Identification of Low Abundance Proteins in a Highly Complex Protein Mixture

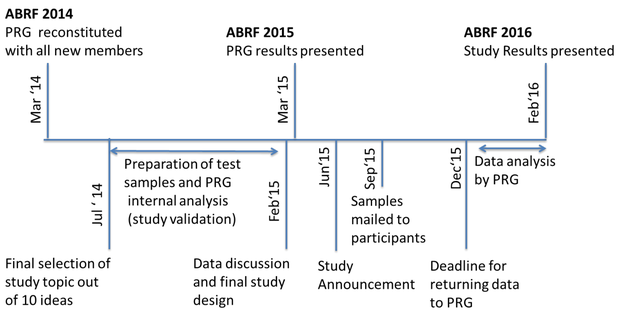
Susan Van Riper[1], Emily Chen[2], Allis Chien[3], Henriette Remmer[4] , Paul Stemmer[5], Yan Wang[6], Pratik Jagtap[7]

[1]University of Minnesota Informatics Institute, University of Minnesota, St. Paul, MN, United States, [2]Herbert Irving Comprehensive Cancer Center & Department of Pharmacology, Columbia Medical Center, New York, NY, United States, [3]Stanford University Mass Spectrometry, Stanford University, Stanford, CA United States, [4]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, United States, [5]Institute of Environmental Health Science, Wayne State University, Detroit, MI, United States, [6]Proteomics Core Facility, University of Maryland, College Park, MD, United States, [7]Center for Mass Spectrometry and Proteomics, University of Minnesota, Saint Paul, MN, United States

INTRODUCTION

Proteome Profiling of whole cell lysates is a proteomic service performed by most, if not all, core facilities as a routine service. However, capture of low-level proteins remains challenging. Sample preparation, overall fractionation time (on the peptide and protein level) as well as instrument sensitivity are main factors that determine the outcome of such an experiment. The ABRF Proteomics Research Group (PRG) decided to conduct a study on identification of low abundance proteins in a highly complex protein sample.

2015/2016 PRG: TIMELINE OF THE 2-YEAR STUDY



MATERIALS AND METHODS

The PRG provided participants with lyophilized HeLa cell lysates with four proteins spiked –in at three different amounts (in the range of 20 fmol to 500 fmol – see Table 1). Each sample set contained four samples: (a) three samples contain the four proteins at different amounts in each of the three samples but equal amounts within each sample; (b) one sample did not contain any spiked-in proteins and functioned as a negative control.

Participants were expected to perform protein identification using user-defined sample preparation and LC methods to detect these four spiked-in proteins (with no homology to human proteins). In addition participants were asked to determine which of the samples is the negative control. Participants did not know identities of the proteins, rather, the PRG provided a database that contains de-identified protein sequences for these four proteins. These proteins were labeled in the database as ABRF-1, ABRF-2, ABRF-3 and ABRF-4 (see Table 2).

Results and methods were reported by individual participants via survey monkey. Where possible, participants provide raw data and standard formatted identification results.

Table 1. Amount spiked-in to samples

Sample Label	fmol Each Protein Spiked -in
Sample 1	100
Sample 2	0
Sample 3	20
Sample 4	500

Table 2. Spiked-in Proteins

ABRF Number	Protein Name	Species
ABRF-1	Beta	Escherichia coli
ABRF-2	Lysozome	Gallus gallus
ABRF-3	Amylase	Aspergillus niger
ABRF-4	Protein G	Streptococcus

RESULTS OVERVIEW

Participants reported a diverse range of protocols and results for this limit of detection study. For each participating laboratory, we report the overall number of proteins identified, the number of spectra and unique peptides identified for spiked-in in proteins at the different amounts. Preliminary results (with outliers removed) indicate that participating laboratories were able to identify all four spiked-in proteins at 500 fmol. As expected, the identification of all four spiked-in proteins varied in different laboratories at the lower levels of spiked-in amounts. We found a significant advantage of performing fractionation on complex samples to detect proteins at an extremely low amount (e.g. 20 fmol in our study). The utility of fractionation on detecting spiked-in proteins at higher amounts appears to be protein-dependent. Unexpectedly, when .mzid files were loaded into Scaffold for validation of self-reported values, significant differences in participant reported and study validated values were found.

ACKNOWLEDGEMENTS

The **ABRF Executive Board** for support and scrutiny of study proposal. All **Participating Labs** for analyzing samples and returning data. We would like to thank Sigma-Aldrich (for providing the spiked-in proteins), ThermoFisher Scientific (teleconference support) and Laura Burr (Anonymizer, University of Michigan).

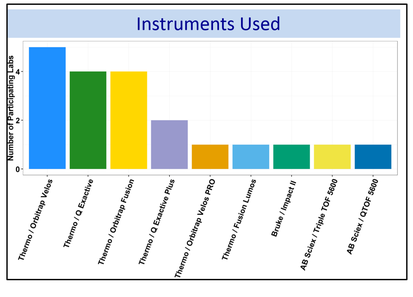
SURVEY AND INFORMATICS ANALYSIS RESULTS

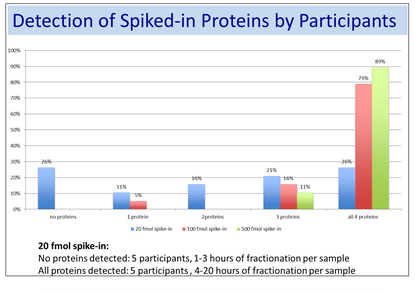
ID	Protein IDs @ 1% FDR - Sample 1	Protein IDs @ 1% FDR - Sample 2	Protein IDs @ 1% FDR - Sample 3	Protein IDs @ 1% FDR - Sample 4	# Spectra - Beta Gal- 0 fmol Spike	# Unique Peptides - Beta Gal- 0 fmol Spike	# Spectra - Lysozome - 0 fmol Spike	# Unique Peptides - Lysozome - 0 fmol Spike	# Spectra - Amylase - 0 fmol Spike	# Unique Peptides - Amylase - 0 fmol Spike	# Spectra - Protein G - 0 fmol Spike	# Unique Peptides - Protein G - 0 fmol Spike	# Spectra - Beta Gal- 20 fmol Spike	# Unique Peptides - Beta Gal- 20 fmol Spike	# Spectra - Lysozome - 20 fmol Spike	# Unique Peptides - Lysozome - 20 fmol Spike	# Spectra -Amylase -20 fmol Spike	# Unique Peptides - Amylase - 20 fmol Spike	# Spectra - Protein G - 20 fmol Spike	# Unique Peptides - Protein G - 20 fmol Spike	# Spectra - Beta Gal- 100 fmol Spike	# Unique Peptides - Beta Gal- 100 fmol Spike	# Spectra - Lysozome - 100 fmol Spike	# Unique Peptides - Lysozome - 100 fmol Spike	# Spectra -Amylase -100 fmol Spike	# Unique Peptides - Amylase - 100 fmol Spike	# Spectra - Protein G - 100 fmol Spike	# Unique Peptides - Protein G - 100 fmol Spike	# Spectra - Beta Gal- 500 fmol Spike	# Unique Peptides - Beta Gal- 500 fmol Spike	# Spectra - Lysozome - 500 fmol Spike	# Unique Peptides - Lysozome - 500 fmol Spike	# Spectra - Amylase -500 fmol Spike	# Unique Peptides - Amylase - 500 fmol Spike	# Spectra - Protein G - 500 fmol Spike	# Unique Peptides - Protein G - 500 fmol Spike	Blank_Sample
20DX	2510 (2322)	2593 (2572)	2195 (2463)	2199 (2365)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	7 (4)	1 (3)	1 (0)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	50 (25)	15 (10)	5 (5)	2 (4)	10 (6)	3 (3)	3 (2)	2 (2)	132 (126)	18 (27)	27 (20)	4 (11)	43 (45)	10 (11)	14 (8)	2 (4)	Sample 2
24BZ	3849	3844	3785	3812	0	0	0	0	0	0	0	0	21	16	4	2	11	5	13	3	105	25	12	2	60	12	22	5	248	37	25	3	94	13	44	5	Sample 2
26Z	1594	1841	1635	1333	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	4	2	0	0	1	1	4	2	29	23	4	4	11	6	5	2	
28B	1847	1888	1888	1790	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	23	2	2	14	8	2	2	143	32	11	4	35	11	18	6	2 or 3
30R	1043	1155	1116	954	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	1	1	1	1	1	1	10	8	4	4	3	3	3	3	
32W	NA (1657)	NA (1785)	NA (1722)	NA (1650)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)	0 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	1 (0)	7 (5)	7 (5)	2 (0)	2 (0)	2 (1)	2 (1)	4 (2)	4 (2)	31 (23)	25 (18)	3 (3)	3 (2)	8 (7)	7 (4)	8 (8)	5 (5)	Sample 2
34F	888 (987)	921 (930)	914 (921)	881 (903)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	7 (8)	4 (4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0	57 (65)	27 (25)	3 (3)	2 (2)	18 (7)	7 (4)	0 (0)	0 (0)	2 or 3
36Q	2588	2637	2862	2698	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	6	3	3	2	2	1	1	38	37	7	6	14	12	5	4	
40Y_1	2807	2756	2802	2639	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	22	14	5	4	8	7	5	4	81	36	8	4	28	13	14	5	
40Y_2	5453	4892	5209	5536	0	0	0	0	0	0	0	0	19	17	0	0	5	5	2	2	78	39	6	5	19	4	8	5	206	51	23	11	59	14	13	5	
44G	10632 (2675)	9359 (2286)	9951 (2425)	9398 (2388)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	43 (13)	32 (14)	8 (3)	6 (3)	16 (6)	11 (6)	6 (1)	4 (2)	0 (38)	0 (37)	0 (8)	0 (8)	0 (16)	0 (16)	0 (5)	0 (2)	Sample 2
46GV	1594	1593	1597	1598	0	0	0	0	0	0	0	0	3	2	0	0	0	0	2	1	20	10	3	2	9	5	5	3	70	27	15	6	30	10	18	6	Sample 2
48T**	1604	NA	NA	NA	44	18	0	0	5	2	0	0	44	18	0	0	5	2	0	0	44	18	0	0	5	2	0	0	44	18	0	0	5	2	0	0	
52P	2248 (4550)	2719 (4944)	2717 (4874)	2577 (4747)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (2)	1 (1)	0 (1)	0 (1)	1 (0)	1 (0)	0 (0)	0 (0)	25 (30)	17 (17)	4 (5)	4 (1)	4 (5)	1 (4)	0 (0)	0 (0)	81 (81)	29 (29)	29 (27)	6 (5)	27 (29)	5 (6)	8 (8)	4 (4)	
54L	3653 (3653)	3001 (3001)	3611 (3611)	3289 (3289)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3 (3)	3 (3)	4 (4)	3 (3)	2 (2)	2 (2)	0 (0)	0 (0)	41 (38)	31 (31)	6 (6)	5 (5)	7 (6)	6 (6)	0 (0)	0 (0)	96 (101)	54 (54)	10 (6)	7 (5)	14 (6)	6 (6)	0 (0)	0 (0)	
56FW	2722	2697	2665	2637	0	0	0	0	0	0	0	0	1	7	0	0	1	2	0	1	18	20	2	2	4	3	2	2	57	31	6	5	12	8	6	3	
64D	5956 (4757)	5956 (4542)	5967 (4629)	5692 (4289)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (13)	14 (13)	2 (2)	2 (2)	4 (3)	4 (3)	1 (2)	1 (1)	23 (30)	23 (20)	4 (5)	4 (3)	4 (2)	3 (2)	4 (4)	3 (3)	33 (47)	32 (26)	11 (20)	11 (9)	10 (10)	10 (6)	6 (8)	6 (6)	Sample 2
66JS	811	806	760	811	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	1	1	25	22	6	5	11	11	2	2	
68X	3354	3367	3376	3345	0	0	0	0	0	0	0	0	2	2	1	1	1	1	0	0	19	16	2	2	1	1	1	1	35	26	4	4	4	3	4	3	
70C**	1650	627	519	701	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	
74V	5723 (6216)	5750 (6249)	5782 (6312)	5728 (6250)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	32 (26)	22 (23)	4 (4)	2 (2)	17 (11)	9 (6)	8 (6)	4 (4)	68 (53)	28 (36)	9 (9)	4 (4)	44 (31)	10 (7)	15 (14)	5 (5)	241 (177)	40 (51)	13 (13)	5 (5)	117 (76)	14 (10)	30 (24)	6 (5)	
99X	3711 (6544)	4159 (6396)	4552 (6453)	4350 (6517)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (2)	3 (2)	1 (2)	1 (1)	2 (0)	1 (0)	1 (0)	1 (0)	16 (70)	9 (9)	4 (10)	1 (4)	9 (7)	4 (3)	6 (4)	3 (1)	51 (43)	17 (15)	8 (25)	3 (11)	25 (16)	11 (5)	16 (7)	5 (3)	

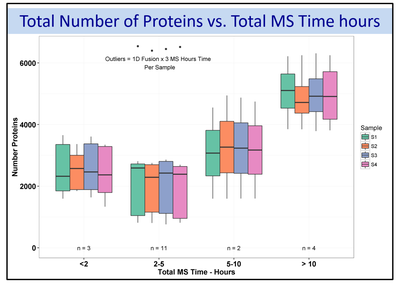
() Items updated with standardized Scaffold information
** Performed TMT analysis - exclude from further analysis
** Performed MRM analysis - exclude from further analysis

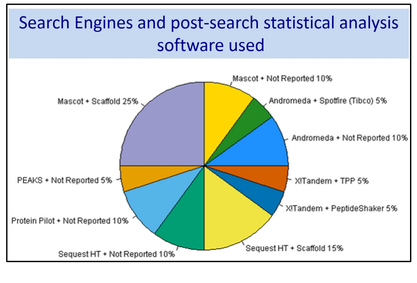
Self Reported Summary and Scaffold Corrected Results – Participants self-reported results through a Survey Monkey questionnaire and submitted results via DropBox. Where possible, search results (deposited as mzIdentML or Scaffold files) were re-analyzed via Scaffold at 1% Global FDR, 1 minimum peptide and 0% peptide confidence. (Unfortunately, mzIdentML files generated by 4 participating labs generated errors when loaded into Scaffold.) Updated results are in parenthesis next to the self-reported results. All but one participant's results re-analyzed via Scaffold changed.

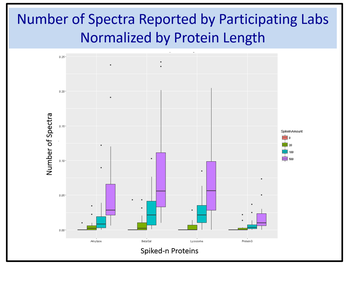
INFORMATICS ANALYSIS

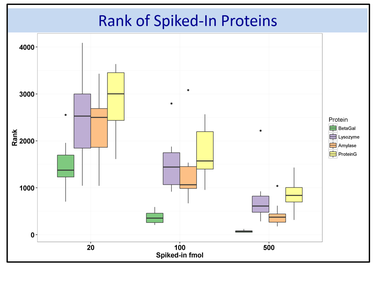












CONCLUSIONS

- **Participants reported a diverse range of protocols and results for this limit of detection study.**
- **While detection of the spiked-in proteins at the 100fmol and 500fmol levels was achieved by most participants, detecting proteins at 20 fmol level remains challenging.**
- **Depth of coverage increased with MS acquisition time for the majority of participants.**
- **Efficiency, as indicated by the Peptide Spectral Match (PSM) to Total Spectra ratio, did not predict Hela proteome depth of coverage, the PSM to Unique Peptide ratio or the number of unique peptides per protein identified.**
- **Acquisition rate did not have major effect on either the number of PSMs or the Efficiency.**
- **In order to identify all four spike-in proteins at the 20 fmol amount in 25 ug of Hela lysate, the Hela proteome needs to be profiled to a depth of 3,000 to 3,500 proteins. Consequently, fractionation at the protein or peptide level prior to LC-MS/MS analysis proved very helpful in achieving low level detection.**
- **Success in detecting low level proteins in a complex sample requires an optimized workflow and proven protocols. Highly sensitive, state-of-the art instrumentation is helpful, but in itself not sufficient to achieve this goal.**