Poster Note 64491

Quantitative Analysis of Large Phosphopeptide Datasets Using Proteome Discoverer 2.0

David M. Horn,[1] Ilyas Singec,[2] Laurence Brill[2]
[1]Thermo Fisher Scientific, San Jose, CA, USA ; [2]Sanford-Burnham Research Institute, La Jolla, CA

Overview

Purpose: Interpretation of a complex quantitative phosphoproteomic dataset in the Proteome Discoverer 2.0 platform.

Methods: Phosphopeptides were enriched using IMAC or TiO2 from hESC and hNSC samples and the flow through, wash and enriched fractions were collected. All MS data were acquired on a Thermo Scientific™ LTQ Orbitrap™ Mass Spectrometer system equipped with ETD. Data were analyzed using the new study management features in Thermo Scientific™ Proteome Discoverer™ software 2.0 using a label-free quantification approach

Results: Proteome Discoverer 2.0 software processed the 4 stem cell samples in ~7 days with identification and quantification of phosphoproteins and phosphopeptides. Several proteins and phosphopeptides are shown to be differentially expressed and these are known to be regulated in stem cells as well as some novel proteins not known to be differentially regulated.

Introduction

With advances in liquid chromatography/mass spectrometry (LC/MS) technology, The complexity of proteomics data is increasing rapidly. It is becoming increasingly common to find datasets with 100's of GB of raw data files for complex scientific studies, putting an increased burden on downstream software tools for interpretation of such datasets.

The latest release in the Proteome Discoverer platform has several new features for analysis of complex datasets. The first major feature is the new Consensus workflow that creates persistent reports that open very large datasets quickly. Secondly, the results are presented in a new hierarchical format with linked views for protein groups, proteins, peptides, and peptide spectrum matches (PSMs). Third, the most critical feature for analysis of large quantitative datasets is the new study management. These will be demonstrated on a large dataset of a quantitative comparison between human embryonic stem cell (hESC) and neural stem cell (hNSC) derivatives

Methods

Proteins were extracted from hESCs and hNSCs, reduced and alkylated using iodoacetamide, digested using trypsin, and separated into 32 fractions using strong cation exchange (SCX) chromatography. Phosphopeptides enrichment was preformed, also collecting the flow through and wash fractions from the SCX fractions. More details on the cells and the sample preparation will be available in a forthcoming publication. Each of the fractions were analyzed in duplicate using a data dependent decision tree LC/MS/MS method on an LTQ Orbitrap Velos mass spectrometer equipped with ETD.

These study factors will be used by Proteome Discoverer to determine which quantification values will be calculated and shown in the final report.

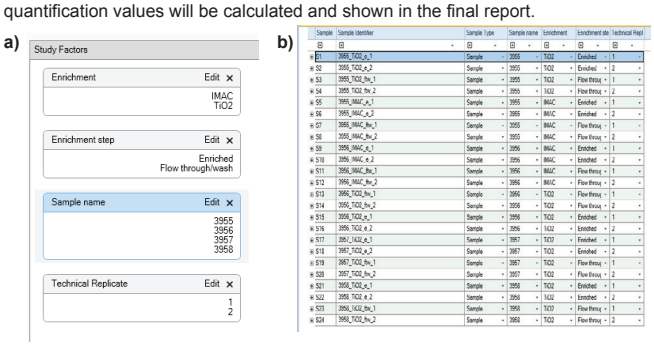


FIGURE 1a) Study factors created for this analysis. b) List of samples imported into Proteome Discoverer software and the study factors assigned to those samples. Each of these samples corresponds to at least 25 raw data files.

The next step is to choose the raw files. Proteome Discoverer 2.0 software includes a new "Add Fractions" feature that groups a set of raw data files as a single sample. For this study, each set of raw data files that fit the criteria of the study factors above such as Sample 3955, TiO2-enriched phosphopeptides, technical replicate 1 are loaded simultaneously as a single sample. A typical group has 25-50 data files, with the larger collection of data files corresponding to samples where the flow through and wash samples from phoshpopeptide enrichment were run as separate LC/MS/MS runs. Once the datasets were imported into the software, the study factors defined in Figure 1a were assigned to each sample as shown in Figure 1b above.

Thermo SCIENTIFIC

The next step is to choose the raw files. Proteome Discoverer 2.0 software includes a new "Add Fractions" feature that groups a set of raw data files as a single sample. For this study, each set of raw data files that fit the criteria of the study factors above such as Sample 3955, TiO2-enriched phosphopeptides, technical replicate 1 are loaded simultaneously as a single sample. A typical group has 25-50 data files, with the larger collection of data files corresponding to samples where the flow through and wash samples from phoshpopeptide enrichment were run as separate LC/MS/MS runs. Once the datasets were imported into the software, the study factors defined in Figure 1a were assigned to each sample as shown in Figure 1b above.

The third step is to create a new analysis, which includes two node-based workflows as well as the selection of the quantification details. Proteome Discoverer 2.0 software introduces a new dual workflow setup that includes a Processing Workflow for peptide identification, peptide quantification, FDR calculation and PTM site localization and adds a new Consensus workflow that is used to perform protein inference, filter the results based on FDR and other calculations, and summarize quantification results. The two workflows used for this study are shown in Figures 2a-b below.

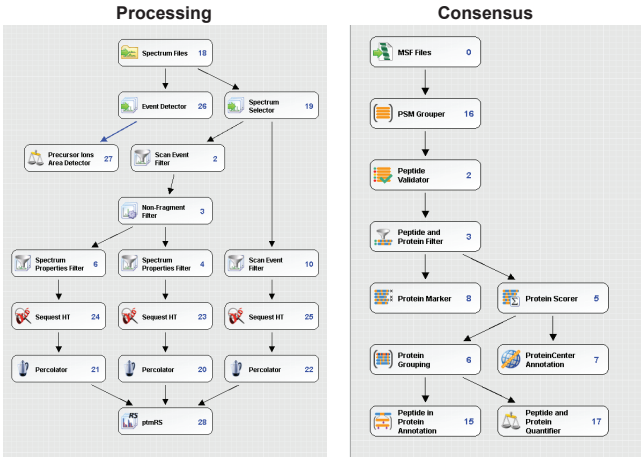
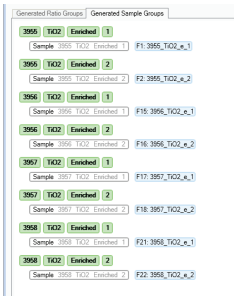


FIGURE 2. Processing and consensus workflows for the phosphopeptide analysis.

The processing workflow in Figure 2 includes Sequest HT nodes to interpret both CID and ETD MS/MS and includes the Event Detector and Precursor Ions Area Detector nodes to calculate peak areas for identified peptides. All Sequest HT searches used 10 ppm precursor mass tolerance, while the CID data used 0.6 m/z fragment tolerance and the ETD data used 1.2 m/z fragment tolerance. All searches used fixed carbamidomethylation, variable phosphorylation (S,T,Y), oxidation (M), and pyro-Glu (peptide N-term Q). Finally, the ptmRS node was appended to the end of the workflow to calculate modification site localization probability.

The Consensus workflow is used to calculate protein groups, filtering peptides and proteins by false discovery rate, and rolling up the quantification results from individual peptide spectral matches (PSM's) to peptide groups and proteins.

a)



b)

FIGURE 3. Study factors for the enriched (a) phosphopeptide search and the flow-through and wash fractions (b). There are as many columns in the final report as are shown in green on this table.

The final step is to set up the columns to be displayed in the final result for quantification. The columns that are highlighted in green as in Figure 3 are those that will show up in the final report with the quantification values. This setup also works similarly for reporter ion quantification and isotopically-labeled precursor quantification.

For these data, two different analyses were run. The first analyzed the samples from the enriched fractions to compare phosphopeptides while the second analyzed the combined flow-through and wash samples to compare changes in protein abundances across the samples.

Results

Phosphopeptide search results

The IMAC enrichment was less successful for these samples and thus the IMAC enrichment data for samples 3955 and 3956 were removed from the analysis for clarity. As a result, there were two technical replicates for the TiO_2-enriched phosphopeptides for each of the original samples leading to 8 different quantitative categories as shown in Figure 4 for the 242 raw files. Figure 4 shows a screenshot of the Proteome Discoverer results of the phosphopeptide data.

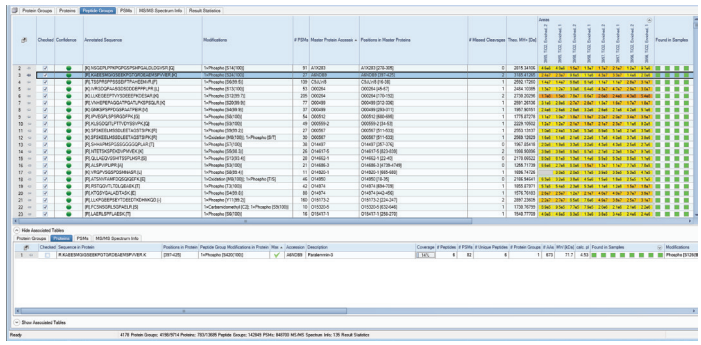


Figure 4. Identified peptide groups for TiO_2-enriched phosphopeptides. The fourth column shows the modification and the site localization probability calculated by ptmRS. The "Area columns" show the peak area calculated for that identified phosphopeptide across all 8 multidimensional LC/MS/MS runs. The selected peptide is about 20x more abundant in the hESCs compared to the hNSCs with good biological and technical reproducibility.

Table 1 shows a summary of the number of PSM's, peptides, and phosphopeptides identified across the 4 samples. To find phosphorylation sites that are unique to the two hESC samples (3955 and 3957), the row filters were selected to show phosphopeptides that appear in one of the two replicates in each case that were not identified in the two hNSC samples (3956 and 3958). In total, there were 811 unique phosphopeptides identified only in the hESC samples versus 253 unique phosphopeptides identified only in the two hNSC samples.

Sample	PSM's	Unique peptides	Unique phoshpopeptides	Unique phosphopeptides (ptmRS isoform score >60)
hESC 1	38174	7715	5575	4151
hNSC 1	21987	4562	3191	2388
hESC 2	47338	7835	6223	4526
hNSC 2	35350	7086	4898	3677

TABLE 1. Identification summary for enriched phosphopeptide samples.

For the unique hESC phosphopeptides, overrepresentation analysis of pathways was performed and insulin signaling, MAPK, ErbB, and AMPK pathways were identified. For the peptides unique to the hNSC samples, there were several phosphorylation sites detected on the protein MAP2, which is known to be involved in neurogenesis, SOX-5, which is involved in chondrogenesis, and neuron-navigator 1.

Some phosphopeptides were also identified in all samples but are also differentially abundant. There were 38 phosphopeptides that appear to be up-regulated by a factor of ≥2 in hNSCs while there were 101 phosphopeptides that were up-regulated in hESCs. The phosphopeptides with the highest differential expression ratios between the hESC and hNSC samples are shown in Table 2.

Peptide	Protein	Ratio hNSC/hESC
LKCGSGPVHISGQHLVAVEEDAE**S**EDEEEEDVK	Rho GTPase activating protein 17	0.025
RPTPNDDTLDEGVGLVHSNIATEHIP**S**PAK	Hydroxymethylglutaryl-CoA synthase	0.032
MAPTPIPTR**S**PSDSSTASTPVAEQIER	Drebrin	0.055
SSM**S**GLHLVK	Acetyl-CoA carboxylase 1	0.058
DME**S**PTKLDVTLAK	Microtubule associated protein 4	0.068
TTR**T**PEEGGYSYDISEK	Microtubule associated protein 1B	18
RPA**S**PSSPEHLPATPAESPAQR	Sin3 histone deacetylase corepressor complex component SDS3	19
VAL**S**DDETKETENMR	DNA polymerase delta subunit 3	21
RS**T**QGVTLTDLQEAEK	Protein phosphatase 1 regulatory subunit 12A	31
IED**S**EPHIPLIDDTDAEDDAPTKR	Plasma membrane calcium-transporting ATPase 1	120

TABLE 2. Phosphopeptides with the lowest and highest quantitative ratios between hESCs and hNSCs.

Protein Quantification

For protein quantification, another set of searches were performed on the flow-through and wash fractions from the phosphopeptide enrichment steps. The same workflows were used as shown in Figure 2, but phosphorylation was removed as a variable modification from the Sequest HT searches and Protein FDR threshold node set to 1% was added. For these data, the "Proteins" tab in the final results show the average of the top 3 most abundant peptides detected across all of the samples as seen in Figure 5 below.

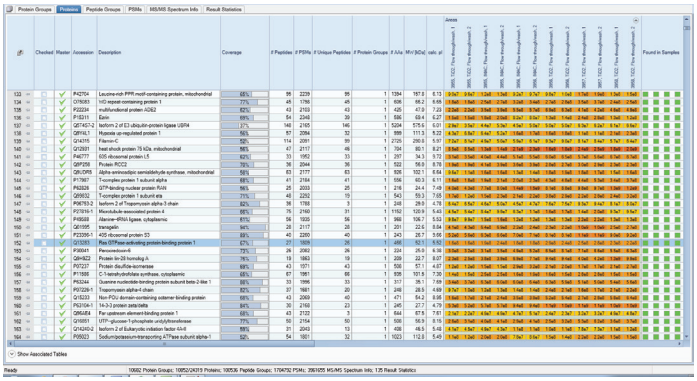


Figure 5. Identified proteins for the flow-through and wash fractions across the samples. The "Areas" table shows the average of the peak areas of the top 3 identified peptides from that protein for each sample.

A total of 10682 unique proteins were identified across all of the samples. Table 3 shows the proteins identified for each of the samples, with roughly 6500-7000 proteins identified in each sample. Row filters were applied to show only those proteins identified in the various hESC samples, with a total of 285 proteins with two or more unique peptides identified in at least one of the hESC samples but none of the hNSC samples (data not shown). This includes proteins such as cadherin-3, PR domain zinc finger protein 14, Tyrosine-protein kinase Lck, and Oct4. A second set of filters were used to show proteins that only appear in one or more hNSC samples, which produced a list of 458 proteins with at least 2 peptides. Selected proteins include known nervous system proteins ephrin type-A receptor 4 isoform a precursor, contactin-associated protein 1 precursor, and Spondin-1.

Sample	Enrichment	Replicate	Protein groups	Peptide Groups
hESC 1	TiO_2	1	6240	47667
hESC 1	TiO_2	2	6303	48105
hESC 1	IMAC	1	5854	35924
hESC 1	IMAC	2	5991	36990
hNSC 1	TiO_2	1	7047	54935
hNSC 1	TiO_2	2	7044	55473
hNSC 1	IMAC	1	6551	43585
hNSC 1	IMAC	2	6654	44475
hESC 2	TiO_2	1	6701	54023
hESC 2	TiO_2	2	6776	54385
hNSC 2	TiO_2	1	6605	50303
hNSC 2	TiO_2	2	6639	50887

TABLE 3. Summary of protein identifications for each sample.

The table of protein identifications and quantitative information was exported to Excel. To account for differences in the injection amounts for each sample, the abundances for each protein were normalized to the summed abundance of all proteins in the sample. The results were imported into ProteinCenter and profiled. The top and bottom clusters showed overrepresentation in hNSCs and hESCs respectively (Figure 8).

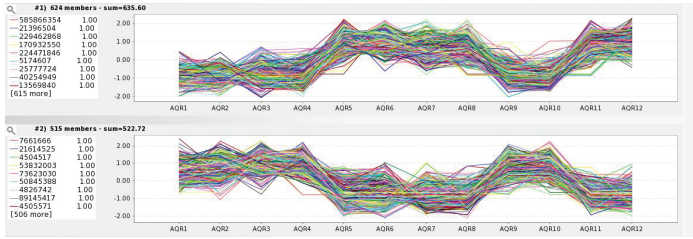


FIGURE 8. Top two clusters from ProteinCenter profiling. The upper cluster corresponds to proteins more abundant in the hNSCs whereas the lower cluster corresponds to proteins that are more abundant in the hESCs.

Some selected proteins in cluster 1 include several ephrin receptors, PAX-6, and frizzled-3 precursor all of which are involved in neuronal development in animal models, and SOX-2, which is known to be involved in stem cell pluripotency and differentiation. Cluster 2 contains dozens of highly up-regulated (≥10x) proteins including gamma-synuclein, CD9 antigen, calveolin-1 isoform alpha, HRAS-like suppressor 3, and nocturnin.

Conclusion

- Proteome Discoverer 2.0 software is well equipped to analyze highly complex datasets, in this case a dataset with well over 800 RAW data files.
- The study management feature was used to produce peak area quantification values for phosphopeptides and the proteins from the flow-through and wash fractions from TiO_2 and IMAC enrichment.
- The precursor ion quantification can be used to find differentially expressed phosphopeptides as well as proteins between the hESC and hNSC samples.
- ProteinCenter aids in the biological interpretation of the Proteome Discoverer software results.

Thermo Fisher Scientific, San Jose, CA USA is ISO 13485 Certified.

ISO 13485

Africa +43 1 333 50 34 0
Australia +61 3 9757 4300
Austria +43 810 282 206
Belgium +32 53 73 42 41
Canada +1 800 530 8447
China 800 810 5118 (free call domestic)
400 650 5118
Denmark +45 70 23 62 60
Europe-Other +43 1 333 50 34 0
Finland +358 10 3292 200
France +33 1 60 92 48 00
Germany +49 6103 408 1014
India +91 22 6742 9494
Italy +39 02 950 591
Japan +81 45 453 9100
Korea +82 2 3420 8600
Latin America +1 561 688 8700
Middle East +43 1 333 50 34 0
Netherlands +31 76 579 55 55
New Zealand +64 9 980 6700
Norway +46 8 556 468 00
Russia/CIS +43 1 333 50 34 0
Singapore +65 6289 1190
Spain +34 914 845 965
Sweden +46 8 556 468 00
Switzerland +41 61 716 77 00
UK +44 1442 233555
USA +1 800 532 4752

Thermo SCIENTIFIC
A Thermo Fisher Scientific Brand

PN64491-EN 0615S