Quantitative Analysis of Large Phosphopeptide Datasets Using Proteome Discoverer 2.0

David M. Horn,¹ Ilyas Singec,² Laurence Brill² ¹Thermo Fisher Scientific, San Jose, CA, USA ; ²Sanford-Burnham Research Institute, La Jolla, CA

а

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.

			Sample	Sample Identifier	Sample Type	St	npie nam	Enrichment	Enrichment ste	Technical Rep
du Enctore		b) -	Θ	•	• •	E	•	· •	Θ ·	
ay r accora			\$1	3955_TIC2_0_1	Sample	· 33	55	- TO2 -	Enriched ·	1 -
			52	3555_TiO2_e_2	Sanple	• 33	55	T02	Enided ·	2 .
Enrichment	Edit 🗙		53	3955_TK02_6v_1	Sample	• 35	55	TO2 -	Flow throug .	1 .
	IMAC		i S4	3955_TIQ2_tbv_2	Sample	• 35	55	102	Filter throug .	2.
	TiO2		95	3555_MAC_#_1	Sample	· 35	55	MAC -	Enriched +	1 .
	1102		56	3955_M4C_e_2	Sample	· 31	55	- M/C -	Enriched •	2 -
			\$7	3955_IM4C_thr_1	Sample	- 33	55	- MAC -	Flow throug .	1 -
Enrichment sten	Edit 🗙		58	3555_MAC_bv_2	Sanple	• 33	55	MAC -	Flow throug .	2 -
	//		59	3556_IM4C_e_1	Sample	• 35	56	MAC .	Enriched ·	1 .
	Enriched Flow through/wash		570	3556_IM4C_e_2	Sample	• 3	56	MAC .	Enriched ·	2.
			S11	3556_M4C_thr_1	Sample	• 35	56	- M4C -	Fine throug +	1 .
			\$12	3956_M4C_thr_2	Sample	 35 	56	- M/C	Flow throug +	2 -
			St3	3956_TiO2_bv_1	Sample	. 3	56	- TO2 -	Flow throug -	1 -
ample name	Edit ×		\$14	3956_TK02_bv_2	Sample	· 35	56	- TO2 -	Flow Broug +	2 .
	3955		\$15	3556_TK02_e_1	Sample	• 35	56	TO2 •	Enriched ·	1 .
	3956		S16	3556_TIQ2_6_2	Sample	• 35	56	102	Ennched ·	2.
	3957		S17	3957_TIC2_e_1	Sample	• 3	57	TO2	Enriched +	1 •
	3958		\$18	3957_TK02_+_2	Sample	· 35	57	TO2 -	Enriched +	2 -
			\$19	3967_TiC2_tw_1	Sample	· 31	57	- TO2 -	Flow throug +	1 -
			520	3957_TK02_6v_2	Sample	· 31	57	- TO2 -	Flow throug .	2 .
echnical Replicate	Edit ×		\$21	3958_TIC2_e_1	Sample	• 33	58	- TO2 -	Enriched ·	1 •
			S22	3958_TIQ2_e_2	Sample	• 35	58	102	Enriched ·	2 .
	2		573	3558_TK02_thv_1	Sample	• 3	58	TO2 -	Fine throug +	1 .
			524	3958 TiO2 ftw 2	Sample	. 25	52	TO2 -	Fige throw +	2 -

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 1 a were assigned to each sample as shown in Figure 1b above.



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.



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 TiO₂-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
	Idontificat	ion cummany fo	r opriched phosphope	ntido complos

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 <u>S</u> EDEEEEDVK	Rho GTPase activating protein 17	0.025
RPTPNDDTLDEGVGLVHSNIATEHIP <u>S</u> PAK	Hydroxymethylglutaryl-CoA synthase	0.032
MAPTPIPTR S PSDSSTASTPVAEQIER	Drebrin	0.055
SSM <u>S</u> GLHLVK	Acetyl-CoA carboxylase 1	0.058
DME <u>S</u> PTKLDVTLAK	Microtubule associated protein 4	0.068
TTR <u>T</u> PEEGGYSYDISEK	Microtubule associated protein 1B	18
RPA S PSSPEHLPATPAESPAQR	Sin3 histone deacetylase corepressor complex component SDS3	19
VAL <u>S</u> DDETKETENMR	DNA polymerase delta subunit 3	21
RS <u>T</u> QGVTLTDLQEAEK	Protein phosphatase 1 regulatory subunit 12A	31
IED <u>S</u> 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.

ø	Checke	ć Master	Accession	Description	Coverage	I Papidas	2 PSH4	# Unique Peptides	# Protein Groupe	1.1/4	Mir/BOx	calc p	MS. TEC: Film Becoghimedi. 1	85, TEO. Plan Broughinstei, 2	RS, SHC, Paul Record March 1	85, 896C, Flow RecogNismet, 2	06, 504C, Flue Evolutivesh, 1	08. INC. Pow tecophysics, 2	96. TLCL / Env Broughinest). 1	RM, TCD, Flow Broughtman, 2	87, TGC. Plan throughings). 1	RR, TCC, Flow throughtenesis, 1	88. T.C.C. Plan Broughinson, 2 🛞	Found	l in San
10		1.1	8/770/	I want of \$20 and anticipation and a mindradial	115		1110			1184	167.0		8	R	8	8		8	8		8 1	8 R	8		-
		15	005087	https://www.comparing.gover.meconomia	100		1780			676	66.7				1.000	100	1.00	1.4					1000	-	22
		15	8222214	m Nilverfreed enters 5/923	17%	43	2100	0		116	47.6	2.95	1.1.1	1.54	114	104	114	1.54			141.41		100	- 12	22
		15	016315	Enie	687		2540	20		600	65.4	6.22	114	1144	104	104	1101	1.747			440		1.54	-	22
0.0		13	067453-3	Jackwar 2 of E3 altimatic contain linese 1004	1977	142	2165	145	1	5204	121.0	6.01	224	167	44.	55.7	450	50.7		1.00	22.01		0.547	-	22
		15	Official 1	Minerala un remulated contain 1	625	87	2054	32	1	000	111.3	\$ 22	130	4147	1447	104	164	1.74		1	141.11	-	2348		6 B
		13	014115	FaminC	425	114	2091		1	2726	210.8	5.93	7247	1.147	45.7	5.647	141	6.347	1.52	1.0.7	147 1.1		5407	-	22
		15	012881	hast shock rooters 75 kDa, mitschoodrial	803	10	2117	48	1	354	80.1	8.21	1.1.1	1141	124	144	214	2348			40.27	1111	204	-	22
		13	046777	610 observat contain 1.6	(PAN)	22	1967	11		297	34.3	0.71	1.1.1	11-0	404	44.4	1.1.4	11.4	100	1.0	24.11		1.74	E 2	22
		15	000050	Potein BCC2	395		2044			432	56.5	8.76	110	1144	4.141	100	144	1141			~ 11			-	44
1.2		15	OBLIDES	Mitha antionation samplifabula surthana mitochondrial	68%	61	2177	41	1	876	117.1	6.64	0.047	114	164	164	114	1.64		-	-	1.154	104	-	41
		13	817987	Turometer centers 1 automit eintre	615	41	2184	41	1	114	60.3	6 11	104	114	1144	214	104	234		100	441 51	1 100	124		21
		1.5	841014	OTRAL data and an entral and an Article	443	- 14	3033			114	34.4	2.0	1.000	1.24	1110	100		-			-				22
		15	0696333	Terrenter contrin 1 and without	755	48	2262	10	1	643	66.3	746	1.54	1.54	1104	124	21.4	114		-	-		11.0	-	212
		1.5	80/763.2	hadren 2 of Tennemenia shike 2 shale	825	14	1700	1		148	16.5	4.76	1.00	11.0	11.1	100	11.1	1.747			-		1.00	-	22
1.		15	@27016.1	Month is in appropriate option of	115	15	2160	21		1150	120.9	6.0	1.00	140	1.4.5	10.0	100	1.54			-		35.7		44
		13	049500	Maximu 40% Irona crteshamir	612	14	1925	56	1	900	116.7	5.57	10.007	11.7	154	104	114	114			24.11		1.54	-	21
		13	0011995	transmin	647	28	2117	20	1	201	22.6	8.64	1.00	1.50	1.44	100	114	244		-	-		234	-	21
		15	#111164.1	ANR observed and do P1	445	45	1100			3/3	14.1	2.00					10.0	-			-				212
	-	1.0	(111181	Res (178 and only a particle binding particle 1	1 125	- 11	1976			180	63.5	6.67	1.0.0	10.0	10.0	244	10.0						-	-	20
	10	1.5	£30045	Pandia finia 6	100	14	1000	10		234	16.6	6.16	1.54	154	11.4	114	124	1.54			1.0		1.24		27
		15	094972	Pentain Sec.20 househos &	745	12	1063	10		200	22.7	9.07	114	114	114	104	1141	1.54			-		1.00	-	212
		13	007217	Dotale de ilide increase	697	43	9971	40	1	500	57.1	407	1100	1144	100	100	164	124		-	141.11		234		21
		13	#11655	C. Lambularibles authors outsilamic	655	67	1941	66	1	876	101 6	2.50	100	154	254	254	1648	104	1.0	541 2	141 20	1 154	154	-	22
		15	P63244	Outpring our factories his visus suburit help 2, like 1	60.5	33	2996	11	1	317	36.1	7.65	3.645	3.345	1.14	S Call	tint.	A 645			147 50		5.545	-	81
		1.5	607276.5	Toronomous ables of share	ETV.	97	1001	20		140	10.5	1.65	1.20	1.04	12.0	124	100	100					114	-	212
1		15	015222	Has D'VI Assais comision asterna history potale	685	43	2009	40		471	64.3	0.00	1.04	1.54	114	144	114	-			-			- 10	22
		15	001106-1	14-3.3 contain satisfaits	803	30	2160	23	1	245	27.7	470			10.04	100	144	200			-	100	1000	-	21
		15	COSLEA	Ear contrasm dismost biodice contain 1	682	41	2122	1		644	67.5	7.61	2.54	2247	1147	4.047	4.747	1.52	4.7	1.1	10.11	1 4 50	110		22
		15	015551	ITP-starset Ishoriate with threatene	775	10	2154	50	1	505	56.9	8.16	200	1.14	100	4140	254	1.54	-	1241	-	-	1334		11
		15	014340-2	hadrow 2 of Eulerootic indution factor 42x8	885	31	2023	13		416	46.5	5.0	1.147	1507	1.547	4.767	114	114	Test.	141 1	AC 11	1 1 1 1	124	88	27
		15	006010	Conf. on humanian descenation ATDass or de sit ability 1	100	64	1011	12		1003	112.8	1.0	100	1.54	104	104	10.1	11.7	1.1.1		-		1104	-	21

Ready 10062 Protein Group: 10052/24319 Protein: 10050 Peptide Group:: 1704732 PSMI: 305165 MS-MS Spectrum Into: 125 Result

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 ephrin type-A receptor 4 isoform a precursor, contactin-associated protein 1 precursor, and Spondin-1.

Sample Enrichment Replicate Protein groups Pep	ptide 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 TiO₂ 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.

www.thermoscientific.com

©2015 Thermo Fisher Scientific Inc. All rights reserved. ISO is a trademark of the International Standards Organization. All other trademarks are the property of Thermo Fisher Scientific and its subsidiaries. This information is presented as an example of the capabilities of Thermo Fisher Scientific products. It is not intended to encourage use of these products in any manners that might infringe the intellectual property rights of others. Specifications, terms and pricing are subject to change. Not all products are available in all countries. Please consult your local sales representative for details.



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
 +41 161 716 77 00

 UK
 +44 1442 233555

 USA
 +1 800 532 4752



A Thermo Fisher Scientific Brand