

# Proteomics Of The Microbial Community In An Acid Mine Drainage Biofilm

N. C. VerBerkmoes<sup>1</sup>, R. J. Ram<sup>2</sup>, M. P. Thelen<sup>2</sup>, Manesh Shah<sup>1</sup>, G. W. Tyson<sup>2</sup>, B. J. Baker<sup>2</sup>, J. F. Banfield<sup>2</sup>, R. L. Hettich<sup>1</sup>

1) Oak Ridge National Laboratory, Oak Ridge, TN; 2) University of California, Berkeley, Berkeley, CA

## OVERVIEW

- Here we report the first combined proteogenomic characterization of a natural microbial community.
- Mass spectrometry-based "shotgun" proteomics was combined with community genomics to identify 2,036 proteins from five dominant species in an acid mine drainage (AMD) biofilm.
- We validated 216 conserved and 407 unique hypothetical proteins. Entire operons encoding expressed, novel, lineage-specific proteins, which may be important for acid and metal tolerance, were detected.
- An extracellular fraction was dominated by a novel protein theorized to be a cytochrome central to iron oxidation and AMD formation. Sequencing of DNA encoding cytochrome regions for which peptides were not recovered resulted in 100% sequence coverage of the mature protein.
- This study demonstrates that community genomics and proteomics can be combined to enable comprehensive *in situ* analyses of natural consortia.

## INTRODUCTION

- Microbial communities play key roles in the Earth's biogeochemical cycles. Our knowledge of the structure and function of these communities is limited because analyses of microbial physiology and genetics have been largely confined to studies of organisms from the few lineages for which cultivation conditions have been determined. Recent acquisition of genomic data directly from natural samples has begun to reveal the genetic potential of communities (G. W. Tyson et al., Nature 428, 37, 2004) and environments (J. C. Venter et al., Science 304, 66, 2004). Here we combined cultivation-independent genomic and proteomic analyses to validate predicted genes, determine relative abundance and cellular localization of expressed proteins, and provide clues to protein function.

### Challenges for Proteome Analysis of Microbial Communities

Proteome analysis of any microbial community will be difficult with any current technology. The primary theoretical and practical concerns are:

- The level of DNA sequence information and quality annotation available on the community.
- The level of diversity and dynamic range associated with the species of interest in the community.
- The quantity of biomass available for study.
- The level of interrelatedness and/or diversity at the base pair level amongst members of the same species in the community and between species in the community.

## EXPERIMENTAL

### Biofilm Sample Collection and Protein Extraction/Fractionation

- The biofilm samples used in this study and prior community genome sequencing (Tyson, 2004) were collected from the underground regions of the Richmond Mine at Iron Mountain near Redding, California (USA) (Figure #1). These pink biofilms grew on the surface of acidic (pH ~0.8) sulfuric acid-rich, hot (~42°C), metal-contaminated solutions.
- The biofilm formed a continuous, paper-thin film on the surface of a pool of slowly flowing acid mine drainage (AMD) (Figure #2, A and B). This biofilm was much thinner than the biofilm present at the same location six months later (Figure #2, C), indicating that it comprised an actively growing community.
- Oligonucleotide probe-based studies (FISH) demonstrated that *Leptospirillum* group II dominated the sample, but it also contained *Leptospirillum* group III, *Sulfobacillus*, and archaea related to *Ferroplasma acidimanus* (Figure #3).
- For proteomic analyses, ~8 ml of biofilm were fractionated by washing, sonication, and centrifugation to yield extracellular proteins and samples enriched in proteins from whole cells, membranes (two different preparations), and cytoplasm. Extracellular proteins were collected after treatment of the biofilm by cold osmotic shock, which releases mostly periplasmic proteins. The five fractions were stored at -80°C until used in mass spectrometry experiments. Proteome fractions were denatured and reduced, digested with trypsin, desalted, and concentrated.

### LC/LC-MS/MS Analysis and Database Searching

- We combined two proteomic datasets (LCQ and LTQ) that were generated by triplicate analyses of the samples listed above. The LCQ dataset was generated on a 3-dimensional quadrupole ion trap mass spectrometer (LCQ-DECA XP plus, Thermo Finnigan, San Jose, CA). The LTQ dataset was generated on a 2-dimensional linear ion trap mass spectrometer (LTQ Thermo Finnigan). Both analyses used an identical "shotgun" proteomics approach via a two-dimensional (2D) nano-LC MS/MS system with a split-phase column (RP-SCX-RP).
- From the genomic dataset, we created a database of 12,148 proteins (Biofilm\_DB1) that was used to identify MS/MS spectra. All MS/MS spectra from the LCQ and LTQ datasets were searched with the SEQUEST algorithm (Thermo Finnigan), and filtered with DTASelect at the peptide level [Xcorr of at least 1.8 (+1), 2.5 (+2) 3.5 (+3) were used in all cases]. Results of all replicate runs were compared with the Contrast program and evaluated based on matching of one peptide, two or more peptides, or three or more peptides per protein. Only proteins identified with two fully tryptic peptides were considered for further biological considerations.
- Alternative database searches were conducted to estimate false positive rates and matches against very large databases. For all of these searches, a subset dataset was generated from both the LCQ and LTQ datasets. The LCQ and LTQ sub-datasets contained a single 24-hour 2D analysis of each of the five fractions from the proteome. The subset datasets were searched with SEQUEST. The subset datasets were searched against a variety of databases, including the genomic database (Biofilm\_db1), Biofilm\_db1 with an appended reversed database, termed Biofilm\_db1R, Biofilm\_db3 was created by the addition of *S. onaidensis* MR-1, *R. palustris* CGA009, *E. coli* K-12, and *S. cerevisiae* public protein databases to Biofilm\_db1. Biofilm\_db1\_Large contained over 200 microbes/plant plus biofilm\_db1 (978,849 entries).
- All DTASelect files and Contrast files used in this study, databases and resulting MS/MS identifications, can be downloaded from the AMD Proteome Website Analysis Page ([http://compbio.ornl.gov/biofilm\\_and/](http://compbio.ornl.gov/biofilm_and/)). Linkable spectra for identified peptides are downloadable, a step towards open access proteome results (Ram, Science, 2005).

Figure 1

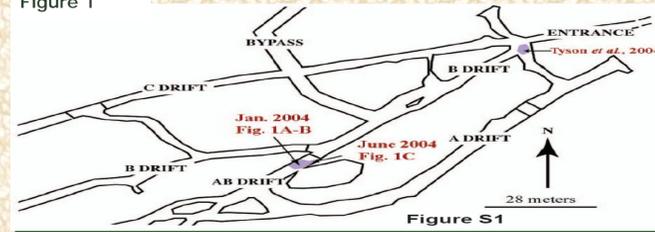
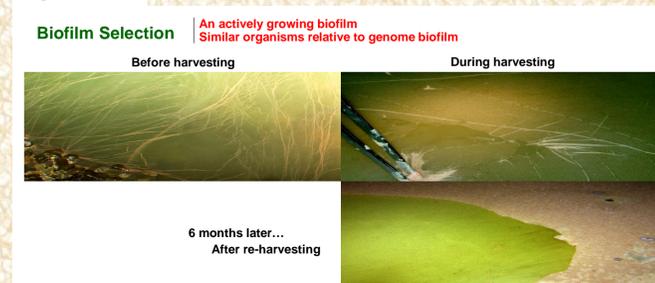


Figure 2



## RESULTS

### Abundant biofilm proteins

- 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. Percent sequence coverage, the number of unique peptide hits, and MS/MS spectral counts are all indicators for protein abundance.
- Of the proteins enriched in the extracellular fraction (top ten based on spectral count from one LTQ dataset shown in Table 2), the one with the highest sequence coverage and spectral count was encoded by a hypothetical gene from *Leptospirillum* group II.
- 67% of the hypothetical protein sequence from the community genomics dataset could be reconstructed from multiple overlapping peptides. No peptides were recovered from three discrete regions of the protein (Figure 5). The first is a 23 amino acid region predicted to be a signal peptide. Sequences for the gene determined by PCR amplification differed from that in the community genome dataset in substitutions of glutamate for aspartate at position 76 and serine for glycine at position 139.
- Peptide sequences were re-analyzed with a database containing the corrected amino acids. The protein was fully recovered with 100% sequence coverage after this modification, taking into account the predicted signal peptide (Figure 5).
- The abundant protein in the extracellular fraction is weakly similar (e-6 by BLASTp) to previously studied c-type cytochromes and Fe/Pb permeases. The presence of a heme-binding consensus sequence suggested a role in electron transport. Based on its distribution, abundance, and the ability of its *L. ferriphilum* homolog to oxidize iron, we hypothesized that the identified hypothetical (now called cyt579) was central to iron oxidation by *Leptospirillum* group II (Figure 6).

Figure 3

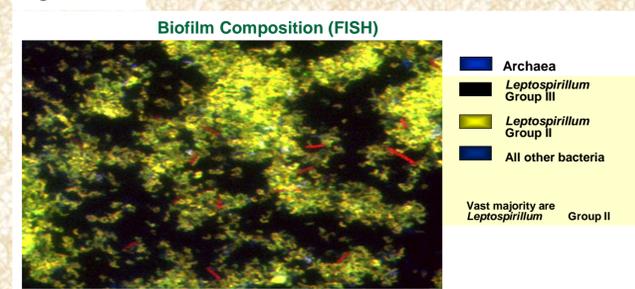


Table 1

Filtering Level	LCQ Dataset	LTQ Dataset	Combined Dataset
Liberal filters*	3127	5,534	5,994
Conservative filters**	1160	2,077	2,146
Ultra-conservative filters***	837	1,419	1,435

\*Liberal filters requiring at least 1 peptide per gene;  
 \*\*Conservative filters requiring at least 2 peptides per gene;  
 \*\*\*Ultra Conservative filters requiring at least 3 peptides per gene.  
 Xcorr of at least 1.8 (+1), 2.5 (+2) 3.5 (+3) were used in all cases.

Figure 4

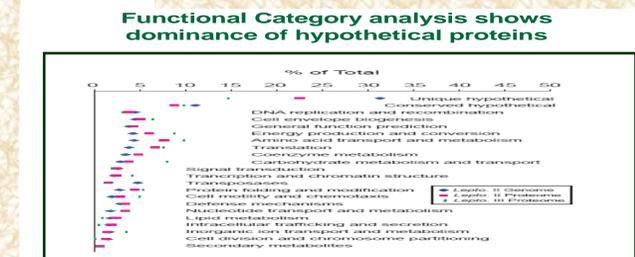


Figure 5

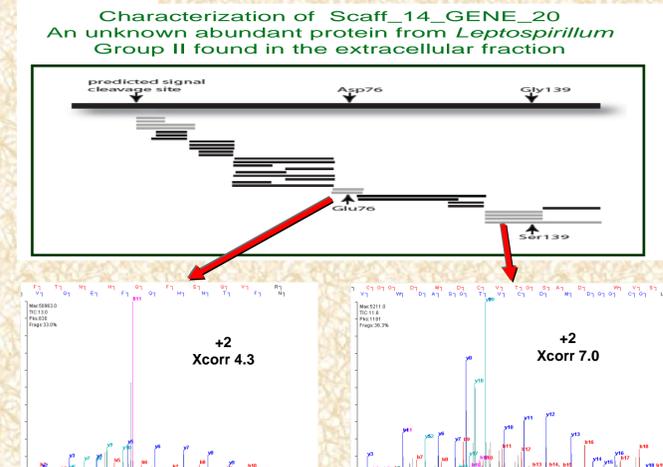
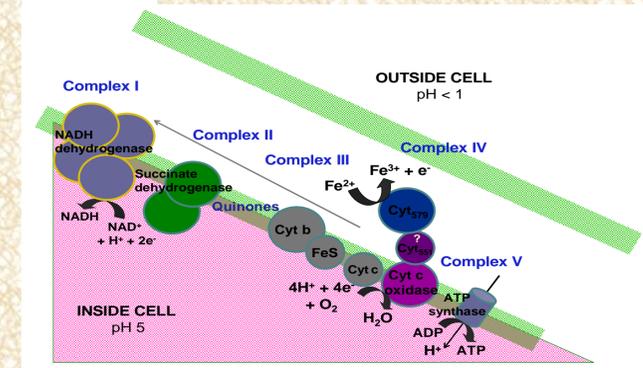


Figure 6



### Database tests

- We performed database analyses to test the likelihood of matching unique peptides from proteins not present in the biofilm samples ("false positives").
- Three appended databases were used: a reversed database, a database with 4 other microbes, and a database with 260 microbial/plant species.
- We estimated false positive rates by comparing the number of unique peptides identified from the AMD database (Biofilm\_db1) to unique peptides identified from proteins not from the biofilm genome database.
- This slightly overestimates the false positive rate, since peptides which were originally unique to the AMD database can become non-unique as other peptides in the database have the same sequence. A subset of these 'false positives' may correspond to homologs for which we have no genome sequence, yet are present in the biofilm.
- The results from each search are illustrated in Table 3.
- The results indicate that the rate of spuriously matching peptides is acceptable for the dynamics of an environmental sample, taking into consideration that some of the false positives could have derived from mine organisms for which we have incomplete or no genomic information.
- The results indicate that with ion trap data it is necessary to filter at least the two peptide level.
- The resultant increase in MS/MS spectra with LTQ data will lead to an increase in false positive levels with the same filtering criteria, emphasizing the need for two peptide filter level.

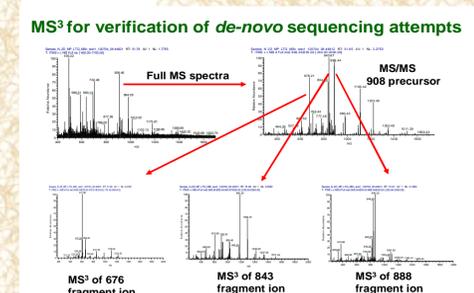
Table 3

	True Positives Peptides	False Positives Peptides	% False Positive Peptides
DB1 LCQ_1pep	6055	N/A	N/A
DB1 LCQ_2pep	5380	N/A	N/A
DB1 LTQ_1pep	17648	N/A	N/A
DB1 LTQ_2pep	16264	N/A	N/A
DB1FR LCQ_1pep	5879	411	6.5
DB1FR LCQ_2pep	5296	55	1.0
DB1FR LTQ_1pep	16940	1719	9.2
DB1FR LTQ_2pep	15791	507	3.1
DB3 LCQ_1pep	5778	827	12.5
DB3 LCQ_2pep	5247	150	2.8
DB3 LTQ_1pep	16476	3197	16.3
DB3 LTQ_2pep	15440	1035	6.3
DBLarge LCQ_1pep	5019	3254	39.3
DBLarge LCQ_2pep	4727	213	4.3

## CONCLUSION

- We confidently identified 2,036 unique proteins from the Acid Mine Drainage biofilm community. This is the first example of a large-scale proteome analysis of a natural microbial community combining genomics and proteomics.
- The large number of detected unique and conserved novel proteins underscores the importance of proteins of unknown function in the community. Novel proteins that were detected and abundant may be targeted for purification from biofilm samples and subsequent functional analysis.
- A hypothetical protein, now named Cyt<sub>579</sub>, was detected with very high spectral counts and sequence coverage in the extracellular proteome and is hypothesized to be central to iron oxidation processes.
- A strain variant of the abundant hypothetical protein was detected by combined genomic/proteomic approach.
- Future plans will include deeper coverage through 3D-LC and increase dynamic range with FT-ICR. Sequence tagging or *de-novo* sequencing approaches with combined multiple MS<sup>3</sup> spectra (Figure 7). Absolute and relative quantitation for exploring spatial and temporal changes in the community.

Figure 7



## 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 informatics tools used in analysis of the MS data, and Dr. Frank Larimer at ORNL for providing computer resources.