PIGT: Proteomics Inferred Genotyping in Natural Environments

Nathan VerBerkmoes[1], Vincent Denef[3], Mark Lefsrud[1], Paul Abraham[1], Manesh Shah[1], Brett Baker[3], Ian Lo[3], Michael Thelen[2], Jillian Banfield[3], Robert Hettich[1]

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

OVERVIEW

- Recent acquisition of genomic data directly from natural samples has begun to reveal the genetic potential of communities (Tyson, *Nature* 2004) and environments (Venter, *Science* 2004).
- The combination of mass spectrometry-based proteomics with community genomic analyses has shown promise for characterizing microbial communities proteomes (Ram, *Science* 2005).
- The Acid Mine Drainage System at Richmond Mine (Iron Mountain, Redding, California) is inhospitable place to live (pH less than 1.0, Molar Metals Conc., 42° C, with no sunlight, fixed carbon or fixed nitrogen). Never the less, low diversity microbial biofilms of bacteria of archaea, as well as fungi, protists and viruses, thrive in these conditions.
- The bacterium *Leptospirillum* Group II is a dominate species in many AMD biofilms. Two environmentally derived complete genomes have been sequenced from two locations within the mine (5-way and UBA sites). The genomes differ by 0.3% at the 16S RNA level, are highly syntenous and share 80% of their genes with 95.24% AAI.
- We recently have demonstrated the use of shotgun proteomics to map, genome wide, strain variation, strain-specific expressed protein variants in a third community, for which no genome sequence is available (Lo, et al. *Nature* 2007) Figure 1.
- Here we illustrate this technique across multiple samples, with smaller sample sizes and a more rapid approach (1-2 weeks vs. 24 hours).

EXPERIMENTS

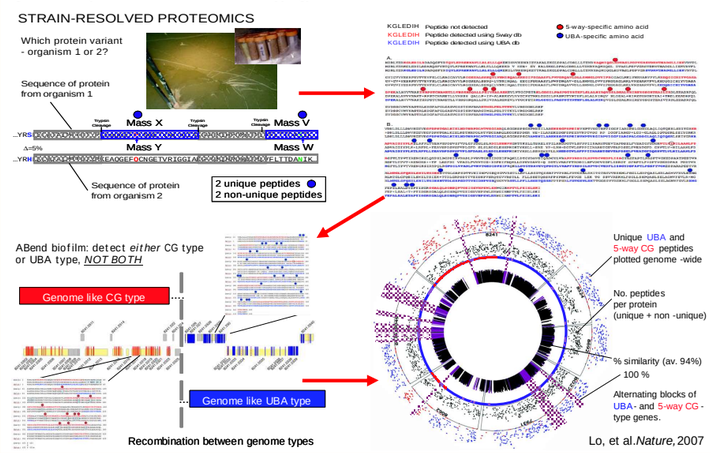
Samples and Sample Preparation

- The Acid Mine Drainage (AMD) biofilm samples used in this study were collected from the underground regions of the Richmond Mine at Iron Mountain near Redding, California (USA). These biofilms grew on the surface of acidic (pH ~0.8), hot (~42°C), metal-contaminated solutions.
- For deep proteome analyses samples were fractionated via centrifugation techniques and analyzed in technical triplicates (~15 days per proteome with grams of raw biofilm starting material).
- For rapid-PIGT analyses only a whole cell fraction was prepared, by small sample lyses and preparation, from many biofilms spatially and temporally collected from across the mine. All samples were analyzed in technical triplicates (3 days per proteome with 10-50 mg of raw biofilm starting material).

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 or LTQ-Orbitrap (Thermo Finnigan) with 22 hour runs per sample. All rapid-PIGT analyses were done on the LTQ-Orbitrap.
- 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. LTQ analyses were run in the exact same method except full scans were in LTQ.
- 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)]. Only proteins identified with *two fully tryptic peptides* were considered for further biological study.
- Monoisotopic theoretical masses for all peptides identified by SEQUEST were generated and compared to observed masses. Observed masses were extracted from .raw files, from full scan preceding best identified spectra.
- Database included the predicted AMD proteome from the original genome sequencing project (5-way CG site Tyson, Nature 2004 and UBA site, Lo et al. Nature, 2007). This database structure was 2846 5way_CG Leptospirillum Group II, 2603 UBA Leptospirillum Group II, 3183 UBA Leptospirillum Group III, 2140 Ferroplasma type I, 2049 Ferroplasma type I, 1445 G plasma type I, 916 Unassigned archaeal, 592 Unassigned bacterial, 36 common contaminants proteins.
- False positive levels were determined by conventional reverse database searching methods and compared with high mass accuracy results from the Orbitrap.

Figure 1 PIGT (Proteomics Inferred Genome Typing)



RESULTS AND DISCUSSIONS

Deep Proteome Analyses

- Deep Proteome analyses (at least 5 five fractions in technical triplicate) was performed on 7 natural AMD biofilm samples and one bioreactor sample (bioreactor mimics living conditions in the mine).
- These samples were spatially and temporally separated (Figure 2).
- Totals for proteins, peptides and spectra identified are shown in Table 1 and the false positive rate derived by reverse database searching is shown in Table 2 (for Proteome 1 AB End)
- Some Simple Metrics

  For the most abundant organism:
  - ~ 52% predicted proteins detected in a single sample
  - ~ 49% hypothetical genes (35% of the predicted genes)
  - ~ 52% conserved hypothetical genes

  Overall, proteins identified across multiple samples:
  - ~ 65% Leptospirillum group II
  - ~ 41% Leptospirillum group III
  - ~ 37% G plasma
  - ~ 20% Ferroplasma type I
  - ~ 20% Ferroplasma type II

Table 1

Sample	Proteins IDs	Unique Peptides	Non-Unique Peptides	Hypothetical Proteins
AB End	3,234	8,737	30,560	1,108
AB Front Sample 1	2,369	11,019	9,994	930
AB Front Sample 2	3,989	14,901	25,063	1,309
UBA (Ultraback)	3,569	12,643	27,115	1,164
C-Pink	1,862	4,583	7,961	854
AB Muck GSII	3,522	11,805	24,371	1,110
AB Muck Friable	3,572	11,077	28,869	1,152
Bioreactor	2,675	5,478	16,261	752

Figure 2

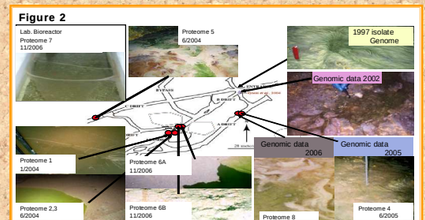


Table 2

Reverse Database Searching for False Positive Levels AB End Dataset

1 Peptide Filter

Fraction	True Peptides	False Peptides	Percentage	True Proteins	False Proteins	Percentage
Membrane A	16372	581	3.42	3136	566	15.29
Membrane N	24387	661	2.64	3806	630	14.21
Extracellular	6628	406	5.77	1393	393	22.01
Whole Cellular	18524	654	3.41	3229	626	16.23
Intracellular	13413	454	3.27	2821	433	14.18

2 Peptide Filter

Fraction	True Peptides	False Peptides	Percentage	True Proteins	False Proteins	Percentage
Membrane A	15201	30	0.19	1980	15	0.75
Membrane N	23159	64	0.28	2578	33	1.26
Extracellular	5973	29	0.48	738	16	2.12
Whole Cellular	17417	58	0.33	2135	30	1.38
Intracellular	12399	47	0.37	1611	26	1.59

For *Leptosprillium* Group II Proteins
7,526 Peptides were CG type
23,787 Peptides were UBA type
74,483 matched both types

PIGT in Deep Proteome Analyses

- Previous data obtained by MLST (Multiple Locus Sequence Tagging) suggested that large scale recombination might be occurring between two ancestral genomes (Lo, et al. *Nature* 2007).
- Further technological advancement led to the development of strain resolved FISH (Fluorescent in-situ hybridization), which allowed direct rapid analyses of *Leptospirillum* II strain types (Figure 4). This method a major advancement over typical FISH in that strains of *Leptospirillum* can be distinguished (Figure 3).
- Neither MLST or strain resolved FISH is high resolution, thus neither can exactly determine stain types, level of recombination or sites of recombination.
- Figure 2 illustrates the methodology of Proteome Inferred Genome Typing (PIGT).
- Figure 5 shows the strain resolution at the protein level via PIGT analyses after deep proteome analyses for the same biofilms as in Figure 3 and 4.
- Recombination was observed in large blocks of 10-100kB and sometimes the exact spot of recombination crossed a single protein and could be clearly detected.
- This methodology clearly provides high resolution strain typing across the entire genome if a deep enough proteome measurement can be made.

Figure 3

Conventional FISH
Example:

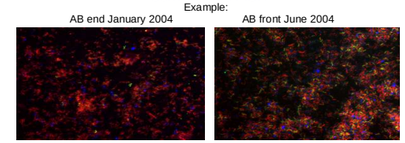


Figure 4

Stain-resolved ecology: Lepto II type-specific FISH
Example:

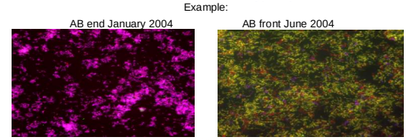
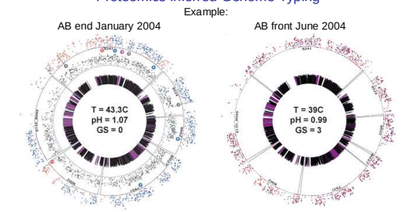


Figure 5

Proteomics Inferred Genome Typing
Example:



Rapid-PIGT Analyses

- The previous PIGT analyses were all done from large proteome datasets acquired on LTQs with large sample amounts and fractionation schemes (~15 days of data per proteome).
- Our goal was to test Rapid PIGT Analyses on an LTQ-Orbitrap (24 hours per sample). For this set of analyses AMD biofilms were processed by single tube lyses with smaller quantities, only a whole cell fraction was prepared. Samples were run in technical duplicates or triplicates on an LTQ-Orbitrap. Full scans were collected at 30K resolution and MS/MS scans were acquired in the LTQ.
- Table 3 highlights the number of proteins, peptides and spectra from each run, as well as sample location and growth stage.
- Table 4 highlights the number of peptides found in three PPM categories from the Orbitrap full scans from one of the Rapid-PIGT runs. The three categories include between 10 and -10 ppm, over 10 and -10ppm and undefined parent mass (Orbitrap full scan was not resolved but good MS/MS spectra were obtained that was matched by SEQUEST).
- The category of below 10 to -10 ppm has the highest level of accuracy, with false positives rates, determined by reverse database searching, of less than 0.1% (See Lefsrud Poster WP425 and Abraham Poster TP529).
- Figure 6 a, b and c shows the strain resolution at the protein level via Rapid-PIGT analyses for C75 sampling location spatially separated.
- While the depth of coverage is not as dense as with deep proteome analyses, sufficient unique peptides are identified to resolve areas of recombination and the strain variation within samples.

Table 3

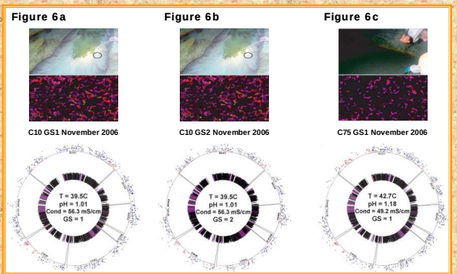
Sample Locations	Rapid-PIGT Run #	Proteins	Unique Peptides	Unique Spectra	% of Peptides within 10 to -10 PPM
C10m GS1 Nov2006	Sample4/Run2	1784	9299	27079	N/A
	Sample4/Run3	1684	8659	39160	87.20
	Sample4/Run4	1674	8743	35946	86.26
C10m GS2 Nov2006	Sample7/Run2	1627	7107	25533	N/A
	Sample7/Run3	2116	10258	29600	85.02
	Sample7/Run4	1914	8929	26987	84.58
C75m GS1 Nov2006	Sample8/Run2	2363	12950	28384	N/A
	Sample8/Run3	2181	10824	21477	82.54
	Sample8/Run4	1967	9545	18023	82.42
C75m GS1 Aug2006	Sample9/Run1	1650	9000	23666	85.85
	Sample9/Run2	1524	8171	22804	85.30
C75m GS1 Jun2006	Sample10/Run1	1462	7250	24870	83.03
	Sample10/Run2	1618	8481	25651	82.48
C75m GS1 Jun2006	Sample10bis/Run1	1810	9194	18071	N/A
	Sample10bis/Run2	1701	8466	18941	N/A
	Sample10bis/Run3	1759	8557	15393	N/A
AB drift GS0 Jun2006	Sample11/Run1	1882	11196	39992	87.62
	Sample11/Run2	1561	8444	36650	86.01
AB drift GS1 Jun2006	Sample14/Run1	1595	8377	36124	85.77
	Sample14/Run2	1399	7173	34887	86.66
	Average	**1763.55**	**9031.15**	**27461.90**	**85.05**
	Std Deviation	**247.10**	**1415.06**	**7495.25**	**1.79**

Table 4

C10m GS1 Nov2006 Sample 4 Run 4

PPM	Total Peptides	PPM	%
# of < 10 through -10 ppm	11472	% of < 10 through -10 ppm	86.26
# of > 10 through -10 ppm	615	% of > 10 through -10 ppm	4.62
# of unresolved	1213	% of unresolved	9.12

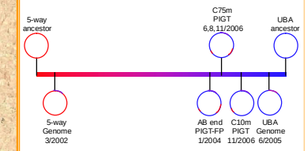
Figure 6a **Figure 6b** **Figure 6c**



CONCLUSION

- We were able to uncover the pattern of recombination around the genome of an uncultivated, uncharacterized microorganism (shadow of recombination).
- With the PIGT method developed it is possible to partially or completely deduce the sequences of gene variants, genome wide, in samples lacking prior genomic characterization, so long as genomic data from relatively closely related organisms are available.
- We have shown that the method can be reduced from deep proteome analyses (15 days) to rapid analyses (1-3 days) without massive loss of signal.
- High dynamic range and high resolution MS instruments will make this methodology more applicable to more complex systems, where false positive rates will be a major concern.
- Figure 7 illustrates our current hypothesis on recombination in the AMD system moving from two ancestral genome types to recombinants that expand clonally.

Figure 7



Recombination as a strategy for adaptation to environmental gradients

ACKNOWLEDGMENTS

- This research was supported by grants from the DOE Microbial Genome Program, NSF Biocomplexity Program, DOE Energy Biosciences Program, and DOE Genomes to Life Program.
- 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.
- Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.