O-062

Enhanced Proteomics Measurements of a Natural Microbial Community to Reveal Information about Community Structure and Metabolic Potential

N.C. VerBerkmoes[1], R.J. Ram[2], V.J. Denef[2], M. Thompson[1], G.W. Tyson[2], B.J. Baker[2], M. Shah[1], M.P. Thelen[3], R.L. Hettich[1] and J. Banfield[2]
1) Oak Ridge National Laboratory, Oak Ridge, TN 2) University of California, Berkeley, CA 3) Lawrence Livermore National Laboratory, Livermore, CA

Contact Information:
Nathan C. VerBerkmoes, Oak Ridge National Laboratory,
P.O. Box 2008, MS-6131 Oak Ridge, TN, 37831
Phone: 865-576-4546, Fax: 865-576-8559,
E-mail: verberkmoesn@ornl.gov

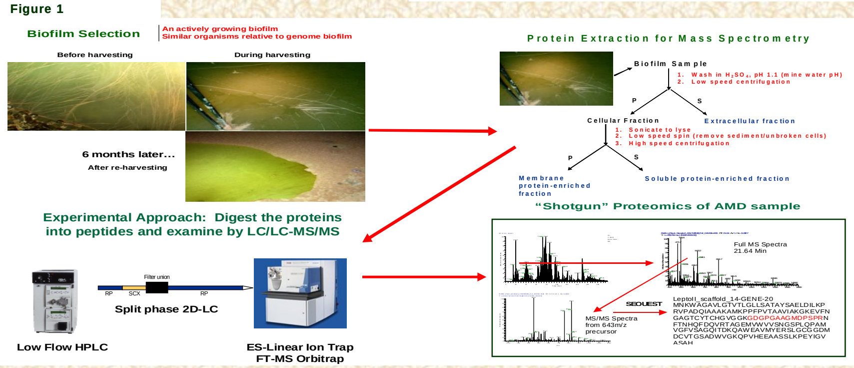
ABSTRACT

Microbial communities play key roles in the Earth's biogeochemical cycles, but for the most part are very poorly understood. For example, microorganisms from a number of lineages thrive in acid mine drainage (AMD), one of the most extreme environments on Earth. We have completed a proteogenomic investigation of a natural AMD consortia sample to determine the biochemical basis for adaptation (Ram et al, Science, 2005). This study illustrated the potential and challenges for MS-based proteome characterization of complex proteomes from environmental samples.
Two major challenges presented by these complex mixtures include 1) processing and analyses of small sample quantities, and 2) microbial strain variation from original community genome sequences. For the complete characterization of the proteome of natural microbial communities, we have developed a systematic approach to push the capabilities of current mass spectrometry (MS) techniques as well as integrate the next generation of MS technology. We have optimized techniques for lyses, digestions and LC-MS/MS analysis of small starting quantities (~1mg) of biofilms and artificial mixtures; this will allow for detailed spatial analysis of biofilm mats. We have successfully identified over 2000 proteins from such mixtures in a single analysis. To investigate the strain variation issue, we have conducted computational simulations on the AMD genome and proteome which indicate that a 2-5% random substitution at the amino acid level results in a major loss of protein identification. While ~2800 proteins can be identified from the original dataset, a 5% random substitution reduces this number to ~1700 proteins and a 20% level results in only 170 identifications. We have implemented new mass spectrometry approaches based on high mass accuracy MS and MS/MS followed by MS3 coupled with *de-novo* sequencing algorithms to address this major problem. For example, the extracellular fraction was dominated by a novel protein shown to be a cytochrome central to iron oxidation and AMD formation. This protein was found to be abundant in multiple AMD samples, with the exact same two sequence variations.

INTRODUCTION

- Recent acquisition of genomic data directly from natural samples has begun to reveal the genetic potential of communities (G. W. Tyson, Nature 2004) and environments (J. C. Venter, Science 2004). The ability to obtain whole or partial genome sequences from microbial community samples has opened up the door for microbial community proteomics.
- The challenges of obtaining useful proteomic information from these samples are diverse and daunting.
- Our methodology for community proteome analyses involves cell lyses, protein fractionation, protein denaturation and digestion with trypsin followed by automated analysis with 2-dimensional nano-LC-MS/MS Level one

Figure 1



METHODS

Samples

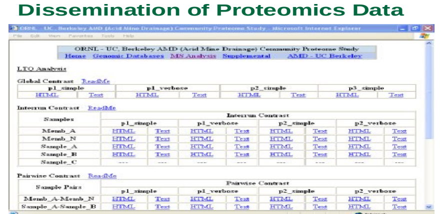
- The Acid Mine Drainage (AMD) biofilm samples used in this study and prior community genome sequencing (Tyson, Nature 2004) were collected from the underground regions of the Richmond Mine at Iron Mountain near Redding, California (USA). These pink biofilms grew on the surface of acidic (pH ~0.8) sulfuric acid-rich, hot (~42°C), metal-contaminated solutions. All proteome samples were denatured and reduced, digested with trypsin, desalted, and concentrated prior to analyses. The exact methodology depended on sample type and experimental goal.
- PCR analyses of Scaff_14_Gene_20 was used to verify SNPs in this protein.

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 either three-dimensional quadrupole ion traps (LCQ) or linear ion traps (LTQ) or LTQ-Orbitrap (Thermo Finnigan) (Figure 1).
- All MS/MS spectra were searched with the SEQUEST algorithm (Thermo Finnigan), and filtered with DTASelect at the peptide level [Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3) were used in all cases]. Only proteins identified with *two fully tryptic peptides* were considered for further biological considerations.
- Databases included the predicted AMD proteome from the original genome sequencing project (Tyson, Nature 2004) or a the *R. palustris* isolate database. Appended reverse database were employed to test false positive levels.
- All data files from the initial AMD study, databases and resulting MS/MS identifications, can be downloaded from the AMD Proteome Website Analysis Page (http://compbio.ornl.gov/biofilm_amd/). ***Linkable spectra for identified peptides are downloadable, a step towards open access proteome results*** (Ram, Science, 2005) (Figure 2).

Figure 2

Dissemination of Proteomics Data



RESULTS

Initial AMD Proteome Characterization

- A combined proteogenomic characterization was applied to a section of AMD biofilm collected from AB drift of the Richmond Mine (Ram et al, Science, 2005).
- We have used the same approach for the characterization of the AMD proteomes from a variety of sites from the mine (Figure 3) (Table 1).
- The UBA site is of great interest since shotgun DNA sequencing and annotation of the exact same site is almost finished allowing for the *first direct comparison of genome and proteome from the same environmental sample.*

Figure 3

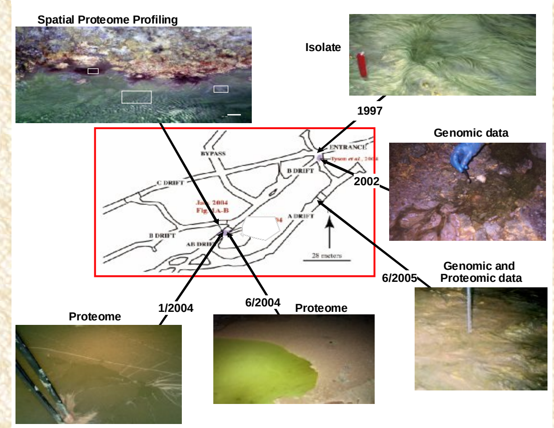


Table 1

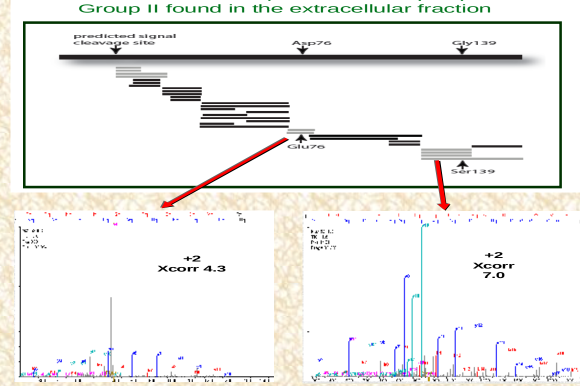
Sample	Proteins IDs	Comment
AB End	2,036	Ram et al, 2005
AB Front Sample 1	2,020	High G-Plasma Abundance
AB Front Sample 2	1,902	Still In Progress
UBA	2,107	Genome Sequence almost ready
C-Pink	1,378	Limited Fractionation for quick screens

Abundant biofilm proteins and SNPs

- The most abundant proteins are likely to be critical to the biofilm community. Determining relative or absolute abundance of individual proteins is a recognized challenge in MS-based proteomics. % sequence coverage, # of unique peptide hits, and MS/MS spectral counts are all potential indicators for protein abundance.
- By looking at proteins with high spectral counts but not complete sequence coverage it is possible to hypothesis about areas of proteins modified by N-terminal cleavages, post-translational modifications and amino acid variants.
- Of the proteins highly enriched in the extracellular fraction was found with very high spectral count but average sequence coverage. Further analyses by PCR and N-terminal cleavage prediction predicted an N-terminal cleaved peptide and two SNPs. Figure 4 describes how these were all verified by MS-proteomics.
- The variant was first characterized in the AB end sample but has also now been found in the other community samples.

Figure 4

Characterization of Scaff_14_GENE_20
An unknown abundant protein from *Leptospirillum* Group II found in the extracellular fraction



LeptoII_scaff_14_GENE_20	
Protein Region Genome Database	**Spectral Counts**
...HQFDQVR...	3
...DWVGKQP...	5
Protein Region SNP Database	
...HQFEQVR...	162
...DWVSKQP...	274

The effects of SNPs on protein identification

- Protein identification from community samples will almost always rely on a composite genome or reference genome related to the environmental sample being analyzed.
- Thus quite often, as shown in Figure #4, amino acids substitutions will occur at some random rate among proteins from the various species.
- Table 2 illustrates the dramatic effect of random substitutions at various percentages would have on the identification of peptides from the original AMD dataset.

Table 2

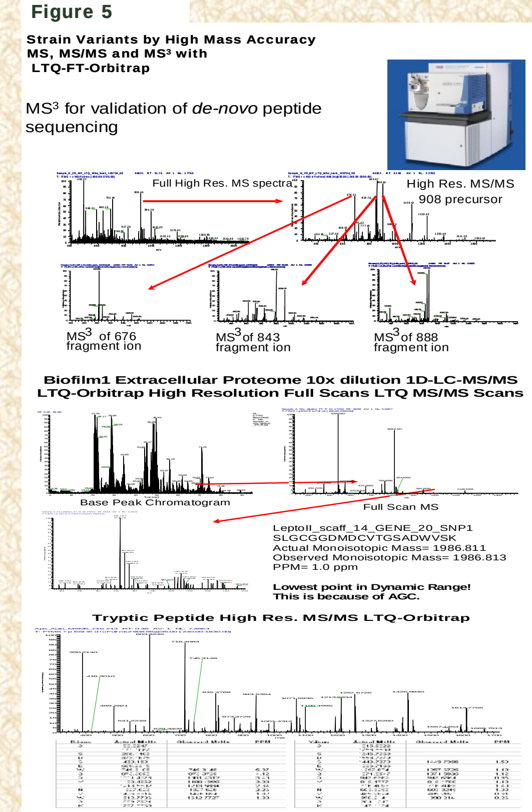
Random Substitutions

Mutated Database %Substitutions	Detected Peptides	Detected Proteins at Filter Level 1	2	3
0	23121	5090	2875	1934
0.1	22955	5087	2868	1927
0.5	21209	4926	2743	1819
1	19418	4708	2592	1719
2	16378	4368	2304	1552
5	9878	3388	1674	1087
10	4580	2195	894	492
20	1015	758	170	58
50	9	9	0	0

How to detect and characterize amino acid substitutions in a high throughput fashion

- Table 2 clearly shows the drastic effects on protein identification a large number of amino acid substitutions can have.
- Methodology needs to be developed for rapid detection and characterization of regions of proteins modified by amino acid substitutions.
- *De-novo* or sequencing tagging algorithms hold promise but are slow and inaccurate.
- MS methodologies must also be advanced to provide more readily interpretable datasets for such programs.
- Figure 5 is our experimental platform for high-throughput amino acid substitution identification with the use of the new hybrid linear ion trap FT-Orbitrap.

Figure 5



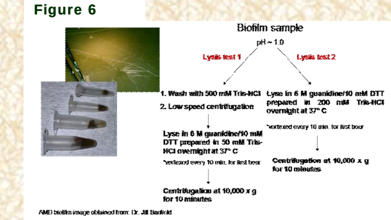
Characterization of Small Starting Quantities

- For many microbial community samples raw cellular starting material will be limited. For the AMD biofilm spatial profiling of biofilms are planned also requiring the processing and analyses of small starting materials (Figure 3).
- Recently we have investigated alternative techniques of cell lyses, protein digestion, sample clean up and LC/LC-MS/MS applying single tube lyses and digestions.
- These tests have been done on isolates of R. *palustris*, a simple microbe mixture and the AMD biofilm. Typical starting materials for isolates range from 5000mg-2g of wet cell weight, for these studies we are aiming for 1mg and less of wet cell weight.
- Table 3 illustrates the results from the *R. palustris* isolate with over 2,000 proteins identified from only 1mg of cells on the LTQ. The AMD biofilm provides a much greater challenge with pH issues, debris, cell lyses challenges and presences of unknown small molecules which lead to column clogging. Figure 6 highlights some preliminary results.

Table 3

Amount of starting material	6 M guanidine/10 mM DTT		8 M urea/10 mM DTT	
	Peptide IDs	Protein IDs	Peptide IDs	Protein IDs
1 mg run 1	3,584	654	2153	473
1 mg run 2	3,989	666	-	-
5 mg run 1	1,880	426	1,584	399
5 mg run 2	2,859	597	–	–
25 mg run 1	5,136	793	3,461	668
1 mg LTQ	16,262	2,146	–	-

Figure 6



CONCLUSION

- We have established that current MS based proteome technologies can characterize simple microbial communities, even with reference genomes.
- There are clearly many challenges that need to be overcome.
- We are currently developing new methodologies in mass spectrometry and proteome informatics to tackle these challenges and further advance proteogenomics approaches for analyzing natural microbial communities.

ACKNOWLEDGMENTS

- This research was supported by grants from the DOE Microbial Genome Program (JFB), NSF Biocomplexity Program (JFB), DOE Energy Biosciences Program (RCB), and DOE Genomes to Life Program (RLH).
- We thank Dr. David Tabb (ORNL) 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. Frank Larimer at ORNL for providing computational resources.