Investigation of Protein Sequence Variants in a Natural Microbial Community by Top-Down Proteomics

Brian Erickson[1]; Mark Lefsrud[1]; Nathan VerBerkmoes[1]; Steven Singer[2]; Michael Thelen[2]; Jillian Banfield[3]; Robert Hettich[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN; [2]Lawerence Livermore Natl. Lab, Livermore, CA; [3]University of California, Berkeley, CA

OVERVIEW

- **The complexity of a natural microbial community presents unique and technical challenges for proteomics based discoveries.**
- **"Bottom-up" and "top-down" proteomics techniques can be integrated to identify the protein members and their isoforms within the extracellular portion of a natural microbial community.**
  - **These protein isoforms can include sequence variants and various post-translational modifications (PTMs).**
- **Within the community, a series of cytochrome related proteins were identified.**
  - **FTICRMS measurement along with IRMPD fragmentation revealed variant versions of the genome-predicted cytochrome-579.**
  - **Characteristic sequence tags from each cytochrome were used to identify the predicted genome cytochrome versus the variant cytochrome**
- **"Top-down" LC–MS identified 333 non-redundant proteins.**
  - **The identifications included 262 proteins with predicted methionine truncations and 36 with signal peptide cleavages.**

INTRODUCTION

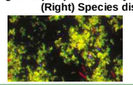
- **Proteome characterization of microbial species from natural environmental communities is challenged by the fact that these consortia are not clonal but rather encompass clades (i.e. bacterial cousins).**
- **This leads to a substantial amount of strain variation in related proteins, precluding the easy identification of peptides by database searching algorithms in many cases.**
- **Combining the advantages of high-throughput "bottom-up" MS with the high resolution intact mass measurements of "top-down" analyses will provide a comprehensive picture of the proteins present.**
- **Furthermore, the identification of post-translation modifications can provide significant insight into protein function.**
- **The identification of signal peptides can provide support for the localization and functional role of the proteins.**

Acid Mine Drainage (AMD)

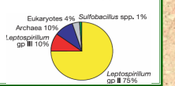
- **Iron Mountain Mine, CA (near Redding, CA)**
- **Biofilm exists in pH < 1, ~42°C, Molar/sub-molar concentrations of Fe, Zn, Cu, As**
- **Genome sequence (Tyson et al., 2004, *Nature*)**
- **Proteome Characterization (Ram et al., 2005, *Science*)**

Figure 1 (A) Acid Mine Drainage (B) Biofilm Collection (C) Location of AMD
(Images: Dr. Jillian Banfield)

Figure 2 (Left) FISH analysis – Yellow: *Leptospirillum* group II Red: *Sulfobacillus*
(Right) Species distribution (Tyson et al., 2004)



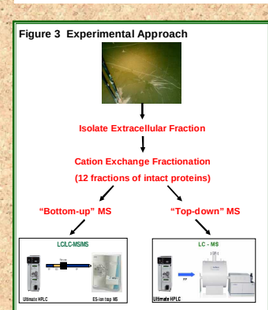
Four abundant members
- ***Leptospirillum* II (most abundant)**
- ***Leptospirillum* III**
- ***Sulfobacillus***
- ***Ferroplasma* (archaea)**



EXPERIMENTAL

- **Sample Preparation**
  - **The extracellular portion of whole biofilms were isolated by centrifugation (without cell lysis).**
  - **The resulting sample was then further separated by cation exchange, resulting in 12 fractions.**
  - **Each fraction was divided in half by volume for analysis by "bottom-up" and "top-down" MS.**
- **"Bottom-up"**
  - **Samples were denatured, reduced, and proteolytically digested into peptides with trypsin.**
  - **Peptides were measured via a 1D (RP)-LC-MS/MS pipeline utilizing a linear ion trap (Thermo Finnigan LTQ).**
- **"Top-down"**
  - **Intact proteins were analyzed by a 1D–LC-MS pipeline with all measurements on a 9.4T IonSpec FTICRMS.**
- **Protein ID**
  - **"Bottom-up" spectra were searched with DBDigger and filtered with DTASelect. (Tabb *et al., Analytical Chemistry*. 2005)**
  - **"Top-down" spectra were identified with PTMSearchPlus.**

Figure 3 Experimental Approach



RESULTS

- **Given the unique and hostile environment of the AMD biofilm, it is expected that a diverse and functionally important complement of proteins will be present in the extracellular portion. Identifying not only the proteins present, but also the degree of protein modification can provide valuable insight into their individual specific functions as well as how the microbial community exists as a whole.**
- **LC-FTMS**
  - **On-line HPLC allowed for spatial protein separation of the complex biofilm samples. Parent ion selection was achieved in a data dependent manner utilizing dynamic exclusion as well as ion intensity for selection. (Figure 4)**
- **Cytochrome identification**
  - **Two separate off-line HPLC fractions (C1 and C2) consisting of cytochrome 579 were isolated and examined by ESI-FTMS for interrogation of the intact proteins.**
  - **The molecular masses for these two variants of cytochrome 579 were distinct (Figure 5A/5B)**
  - **Dissociation of these intact proteins was achieved with infrared multiphoton dissociation (IRMDP), and generated fragment ions that could infer sequence tag information (Figure 5C/5D).**
  - **Through DNA sequencing it has been determined that the genome predicted cytochrome and the newly identified sequence variant share 88% sequence identity.**

Figure 4 – Data Dependent LC – MS (5way_CG_LeptoII_scaff_8_GENE_1)

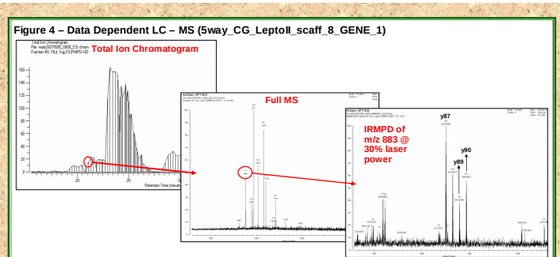
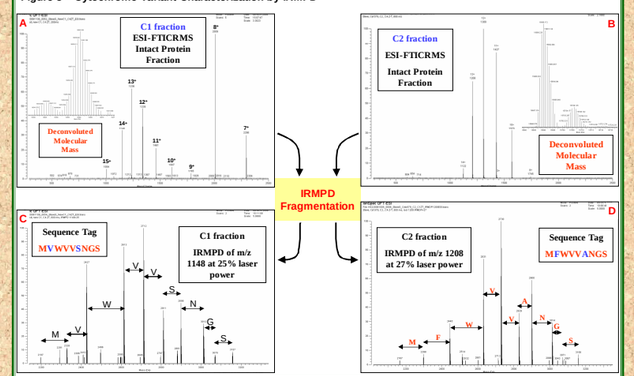


Figure 5 – Cytochrome Variant Characterization by IRMPD



- **Protein Identification**
  - **"Top-down" analysis identified 333 proteins while "bottom-up" analysis identified 215 proteins. Within the identifications, a large number of hypothetical proteins were found. The top-down measurements revealed 262 proteins that contained an N-terminal methionine truncation, and 36 proteins that contained a signal peptide cleavage. (Table 1)**
  - **Protein identifications were verified at high resolution by comparing theoretical vs. experimental isotopic distributions. (Figure 6)**
  - **"Top-down" identifications include proteins from *Leptospirillum* II/III, *Ferroplasma*, *G-plasma*, as well as unassigned archeal and bacterial species. The high mass resolution and accuracy (low ppm) increases the confidence of the protein assignments. (Table 2)**
  - **Spectra were searched with an N-terminal methionine truncation, resulting in hundreds of putative identifications. Examination of the second amino acid residue provided biological support for these identifications. (Table 3)**
  - **The mass spectra were also searched with an appended database containing protein sequences that include signal peptide cleavage. (Table 4)**
  - **The composite protein identifications were analyzed for species of origin and resulted in a distribution correlating well with previous studies of the biofilm. As before, *Leptospirillium* group II was the most dominant. (Figure 7)**

Table 1 Protein Identifications

"Top-down" Identifications	333
"Bottom-up" Identifications	215
Methionine Truncation	262
Signal Peptide Cleavage	36

Figure 6 Theoretical vs. Experimental Isotopic Distribution (5way_CG_LeptoII_scaff_8_GENE_1)

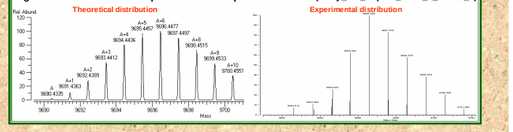


Table 2 Partial List of "Top-down" Identifications (333 Total)

Protein name	Protein description	Calculated Mass (Da)	Observed Mass (Da)	PPM
UBA_Leptoll_Scaffold_6135_GENE_46	Uncharacterized conserved protein	10064.021	10064.021	0.0
fer1_scaff_553_GENE_4	Hypothetical protein	12148.017	12148.017	0.0
UBA_Leptoll_Scaffold_8241_GENE_655	Hypothetical protein	13216.951	13216.951	0.0
fer2_scaff_10_GENE_90	CytidylyltrRNA synthetase	15409.446	15409.446	0.0
UBA_Leptoll_397_37	E-Glycine cleavage system H protein (lipoate-binding)	14042.828	14042.827	0.1
gpt_scaff_105_GENE_7	Dehydrogenase (short-chain alcohol dehydrogenases)	28433.295	28433.289	0.1
5way_CG_LeptoII_scaff_8_GENE_1	Hypothetical protein	9696.871	9696.873	0.2
UBA_Leptoll_Scaffold_8524_GENE_180	Hypothetical protein	9696.871	9696.873	0.2
UBA_Leptoll_Scaffold_8049_GENE_275	Hypothetical protein	12052.827	12052.825	0.2
gpt_scaff_126_GENE_3	5'-3' exonuclease (including N-terminal domain of PolI)	14041.926	14041.925	0.2
5way_CG_LeptoII_scaff_76_GENE_21	Ferredoxin	12844.499	12844.502	0.3
Unass_arch_scaff_8062_GENE_34	Ubiquitin product	13174.130	13174.134	0.3
5way_CG_LeptoII_scaff_303_GENE_2	Adenylosuccinate reductase	12843.874	12843.878	0.3
Unass_arch_scaff_733_GENE_3	Hypothetical protein	12054.098	12054.093	0.4
5way_CG_LeptoII_scaff_374_GENE_1	Hypothetical protein	8973.303	8973.307	0.5
UBA_Leptoll_Scaffold_8135_GENE_15	Hypothetical protein	24829.289	24829.304	0.6
fer1_scaff_249_GENE_1	Amino acid transporters	21167.384	21167.371	0.6
fer2_scaff_307_GENE_2	Hypothetical protein	11640.258	11640.249	0.7
fer2_scaff_119_GENE_5	Hypothetical protein	12053.136	12053.129	0.8
UBA_Leptoll_232_7	Hypothetical protein	14041.414	14041.427	0.9
gpt_scaff_269_GENE_8	RNA-binding protein Rpa4	24828.718	24828.738	1.0
UBA_Leptoll_352_18	Transposase	28432.928	28432.960	1.2
UBA_Leptoll_172_2	Hypothetical protein	20006.233	20006.206	1.3
Unass_arch_scaff_506_GENE_7	Hypothetical protein	19372.482	19372.508	1.4
fer1_scaff_176_GENE_1	Precorrin isomerase	19144.992	19144.968	1.4
UBA_Leptoll_Scaffold_8135_GENE_107	Predicted site-specific integrase-resolvase	7224.451	7224.449	1.5
5way_CG_LeptoII_scaff_59_GENE_15	Cytidinedeoxycytidylate deaminase	18082.600	18082.628	1.6
fer1_scaff_219_GENE_2	Ribosomal protein	17240.265	17240.256	1.7
UBA_Leptoll_274_2	Transposase	16454.881	16454.909	1.7
5way_CG_LeptoII_scaff_79_GENE_9	Transposase	6513.394	6513.408	2.1
UBA_Leptoll_Scaffold_8062_GENE_180	Mg-dependent Dnase	28793.881	28793.865	2.2
gpt_scaff_450_GENE_1	Transferrinbiase	16456.039	16456.096	3.5
UBA_Leptoll_Scaffold_8049_GENE_295	Transposase	16876.303	16876.237	3.9
UBA_Leptoll_Scaffold_8241_GENE_177	Transposase	19049.772	19049.848	4.0
UBA_Leptoll_372_20	Ubiquinone oxidoreductase subunit 3 (chain A)	14325.071	14325.140	4.8
UBA_Leptoll_Scaffold_8027_GENE_48	Biopolymer transport protein	16668.035	16668.234	11.9
gpt_scaff_609_GENE_3	Permease	29567.654	29567.201	15.3
5way_CG_LeptoII_scaff_3_GENE_96	Multiple antibiotic resistance	21450.754	21451.149	18.4

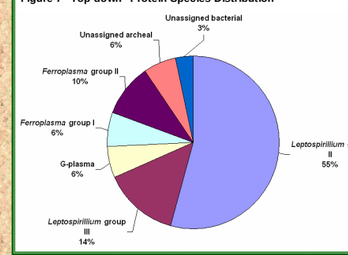
Table 3 Partial List of Proteins with Methionine Truncation (262 Total)

Protein name	Protein description	2nd AA after Met	Calculated Mass (Da)	Observed Mass(Da)	PPM
UBA_Leptoll_Scaffold_8241_GENE_554	Hypothetical protein	G	24828.8325	24828.83239	0.0
fer1_Isolate_689	Conserved hypothetical protein	E	7146.2828	7146.282764	0.0
UBA_Leptoll_379_15	Hypothetical protein	Y	9539.6496	9539.649481	0.0
5way_CG_LeptoII_scaff_86_GENE_19	Hypothetical protein	S	22051.1217	22051.12136	0.0
gpt_scaff_196_GENE_18	Hypothetical protein	N	13314.1218	13314.12151	0.0
fer1_Isolate_151c	Inorganic pyrophosphatase	K	19049.8468	19049.84796	0.1
5way_CG_LeptoII_scaff_150_GENE_15a	Ribosomal protein L34	S	5100.0811	5100.080697	0.1
UBA_Leptoll_Scaffold_8241_GENE_468a	Ribosomal protein L34	S	5100.0811	5100.080697	0.1
UBA_Leptoll_Scaffold_8049_GENE_307	Undefined product	G	15410.3741	15410.37197	0.1
5way_CG_LeptoII_scaff_22_GENE_15	Hypothetical protein	N	12862.5013	12862.50404	0.2
UBA_Leptoll_Scaffold_8062_GENE_162	Lactoylglutathione lyase	K	16510.482	16510.48566	0.2
fer2_scaff_89_GENE_13	Hypothetical protein	A	15409.7883	15409.78468	0.2
UBA_Leptoll_Scaffold_8049_GENE_262a	Hypothetical protein	G	9695.7711	9695.773381	0.2
UBA_Leptoll_Scaffold_8241_GENE_320	Ribosomal protein L27	A	10070.3452	10070.34841	0.3
gpt_scaff_133_GENE_15	Hypothetical protein	D	13028.9276	13028.92287	0.4
fer2_scaff_80_GENE_11	Hypothetical protein	Y	8489.8904	8489.877725	0.4
UBA_Leptoll_Scaffold_8692_GENE_162	Ribosomal protein S17	S	10583.4391	10583.44312	0.4
fer1_Isolate_151c	Undefined product	K	10675.3343	10675.33938	0.4
UBA_Leptoll_Scaffold_8524_GENE_197	Undefined product	T	14110.0122	14110.00514	0.5
5way_CG_LeptoII_scaff_146_GENE_1	Cold shock protein	A	7227.8763	7227.882857	0.6
fer1_scaff_143_GENE_3	Hypothetical protein	R	8837.519	8837.513065	0.7
fer1_scaff_812_GENE_3	Superfamily II DNA and RNA helicases	D	22140.9493	22140.96464	0.7
Unass_bact_scaff_456_GENE_4	Hypothetical protein	K	16625.1411	16625.15336	0.7
fer2_scaff_56_GENE_23	Hypothetical protein	E	13216.1597	13216.14867	0.7
gpt_scaff_689_GENE_1	Hypothetical protein	A	8612.0668	8612.058662	0.8
UBA_Leptoll_398_37	Hypothetical protein	I	6207.1567	6207.151501	0.8
UBA_Leptoll_398_108	Hypothetical protein	T	5883.655	5883.650009	0.8
UBA_Leptoll_371_28	Hypothetical protein	A	6514.4992	6514.493088	0.9
5way_CG_LeptoII_scaff_6_GENE_22	Hypothetical protein	S	13026.8824	13026.89475	0.9
UBA_Leptoll_Scaffold_8524_GENE_126	Undefined product	G	13026.8824	13026.89475	0.9
UBA_Leptoll_345_6	Hypothetical protein	I	18938.27	18938.26164	1.0
fer1_Isolate_176b	Hypothetical protein	Y	10582.9321	10582.92182	1.0
UBA_Leptoll_383_34	Hypothetical protein	P	13027.6922	13027.6922	1.0
Unass_arch_scaff_730_GENE_6	Ribosomal protein L1	R	16436.1151	16436.1956	1.2
fer2_scaff_93_GENE_9	Hypothetical protein	Y	14096.8348	14096.81408	1.5
UBA_Leptoll_364_15	Hypothetical protein	S	7859.8772	7859.965395	1.5
5way_CG_LeptoII_scaff_27_GENE_50	Hypothetical protein	R	10577.1177	10577.13413	1.6
UBA_Leptoll_Scaffold_7931_GENE_426	Hypothetical protein	R	10577.1177	10577.13413	1.6
fer1_scaff_126_GENE_3	Hypothetical protein	D	9693.1208	9693.105504	1.6

Table 4 Signal Peptide Identifications by "Top-down" MS (36 Total)

Protein name	Protein description	Calculated Mass (Da)	Observed Mass (Da)	PPM
UBA_Leptoll_Scaffold_8241_GENE_297_SigP	Hypothetical protein	13216.6417	13216.66745	0.4
fer2_scaff_37_GENE_36_SigP	Putative threonine efflux protein	18092.7265	18092.73637	0.5
Unass_arch_scaff_711_GENE_3_SigP	Putative threonine efflux protein	18092.7265	18092.73637	0.5
Unass_bact_scaff_969_GENE_1_SigP	Hypothetical protein	12198.7605	12198.75423	0.5
UBA_Leptoll_Scaffold_8241_GENE_21_SigP	Putative phospholipid binding protein	18082.616	18082.6285	0.7
UBA_Leptoll_Scaffold_8241_GENE_455a_SigP	Conserved hypothetical protein	6026.1841	6026.178735	0.9
gpt_scaff_142_GENE_10_SigP	Hypothetical protein	6207.4689	6207.474665	0.9
UBA_Leptoll_370_11_SigP	O-Peptide methionine sulfoxide reductase	21401.1278	21401.14755	0.9
UBA_Leptoll_Scaffold_8062_GENE_53_SigP	Hypothetical protein	12413.7429	12413.7647	1.8
UBA_Leptoll_368_4_SigP	Hypothetical protein	6019.3699	6019.385663	2.6
UBA_Leptoll_Scaffold_8524_GENE_128_SigP	Hypothetical protein	18935.7818	18935.72904	2.8
5way_CG_LeptoII_scaff_178_GENE_9_SigP	Hypothetical protein	10583.9861	10584.01693	2.9
UBA_Leptoll_Scaffold_8049_GENE_321_SigP	Hypothetical protein	10583.9861	10584.01693	2.9
UBA_Leptoll_Scaffold_8241_GENE_236_SigP	Hypothetical protein	16047.0234	16046.96653	3.5
fer2_scaff_29_GENE_54_SigP	Hypothetical protein	8973.339	8973.307076	3.6
UBA_Leptoll_389_19_SigP	Hypothetical protein	7408.232	7408.193882	5.1
UBA_Leptoll_Scaffold_8049_GENE_83_SigP	Hypothetical protein	15424.5317	15424.41143	7.8
fer1_scaff_65_GENE_2_SigP	Hypothetical protein	11205.0471	11205.05421	8.3
UBA_Leptoll_369_1_SigP	Hypothetical protein	14022.47	14022.33623	9.8
gpt_scaff_456_GENE_4_SigP	K+-transporting ATPase	20224.8929	20224.68808	10.1
UBA_Leptoll_325_11_SigP	Hypothetical protein	10600.0758	10600.22247	13.8
5way_CG_LeptoII_scaff_62_GENE_26_SigP	Hypothetical protein	21551.1615	21550.62283	15.7
UBA_Leptoll_Scaffold_8241_GENE_148_SigP	Hypothetical protein	25450.5014	25450.05592	17.5
5way_CG_LeptoII_scaff_21_GENE_56_SigP	Hypothetical protein	9677.3588	9677.181859	18.3
UBA_Leptoll_Scaffold_8241_GENE_81_SigP	Hypothetical protein	9677.3588	9677.181859	18.3
UBA_Leptoll_359_19_SigP	Hypothetical protein	14030.0249	14030.30506	20.0
Unass_bact_scaff_930_GENE_4_SigP	Hypothetical protein	14030.0249	14030.30506	20.0
UBA_Leptoll_377_5_SigP	Hypothetical protein	22036.8429	22037.33092	22.1
UBA_Leptoll_391_4_SigP	Hypothetical protein	20009.5075	20009.05831	22.4
5way_CG_LeptoII_scaff_374_GENE_7_SigP	Hypothetical protein	15643.6633	15644.04871	24.6
UBA_Leptoll_Scaffold_8241_GENE_153_SigP	Hypothetical protein	15643.6633	15644.04871	24.6
fer2_scaff_139_GENE_13_SigP	Hypothetical protein	4846.6973	4846.823344	26.0
fer2_scaff_64_GENE_13_SigP	Hypothetical protein	6608.8259	6608.620635	31.1
5way_CG_LeptoII_scaff_830_GENE_1_SigP	Hypothetical protein	10534.7084	10535.12344	39.4
UBA_Leptoll_242_1_SigP	Hypothetical protein	11434.1016	11433.63601	40.7
5way_CG_LeptoII_scaff_490_GENE_3_SigP	Hypothetical protein	8389.6229	8390.103115	57.2

Figure 7 "Top-down" Protein Species Distribution



DISCUSSION

- **Methionine truncation**
  - **N-terminal methionine truncation is an essential modification thought to occur in ~55–70% of proteins, especially extracellular species. Expected residues following the methionine include A, C, G, P, S, T, and V. (Giglione et al., 2004, *Cell. Mol. Life Sci.*)**
  - **Within this study, "top-down" MS identified 262 proteins with possible methionine truncation.**
- **Signal Peptide Identification**
  - **Cellular location is an important functional diagnostic tool. The presence of a cleaved signal peptide supports the extracellular location of the protein and can provide insight into the functional role of the protein.**
  - **Utilizing an "in-house" developed script, all predicted proteins in the biofilm database were searched for possible signal peptides using SignalP (Bendtsen et al., 2004, *J. Mol. Biol.*)**
  - **Proteins predicted to contain a putative signal peptide were appended to the database with the signal peptide sequence removed along with "SigP" appended to the gene name. "Top-down" spectra were searched with the appended signal peptide database using PTMSearchPlus (*in-house software, V. Kertesz*).**
  - **Initial results indicate 36 proteins are predicted to have signal peptide cleavage.**
  - **For all 36 predicted signal peptide cleaved proteins, the genome predicted full length protein was not identified, increasing confidence in the signal peptide prediction.**

CONCLUSIONS

- **Variations in amino acid sequence between proteins can be distinguished by high resolution FTICRMS along with IRMPD fragmentation.**
- **"Top-down" LC-MS identified 333 proteins. Additionally, 262 proteins were identified to contain methionine truncation.**
- **Combining *in-silico* prediction methods along with "top-down" LC-MS resulted in 36 proteins predicted to contain signal peptide cleavage.**
- **Possible post-translational modifications can be identified in a high-throughput manner with the use of LC-FTMS.**
- **The abundance of extracellular proteins appear to correlate well with the abundant microbial membership of the biofilm.**

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.**
- **B. Erickson acknowledges financial support from the UTK-ORNL GST program.**
- **Vilmos Kertesz (ORNL) is acknowledged for development and implementation of the PTMSearchPlus software algorithm for integrating top-down and bottom-up MS data.**