-value	Adult_Adrenalgland_Cad_Elite_40_f	Adult_Brain_Cad_Elite_39_f	Adult_CD3TCells_Cad_Elite_44_f	Adult_Colon_Cad_Elite_51_f	Adult_Frontalcortex_Cad_Elite_30_f	Adult_GallBladder_Cad_Elite_52_f	Adult_Heart_Cad_Elite_54_f	Adult_Kidney_Cad_Elite_55_f	Adult_Liver_Cad_Elite_53_f	Adult_Lung_Cad_Elite_56_f	Adult_Pancreas_Cad_Elite_43_f	Adult_Platelets_Cad_Elite_45_f	Adult_Testis_Cad_Elite_46_f	Adult_Thyroid_Cad_Elite_47_f	Sum
Isoform IIb of Synapsin-2 [OS=Homo sapiens]	0.000					1401.5										
Isoform Tau-D of Microtubule-associated protein tau [OS=Homo sapien	0.000					1400.0										
Isoform 2 of Hepatocyte cell adhesion molecule [OS=Homo sapiens]	0.000					1400.0										
Neuromodulin [OS=Homo sapiens]	0.000		12.9	1.4	5.8	1340.0	1.6	0.1			14.7	2.8	9.5	5.2	16.1	
Isoform 4 of Leucine-rich repeat flightless-interacting protein 1 [OS=Hor	0.000	77.7				1322.2										
Syntaxin-binding protein 1 [OS=Homo sapiens]	0.000	13.8				1303.3	0.6				2.5	14.4	1.7	31.4	32.3	
protein lifeguard 2 [OS=Homo sapiens]	0.002					1299.3	4.6					6.9	3.8		95.5	
Isoform 2 of Proline-rich transmembrane protein 2 [OS=Homo sapiens]	0.000	3.8	16.4	5.0		1269.1		27.1	1.2	5.8		37.4	9.6	22.0	0.8	
Isoform Tau-B of Microtubule-associated protein tau [OS=Homo sapien	0.000			12.7		1266.1						9.8		21.2	99.3	
DnaJ homolog subfamily C member 5 [OS=Homo sapiens]	0.000	1.7				1264.0	2.8					19.0	4.9	78.9	28.8	
Isoform 3 of Regulator of G-protein signaling 6 [OS=Homo sapiens]	0.000					1256.9								143.1		
N(G),N(G)-dimethylarginine dimethylaminohydrolase 2 [OS=Homo sapi	0.001	0.2	82.0		0.2	1234.0	2.3	3.7			5.8		54.1	11.4	1.3	
Synaptonemal-associated protein 25 [OS=Homo sapiens]	0.000	3.8	53.5	14.7	15.0	1229.2	25.9	7.8		23.4	17.2	2.7	6.2	1.4	3.1	
wiskott-Aldrich syndrome protein family member 1 [OS=Homo sapiens]	0.000	7.4	41.8	17.8	6.9	1224.0	3.7	28.4			1.7	17.3	3.4	15.0	32.6	
Calcium/calmodulin-dependent protein kinase type II subunit alpha [OS	0.000	2.6	4.9	8.4	2.2	1220.3	6.2	91.2	7.0	4.9	4.1	9.8	11.8	15.9	11.8	
Aftimin [OS=Homo sapiens]	0.000	0.6	4.1	0.5	23.2	1219.4	66.2	4.8			1.5	5.2	9.1	31.2	43.8	
tyrosine-protein phosphatase non-receptor type substrate 1 [OS=Homo	0.000	9.3	13.1	4.1	8.9	1201.1	10.9	14.4	9.6	39.0	15.0	13.9	10.3	19.9	38.9	
Synapsin-1 [OS=Homo sapiens]	0.000	6.8	17.4	5.9	4.7	1197.8	6.3	5.6	29.5	20.7	19.0	33.3	2.0	19.5	31.4	
Isoform 2 of Syntaxin-binding protein 1 [OS=Homo sapiens]	0.000	7.8	8.5	4.9	4.3	1194.2	2.9	3.3	19.9	11.2	8.5	11.7	7.6	65.2	51.0	

Figure 5. Minora feature detection results for the subset of the Pandey human proteome dataset. The features are sorted by decreasing scaled abundance for the frontal cortex sample.

CONCLUSIONS

A new untargeted label-free quantification workflow based on the Minora algorithm has been demonstrated on a dataset with proteins at known concentration and is shown to be more accurate and sensitive than the previously available label-free quantification approaches from previous versions of Proteome Discoverer software. The combination of the label-free quantification workflow integrated into the scaling, normalization, and study management features of Proteome Discoverer software provide a powerful means for analyzing highly complex proteomics data.

REFERENCES

1. Gemperline, DC *et al*, *Proteomics*, **16**, pp 920-924.
2. Kim, M.-S. *et al*, *Nature*, **509**, 575-581.

PN64857-EN 08/16S