Mass Spectrometry-Based Proteogenomic Characterization of the Human Gut Microflora from Crohn's Patients

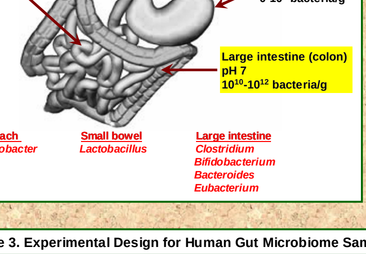
A. Russell[1], M. Rosenquist[2], M. G. Lefsrud[1], M. Shah[1], L. Engstrand[3], C. Tysk[4], J. Halfvarsson[4], N. C. VerBerkmoes[1], R. L. Hettich[1], J. Jansson[2]

1) Oak Ridge National Laboratory, Oak Ridge, TN 2) Swedish University of Agricultural Sciences, Uppsala, Sweden 3) Swedish Institute for Infectious Disease Control, Solna, Sweden 4) Orebro University Hospital, Orebro, Sweden

OVERVIEW

- Inflammatory bowel diseases (IBD) are increasing in prevalence in Europe and North America (current estimates are approximately 0.1% of the population).
- IBD can be divided into two disease categories: Ulcerative Colitis and Crohn's Disease. Crohn's is a chronic, relapsing, immunological mediated disorder that can have severe physical consequences.
- The current hypothesis is that this disease is due to an overly aggressive immune response to a subset of commensal enteric bacteria (Figure 1). Studies to date on IBD have suggested that the disorder may be caused by a combination of bacteria and host susceptibility.
- Currently, no study has been reported to utilize advanced systems biology techniques such as whole community genomics or proteomics for the characterization of the natural microbiome in Crohn's patients.
- In order to understand gut microbial community function and dynamics at a systems level, we propose to integrate information from large-scale molecular fingerprinting with integrated "proteogenomics."
- Fecal samples were obtained from healthy and diseased Crohn's patients. The complex microbiome was then extracted (Figure 2). Shotgun proteomics was used to characterize the microbial metaproteome from these samples (Figure 3).

Figure 1. Microbial Distribution in the Gastrointestinal Tract



EXPERIMENTAL

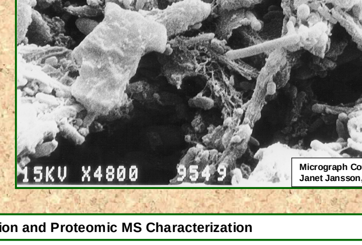
Samples and Sample Preparation

- Human fecal samples from healthy and diseased patients with Crohn's were extracted via density centrifugation to obtain enriched microbial pellets.
- These small pellets (10-100 mg wet weight) were processed via a single tube cell lyses and protein digestion. Briefly, the cell pellet was re-suspended in 6M Guanidine/10 mM DTT to lyse cells and denature proteins. The Guanidine concentration was diluted to 1 M with Tris buffer, and then sequencing grade trypsin was added to digest proteins to peptides. The complex peptide solution was de-salted via C18 solid phase extraction.

LC/LC-MS/MS and Informatics

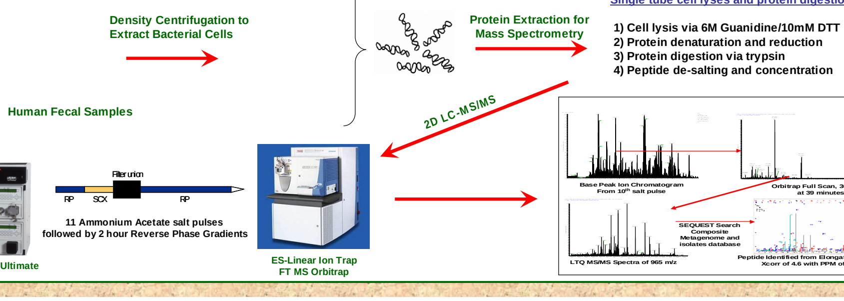
- All samples were analyzed via two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX-RP) on a LTQ-Orbitrap (Thermo Finnigan), with a 22 hour run per sample. All samples were analyzed in technical duplicates.
- The Orbitrap settings were as follows: 30K resolution on full scans in Orbitrap, all data-dependent MS/MS in LTQ (top five), 2 microscans for both Full and MS/MS scans, centroid data for all scans.
- All MS/MS spectra were searched with the SEQUEST algorithm [(enzyme type, trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed (internal lysine and arginine residues), and fully tryptic peptides only (both ends of the peptide must have arisen from a trypsin specific cut, except N and C-termini of proteins)] and filtered with DTASelect/Contrast at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3)]. Liberal filters were used for testing false positive levels [Xcorrs of at least 1.5 (+1), 2.0 (+2) 3.0 (+3)]. Only proteins identified with *two fully tryptic peptides* were considered for further biological study.
- Monoisotopic theoretical masses for all peptides identified by SEQUEST were compared to observed masses. Observed masses were extracted from .raw files, from full scan preceding best identified spectra.
- Database searches were preformed against the gut metagenome database (Gill, Science 2006) as well as a large number of known gut isolate genomes and human proteins (termed large database).
- A subset database search was performed only against the gut metagenome database (termed TIGR Metagenome 1 and 2) as well as two abundant gut isolate genomes.
- False positive levels were determined by conventional reverse database searching methods and compared with high mass accuracy results from the Orbitrap.

Figure 2. Microbial Diversity in Human Feces



Micrograph Courtesy of Janet Jansson, SLU

Figure 3. Experimental Design for Human Gut Microbiome Sample Collection and Proteomic MS Characterization



RESULTS AND DISCUSSION

Proteogenomics of the Human Gut Microbiome

- The long term goal of these studies is to apply proteogenomics techniques to understand the physiology of complex microbial communities in humans with and without Crohn's Disease.
- The short term goal of these studies was to test current shotgun proteomics techniques from sample collection and cell lysis to MS characterization and informatics.
- Current MS techniques are not capable of measuring the entire proteomes of all major and minor microbial species across hundreds of samples.
- Thus, the initial studies focused on determining what steps in the proteomics pipeline need to be optimized and improved to better enable the discovery of biomarkers and other disease related phenomenon in Crohn's disease.
- The general analytical method is described in Figure 3.
- Eight samples, from a double blind study, were extracted via density centrifugation to produce enriched microbial pellets. A fraction of the pellets were characterized for distribution and composition of the GI microbiota using molecular fingerprinting approaches; remaining microbial pellets were lysed via a single tube lyses method.
- The pellets (~10 mg) contained 100s-1000s of bacterial species and were run in duplicate by 2D-LC-MS/MS on a LTQ-Orbitrap. These samples contained chromatograms with 100,000s of peptides, as illustrated in Figure 3. Each replicate was run for 24 hrs.
- The MS/MS spectra were searched against two databases, the large database which includes the GI database, and a subset database (see Experimental).
- Table 1 illustrates the unique proteins, peptides, and spectra identified from each subject and run number from the large database.
- Table 2 (Run1 and Run2) shows the breakdown of each database entry type for a healthy (sample 8) and Crohn's sample (sample 10), along with technical replicates (data well represent entire dataset).
- Technical reproducibility of database hits via total spectral counts was very reproducible, while biological or experimental variability between samples was higher for both total proteins and abundances via spectral counts.

Table 1. Preliminary Identifications from Large Database

Subject ID	Disease ID	Unique Protein IDs	Total Protein IDs	Unique Peptide IDs	Unique Spectra
Subject 7 Run1	healthy	849	1851	1958	3735
Subject 7 Run2	healthy	1108	2407	2605	4077
Subject 8 Run1	healthy	1293	2627	3809	5099
Subject 8 Run2	healthy	1221	2289	2918	4868
Subject 9 Run1	Crohn's	940	2066	2224	3718
Subject 9 Run2	Crohn's	873	1935	2046	3613
Subject 10 Run1	Crohn's	968	1909	2280	3987
Subject 10 Run2	Crohn's	907	1833	1973	4068
Subject 11 Run1	Crohn's	698	1558	1646	2750
Subject 11 Run2	Crohn's	691	1517	1625	2754
Subject 12 Run1	Crohn's	478	1265	1385	3145
Subject 12 Run2	Crohn's	591	1454	1667	3479
Subject 13 Run1	Crohn's	697	1596	1909	3733
Subject 13 Run2	Crohn's	798	1794	2328	4176
Subject 14 Run1	Crohn's	657	1395	1547	2893
Subject 14 Run2	Crohn's	701	1519	1752	2968
Total Non-Redundant Protein Identifications: 3,935					

Table 2. Breakdown of Database Entries for Healthy and Diseased samples

Sample Set 8 Run1

Species	Proteins	Spectra	% of Total Spectra
Bifidobacterium	49	198	1.14
Bacteroides	105	548	3.15
GI	1261	9406	54.00
Metagenome 1	458	2622	15.05
Metagenome 2	754	4645	26.67
Totals	**2627**	**17419**	

Sample Set 8 Run2

Species	Proteins	Spectra	% of Total Spectra
Bifidobacterium	38	168	1.06
Bacteroides	94	480	3.02
GI	1021	8244	51.92
Metagenome 1	427	2531	15.94
Metagenome 2	709	4455	28.06
Totals	**2289**	**15878**	

Sample Set 10 Run1

Species	Proteins	Spectra	% of Total Spectra
Bifidobacterium	83	589	4.26
Bacteroides	21	104	0.75
GI	846	7716	55.84
Metagenome 1	406	2286	16.54
Metagenome 2	553	3124	22.61
Totals	**1909**	**13819**	

Sample Set 10 Run2

Species	Proteins	Spectra	% of Total Spectra
Bifidobacterium	89	557	4.08
Bacteroides	20	121	0.89
GI	796	7428	54.39
Metagenome 1	399	2325	17.02
Metagenome 2	529	3226	23.62
Totals	**1833**	**13657**	

Abundant Proteins and Species in Gut Samples

- From this initial study, a total of 3,935 unique proteins were identified, with ~800-1200 unique proteins per sample.
- The data in Table 1 and Table 2 were compiled from a very large GI database containing microbes thought to be in the intestinal tract, as well as all human proteins, and was merged with two microbial isolates and two human metagenomes from the gut microbiome.
- This database is effective for "fishing" trips to find novel proteins and potential species.
- Table 3 lists some of the representative bacteria identified in all samples.
- Table 4 illustrates the top 25 protein entries for healthy sample (sorted by total spectral count).
- Of the identified proteins from all samples, common metabolic proteins such as ribosomal proteins, glycolsis/TCA proteins, and chaperone proteins were routinely identified.
- Many hypothetical proteins were identified with high spectral counts, such as one very large hypothetical protein (200kDa), with an only known function annotated as a human protein "deleted in malignant brain tumor" (Figure 4).

Table 3. Partial List of Microbial Species Identified in the Human Gut Microbiomes from all Samples

Lactobacillus plantarum WCFS1	*Lactobacillus acidophilus NCFM*
Campylobacter jejuni	*Escherichia coli W3110*
Clostridium perfringens str. 13	*Clostridium perfringens ATCC 13124*
Bifidobacterium infantis	*Bifidobacterium longum NCC2705*
Bacteroides fragilis NCTC 9343	*Bacteroides thetaiotaomicron VPI-5482*

Figure 4. Hypothetical Protein Identified with High Spectral Counts

Locus	Sequence Count	Spectrum Count	% Coverage	Length	Molecular Weight	pI	Descriptive Name
gi|4758170|ref|NP_004397.1|	17	92	10.20%	1785	193941	5.4	deleted in malignant brain tumors 1 isoform a precursor [Homo sapiens]
gi|8923740|ref|NP_060049.1|	17	222	7.60%	2403	259574	5.4	deleted in malignant brain tumors 1 isoform c precursor [Homo sapiens]
gi|6633801|ref|NP_015568.1|	17	247	7.50%	2413	260755	5.4	deleted in malignant brain tumors 1 isoform b precursor [Homo sapiens]

Charge State	Spectral Count	XCorr	DeltCN	Obs_mono_m/z	Calc_mono_m/z	PPM	Delta_amu	Sequence
2	2	2.9123	0.3423	990.554	990.5616	-7.62	-0.0075	R.GRVEVLYR.G
1	2	2	0.2886	990.5557	990.5616	-5.91	-0.0059	R.GRVEVLYR.G
2	12	5.2972	0.6016	2374.949	2374.9588	-4.14	-0.0098	R.GSWGTVCDDYWDTNDANVVCR.Q
2	6	4.0545	0.5097	2930.3674	2930.3809	-4.62	-0.0135	R.QLGCGWAMSAPGNAQFGQGSGPIVLDDVR.C
1	6	2.7814	0.3489	1458.7402	1458.7472	-4.81	-0.007	Q.FGQGSGPIVLDDVR.C
2	26	3.4976	0.2752	1458.741	1458.7472	-4.26	-0.0062	Q.FGQGSGPIVLDDVR.C
2	1	5.6005	0.626	2271.9098	2271.9166	-3.01	-0.0068	R.GSWGTVCDDSWDTSDANVVCR.Q
2	6	4.342	0.5803	1487.6858	1487.6945	-5.82	-0.0087	R.QLGCGWATSAPGNAR.F
1	6	2.4599	0.4485	1487.6868	1487.6945	-5.15	-0.0077	R.QLGCGWATSAPGNAR.F
2	8	4.1651	0.4515	1517.6776	1517.6873	-6.37	-0.0097	R.QLGCGWAMSAPGNAR.F
1	2	2.5751	0.4384	1517.678	1517.6873	-6.11	-0.0093	R.QLGCGWAMSAPGNAR.F
1	1	1.9008	0.1552	N/A	1004.566	N/A	N/A	R.INLGFSNLK.L
2	4	4.5327	0.6056	2736.268	2736.2825	-5.31	-0.0145	R.SDISFQNTGFLAWYNSFPSDATLR.L
2	4	3.3787	0.5737	N/A	1419.6457	N/A	N/A	R.GSFTSSSNFMSIR.F
1	4	2.5288	0.4232	1419.6392	1419.6457	-4.61	-0.0065	R.GSFTSSSNFMSIR.F
1	1	2.008	0.1351	1074.5401	1074.5463	-5.75	-0.0062	R.FISDHSITR.R
2	1	3.7101	0.3528	1958.8776	1958.8798	-1.10	-0.0022	R.SGCVRDDTYGPYSSPSLR.I

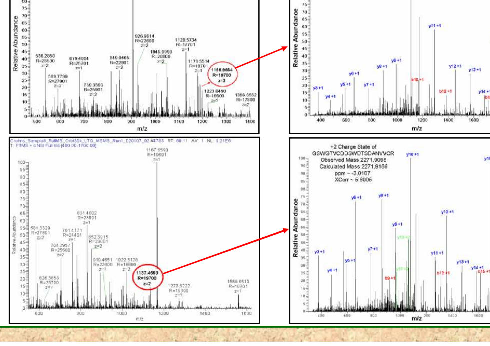


Table 4. Top 25 Protein Entries for Healthy Sample 8

Gene #	Peptide Count	Spectral Count	% Coverage	Length	Mass	pI	Function
gi|6633801|ref|NP_015568.1|	17	247	7.50%	2413	260755	5.4	deleted in malignant brain tumors 1 isoform b precursor [Homo sapiens]
gi|8923740|ref|NP_060049.1|	17	222	7.60%	2403	259574	5.4	deleted in malignant brain tumors 1 isoform c precursor [Homo sapiens]
gi|4758170|ref|NP_004397.1|	17	92	10.20%	1785	193941	5.4	deleted in malignant brain tumors 1 isoform a precursor [Homo sapiens]
TIGR_H8_159007ORF.2	19	66	61.80%	304	32398	5.7	glyceraldehyde-3-phosphate dehydrogenase
gi|52217358|dbj|BAD49951.1|	9	60	36.10%	233	25179	6.1	conserved hypothetical protein [Bacteroides fragilis YCH46]
gi|60493949|emb|CAH08740.1|	9	60	36.10%	233	25179	6.1	conserved hypothetical protein [Bacteroides fragilis NCTC 9343]
gi|109690306|ref|ZP_01377384.1|	1	55	2.20%	718	79192	5.4	hypothetical protein Cjejd_01001901 [Campylobacter jejuni subsp. doylei 269.97]
gi|112804268|gb|EAU01612.1|	1	55	2.20%	718	79192	5.4	polyribonucleotide nucleotidyltransferase [Campylobacter jejuni subsp. doylei 269.97]
TIGR_H8_159007ORF.2	16	54	61.70%	235	24733	5.6	glyceraldehyde-3-phosphate dehydrogenase
TIGR_H7_171636ORF.1	16	48	50.50%	303	32072	5.7	glyceraldehyde-3-phosphate dehydrogenase
gi|110675213|gb|ABG84200.1|	6	46	27.40%	230	24828	5.3	conserved hypothetical protein [Clostridium perfringens ATCC 13124]
gi|110683011|gb|ABG86381.1|	6	46	27.40%	230	24855	5.3	conserved hypothetical protein [Clostridium perfringens SM101]
gi|18144894|dbj|BAB80939.1|	6	46	27.20%	232	25073	5.3	hypothetical protein [Clostridium perfringens str. 13]
TIGR_H8_159766ORF.1	17	46	60.80%	337	37044	7.5	pyruvate ferredoxin oxidoreductase
gi|1685286|gb|AAC26399.1|	13	44	24.70%	445	48374	5.7	NAD(H) glutamate dehydrogenase [Bacteroides fragilis]
gi|52217785|dbj|BAD50378.1|	13	44	24.70%	445	48390	5.7	NADP-specific glutamate dehydrogenase
gi|55977839|sp|P94316|DHE2_BACFR	13	44	24.70%	445	48390	5.7	NAD-specific glutamate dehydrogenase (NAD-GDH)
gi|60494332|emb|CAH09128.1|	13	44	24.70%	445	48374	5.7	NAD-specific glutamate dehydrogenase 1 [Bacteroides fragilis NCTC 9343]
TIGR_H8_173299ORF.1	16	43	70.40%	223	23887	5	electron transfer flavoproteins beta
TIGR_H7_159766ORF.1	14	42	49.10%	330	36723	6.3	pyruvate ferredoxin oxidoreductase
TIGR_H8_GIM4W70TRORF.1	18	42	61.50%	309	35018	5	Chain A, T453a Mutant Of Pyruvate, Phosphate Dikinase
TIGR_H8_160248ORF.1	26	40	52.30%	390	42278	6.9	NADP-specific glutamate dehydrogenase
TIGR_H7_177047ORF.1	16	38	31.80%	481	53566	5.7	Phosphoenolpyruvate carboxykinase [ATP] (PEPcarboxykinase)
gi|50787940|emb|CAH05076.1|	15	36	33.90%	563	61491	8.5	putative plasminogen binding protein precursor [Bacteroides fragilis]
gi|52218629|dbj|BAD51222.1|	15	36	34.50%	553	60278	8.2	conserved hypothetical protein [Bacteroides fragilis YCH46]
gi|60495129|emb|CAH09950.1|	15	36	34.20%	559	61009	8.2	conserved hypothetical protein [Bacteroides fragilis NCTC 9343]
TIGR_H7_170765ORF.1	8	36	25.30%	336	35719	6.1	glyceraldehyde 3-phosphate dehydrogenases
gi|115444|sp|P06702|S10A9_HUMAN	9	35	66.70%	114	13242	6.1	Protein S100-A9 (S100 calcium-binding protein A9)
gi|29126810|gb|AAH47681.1|	9	35	66.70%	114	13242	6.1	S100 calcium binding protein A9 (calgranulin B) [Homo sapiens]
gi|4506773|ref|NP_002956.1|	9	35	66.70%	114	13242	6.1	S100 calcium-binding protein A9 [Homo sapiens]
gi|56205191|emb|CAI19494.1|	9	35	66.70%	114	13242	6.1	S100 calcium binding protein A9 (calgranulin B) [Homo sapiens]
TIGR_H8_177615ORF.1	8	35	18.20%	247	25764	5.4	extracellular glyceraldehyde-3-phosphate dehydrogenase
gi|110675296|gb|ABG84283.1|	8	34	15.20%	448	49008	6.3	glutamate dehydrogenase [Clostridium perfringens ATCC 13124]
gi|110682164|gb|ABG85534.1|	8	34	15.20%	448	48976	6.3	NADP-specific glutamate dehydrogenase [Clostridium perfringens SM101]
gi|18145180|dbj|BAB81224.1|	8	34	15.20%	448	49052	6.3	NADP-specific glutamate dehydrogenase [Clostridium perfringens str. 13]

Subset Database

- The large database was effective for initial proteome searching, but due to large redundancies and non-specific organism entries in some cases, detailed comparisons of database matching was complicated.
- One goal of this initial work was to compare the two Gill *et al*. gut metagenomes with the eight samples from this study (*metagenome samples and metaproteome samples are not at all related*).
- A subset database was made of only the two gut metagenomes and the two abundant sequenced microbes (experimental).
- Figure 5 is the species breakdown based on percentage of total spectra from the two technical replicates of each sample.
- Figure 6 is the same breakdown, but is based on unique spectral counts.

Figure 5. Total Spectra Percentages for Each Entry in Subdatabase

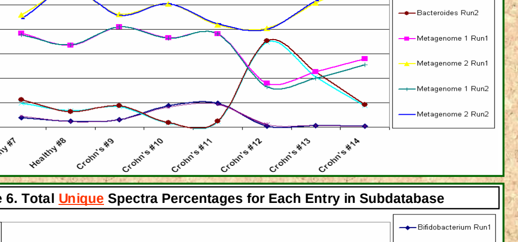
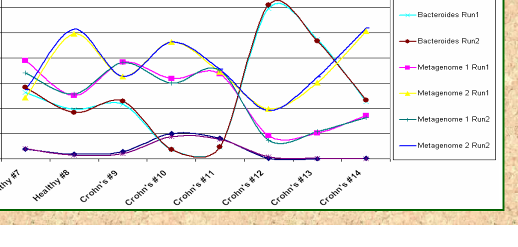


Figure 6. Total Unique Spectra Percentages for Each Entry in Subdatabase



False Positives and LTQ-Orbitrap

- One of the largest challenges in the application of shotgun proteomics to the complex microbial community of the gut will be incomplete or partial genome information.
- Since proteome datasets have to be compared to relevant reference metagenomes, there is the chance that a large number of measured peptides of high quality will not match database entries. This can increase the rate of false positives.
- Detailed discussion of false positives and LTQ-Orbitrap measurements can be found at Lefsrud Poster WP425 and Abraham Poster TP529.
- Similar methods of reverse database searching and filtering were used here. Again, very low levels of false positives were found for peptides with measured mass errors of less than 10 ppm (Table 5).

Table 5. Level of False Positives

Conservative Filters Sample Set 8 Run1	Forward Hits	Reverse Hits	% of False Positive Peptides
# of hits between 10 and -10 PPM	4068	2	0.10
# of hits above 10 and -10 PPM	438	20	8.73
# of hits with Undefined Parent Mass	368	3	1.62
Conservative Filters Sample Set 8 Run2	**Forward Hits**	**Reverse Hits**	**% of False Positive Peptides**
# of hits between 10 and -10 PPM	3828	2	0.10
# of hits above 10 and -10 PPM	406	30	13.76
# of hits with Undefined Parent Mass	432	6	2.74
Liberal Filters Sample Set 8 Run1	**Forward Hits**	**Reverse Hits**	**% of False Positive Peptides**
# of hits between 10 and -10 PPM	5121	26	1.01
# of hits above 10 and -10 PPM	1172	340	44.97
# of hits with Undefined Parent Mass	765	67	16.11
Liberal Filters Sample Set 8 Run2	**Forward Hits**	**Reverse Hits**	**% of False Positive Peptides**
# of hits between 10 and -10 PPM	5381	30	1.11
# of hits above 10 and -10 PPM	1192	330	43.36
# of hits with Undefined Parent Mass	686	70	18.52

CONCLUSIONS

- We have applied shotgun proteomics to 8 complex gut microbiome samples from healthy and diseased subjects (with only indirect metagenome information).
- Preliminary data shows great promise for this technique, with excellent technical reproducibility for both microbial species and proteins identified.
- False positive levels are very low with the LTQ-Orbitrap when mass accuracies less than 10 ppm can be obtained. This will be essential for environmental or community microbial proteome measurements.
- Improvements are necessary to generate reproducible and efficient lyses and protein extraction between samples, as well as increase the depth of proteome measurements.
- These studies have made it clear that it will be critical to integrate molecular fingerprinting with whole community genomics and proteomics in related microbiome samples.

ACKNOWLEDGMENTS

- This research sponsored by U.S. Department of Energy, Genomics:Genomes-to-Life Program under contract DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.
- We thank Dr. David Tabb and the Yates Proteomics Laboratory at Scripps Research Institute for DTASelect/Contrast software, the Institute for Systems Biology for proteome bioinformatics tools used in analysis of the MS data, and Dr. F. Larimer and M. Land of the ORNL Genome Analysis and System Modeling Group for computational resources for proteomic analysis.