LBNL-62935
Preprint

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Title Stain-resolved community proteomics reveals recombining genomes of acidophilic bacteria

Author(s), Ian Lo, Vincent J. Denef, et al

Division Genomics

Journal Name Nature
Volume 446
P. 537-541
March 29, 2007

Strain-resolved community proteomics reveals that recombination shapes the genomes of acidophilic bacteria

Ian Lo[*], Vincent Denef[*], Nathan C. VerBerkmoes[†], Manesh Shah[†],

Daniela Goltsman[*], Genevieve DiBartolo[*], Gene W. Tyson[*], Eric E. Allen[*], Rachna J. Ram[*], J. Chris Detter[‡], Paul Richardson[‡],

Michael P. Thelen[§], Robert L. Hettich[†], and Jillian F. Banfield[*#]

[*]University of California, Berkeley, California 94720, USA

[†]Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

[‡]Joint Genome Institute, Walnut Creek, California 94598, USA

[§]Lawrence Livermore National Laboratory, Livermore, California 94550, USA

[#]Corresponding Author: jill@seismo.berkeley.edu

Submitted to **Nature** for consideration for publication as an **Article**

August 14, 2006

Genetic exchange shapes the distribution of genomic diversity in microbial populations. Recombination is typically documented by sequence comparisons involving a few genes from organisms obtained in pure culture. The objective of this study was to distinguish, genome-wide, strain-specific expressed protein variants so as to resolve genetic evidence for recombination in the dominant member of a natural multi-species biofilm. Mass spectrometry–based proteomic discrimination relied upon cultivation-independent reconstruction of near-complete genomic datasets for two related *Leptospirillum* group II bacteria from the same system. Recombination involves chromosomal regions tens to hundreds of kilobases in length. The genome structure was confirmed by multi-locus sequence typing of isolates and uncultivated natural consortia. The ability to distinguish between proteins that differ by as few as a single amino acid enables strain-specific proteomics studies aimed at resolving the behavior of closely related members of natural communities.

Microorganisms comprise the majority of extant life forms, yet much remains to be learned about the nature and driving forces of microbial diversification. Most research to date has focused on physiological and genomic characterization of a relatively small number of isolated microbial species or strains maintained in monoculture. While providing crucial insights needed to connect genes and function, these studies are unable to capture some aspects of the organism's behavior in its natural environment. Perhaps the greatest knowledge gap is in understanding of how microorganisms function within natural multi-species consortia. The two steps needed to address this challenge are the

assessment of the metabolic potential of the community members and the distinction of their roles as a function of community structure and environmental conditions. Cultivation-independent genomic methods opened a new route to address the first challenge [1-3]. If a relatively complete gene inventory can be reconstructed, the second step, which aims to link biochemical processes to the microorganisms responsible, can be accomplished either by detection of mRNA transcripts or protein products. For example, genomic data from one biofilm [1] was used to identify over 2,000 proteins derived from the five abundant organisms in a similar yet spatially and temporally distinct and genomically uncharacterized biofilm community [4].

Recently, it has become clear that microorganisms that are often grouped together as a single species may play quite distinct roles in natural systems [5-7]. In fact, interactions amongst closely related organisms could be key to ecosystem optimization. Consequently, understanding of how natural microbial consortia function requires methods that can go beyond recognition of protein types to distinguish specific proteins from closely related species. In the current study we show that the exquisite selectivity of mass spectrometry-based proteomics enables discrimination between expressed protein products that differ by as little as a single amino acid substitution. Importantly, this was possible in the absence of any *a priori* knowledge of the detailed genomic composition of the actual sample analyzed. The approach makes high-resolution analysis of the contributions of closely related organisms to community function possible. Mapping of the genomic distribution of protein types uncovered evidence for recombination between closely related bacterial populations. Recombination may play a role in adaptation of these organisms, which are often the dominant microbial species associated with

formation of acid mine drainage (AMD), a major source of environmental contamination[8].

We sampled pink biofilms growing in the Richmond Mine, CA that are dominated by bacteria belonging to *Leptospirillum* group II [1, 9, 10] and contain lower abundances of *Leptospirillum* group III [11] and several archaea. *Leptospirillum* species are chemoautotrophs that obtain their metabolic energy by aerobic oxidation of ferrous iron. The ferric iron metabolic byproduct subsequently reacts with the surfaces of metal sulfide minerals present in ore deposits and mining waste, promoting sulfur oxidation and generating very acidic, metal-contaminated solutions (AMD).

Genome reconstruction and comparative genomics

DNA was extracted from a biofilm (UBA; Fig. S1) growing in a shallow (~ 1 cm) but flowing AMD stream (pH 1.1, 41 °C) in the A drift region of the Richmond Mine, CA (sample collected in June, 2005; Fig. S1). Approximately 117 Mb of genomic sequence obtained from a small insert shotgun library (Supplementary Information) was assembled to generate genome fragments that were assigned to *Leptospirillum* group II, *Leptospirillum* group III, and archaeal organisms based on GC content, sequence depth, the presence of phylogenetically informative genes, and by assembly. All *Leptospirillum* group II sequence fragments of > ~2 Kbp were assembled into seven scaffolds (7931, 8027, 8135_8049, 8062, 8241, 8524, 8692), yielding a near complete composite 2.64 Mb genome with 2,601 genes. The 16S rRNA gene of the UBA-type *Leptospirillum* group II organism shares 99.7% sequence identity with the *Leptospirillum* group II bacterium (5-

way CG-type) sampled from the 5-way region of the Richmond mine in 2002 and previously characterized via cultivation-independent genomic analysis by Tyson *et al.* [1].

The independently assembled UBA genome was aligned to a refined version of the 5-way CG-type *Leptospirillum* group II genome (see Supplementary Information) so that differences in gene content and genome arrangement could be identified (Table S1). The 2,160 genes in the UBA composite genome that have orthologs in 5-way CG dataset almost all have the same genomic context (i.e., the genomes are largely syntenous). Orthologs share an average of 95.24% amino acid level identity (median = 96.69 %). However, this value includes some genomic regions where blocks of 10s to 100s kb (comprising a total of 421 genes) have nucleotide sequence that is essentially identical to that of the 5-way CG composite genome. After exclusion of identical proteins, orthologs share an average identity of 94.11% (median = 95.53%). In some cases, the presence of integrase genes and high numbers of transposases and hypothetical genes suggest that sequence identity is due to insertion of the same plasmid- or phage-like DNA into both genomes (e.g., scaffold 8049, see Table S1). However, other blocks encode primarily core metabolic genes. For example, one region contains 54 genes (63% expressed) with ~ 100% nucleotide-level sequence identity, spans a tRNA (orange bar in Fig. 1), lacks integrases and transposases, and is unlikely to be phage derived (Fig. 1). Regions such as this may arise when recombination between two genome types is followed by a selection event that removes from the population (or reduces to low abundance) blocks of one of the sequence types [12].

Proteogenomics

The UBA and 5-way CG *Leptospirillum* group II genomic datasets were manually edited to avoid unsubstantiated differences in gene length and reading frame. The predicted proteins from both datasets were combined to generate a single database (http://compbio.ornl.gov/biofilm_amd_recombination) containing 16,170 protein entries, including predicted proteins from *Leptospirillum* group III and archaeal genomes known to be in the sample. The database was used for identification of tandem mass spectra collected from peptides generated by enzymatic digestion of proteins extracted from a third biofilm sample (Supplementary Information) obtained from the genomically uncharacterized ABend location within the Richmond Mine (Fig. S1). The proteomic dataset was previously analyzed by searching with only the 5-way CG genomic data [4]. Because MS tandem spectral identification depends on a high quality match between the mass of predicted and measured peptide fragments, most peptides predicted to differ in their amino acid composition and mass can be distinguished (however lysine and glutamine can only be differentiated on high resolution mass spectrometers, and isoleucine and leucine cannot be distinguished on most common mass spectrometers). Specifically, if the appropriate peptide sequence is available, the peptide can be identified and distinguished from other related peptides. For *Leptospirllium* group II proteins, we noted that 7,526 peptides could be identified using the 5-way CG but not the UBA genomic dataset, and 23,787 peptides identified using the UBA but not the 5-way CG genomic dataset. A total of 74,483 non-unique (matching both 5-way CG and UBA datasets) peptides were identified (all peptide counts are cumulative over replicate

fractions). This result suggested that proteins of both the UBA-type and the 5-way CG-type were present in the ABend sample, and that UBA-type proteins dominate (~69%).

If the protein coverage (fraction of the protein sampled by identified peptides) obtained using UBA protein sequences exceeds that achieved using the 5-way CG sequences, the protein is more similar to the UBA than the 5-way CG-type protein. When the ratio of protein coverage achieved using the UBA vs. 5-way CG sequences is plotted as a function of genomic location across region 7931 (Fig. 1, green triangles), it appears that the entire genomic block encodes proteins of the UBA type. This representation assumes that the *Leptospirillum group* II genome in the ABend sample is syntenous with the UBA genome, a result considered likely given the largely shared gene order between the UBA and 5-way CG genome types.

In contrast, Figure 2A illustrates another large genomic region (~698 Kbp in length) with coverage ratios > 1 (UBA/5-way CG) that alternate with regions with coverage ratios < 1 (this fragment contains the origin of replication, see Fig. S2). The genome assembly across the transitions between these regions has been carefully checked to eliminate the possibility that boundaries that arise due to incorrect linkage of DNA sequence fragments. To ensure that the sample does not contain discrete coexisting UBA- and 5-way CG-like populations, we also plotted the number of unique peptides identified for each protein across the same genomic region (Fig. 2B). The result clearly confirms alternation of blocks of UBA and 5-way CG protein types around the genome, with only one protein variant present for each gene. Furthermore, the average number of unique peptides per protein across UBA and 5-way CG blocks is approximately the same, as expected for proteins encoded by a single genome type (also see Supplementary

Information and Fig. S3). These plots shadow the recombined genome structure. We do not directly observe the DNA sequence data that would demonstrate that the 5-way CG and UBA genome types have undergone large-scale homologous recombination. Rather, we see evidence for recombination in the pattern of protein sequence types. Plots for the other large genome fragments (Fig. S3) were used to infer the recombination pattern genome-wide (Fig. 3).

We mapped peptides onto protein alignments (Fig. 4A,B) across the recombination point in the middle of Figure 2 to evaluate the sensitivity and accuracy of peptide-based protein discrimination. Notably, proteins that differ by only a single amino acid can be distinguished if peptides are recovered in the relevant regions. We manually verified 226 unique peptides (Table S2); 206 were found to be true positives. Most 5-way CG-type peptides in otherwise UBA proteins (and v.v.) were false positives found at the edge of the scoring threshold (see Supplementary Information).

Throughout the genome there were 129 ortholog pairs for which two or more unique peptides were detected for both UBA and 5-way CG variants. Approximately 54% of these were due to the inability to discriminate between Ile/Leu, the inability of the instrumentation used to distinguish masses different by less than 1 Dalton (Lys/Gln), and inconsistent discrimination of the similar pairs Asn/Ile Asp/Asn, and Glu/Lys. Therefore these should not be considered as unique peptides in this study. New hybrid mass spectrometers are capable of rapid scanning as well as high resolution and high mass accuracy MS and MS/MS acquisitions on liquid chromatography time scales [13]. A hybrid instrument, an LTQ-Orbitrap, can routinely provide mass accuracy of ~1-3 parts per million (ppm) on the parent peptides and ~ 7-10 ppm on fragment ions, with external

calibration (see Fig. S4 and Table S3 for results for one ABend biofilm fraction), allowing discrimination of all single amino acid polymorphisms, except for leucine/isoleucine. These instruments will greatly reduce the false positive identifications.

After accounting for mass discrimination limitations, the spectra that indicated ortholog pairs with peptides of both UBA and 5-way CG protein variants were manually checked to eliminate false positives. Only 24 cases (18.6%) were verified as examples of proteins with unique peptides for both protein variants. These proteins are distributed throughout the genome and generally the cumulative number of unique peptides detected of one variant is far greater than the other one. This result suggests that the second protein variant is present in a minor fraction of the population. At least one of the three cases where we detected a high number of unique peptides for both variants is clearly a recombinant protein (a RuBisCO-like protein; Fig S5).

Verification of recombination

As this is the first mapping of genome-wide recombination structure via proteomic analysis, we undertook targeted isolation (Supplementary Information) of *Leptospirillum* group II strains (see Table S4) so that we could to test for recombination between the UBA and 5-way CG genome types using a standard approach involving PCR-based multilocus sequence typing (MLST) analysis [14] (see Table S5). Isolation by colony formation required use of a plate overlay method [15]. Four isolates were obtained from biofilms collected from different sampling sites at least 30 meters apart (Fig. S1). MLST results revealed that three isolates share most of their gene sequences with the 5-way CG

population and one isolate shares most gene sequences with the UBA population (Table 1B). However, linkage patterns in the UBA-like isolate (strain CF-1) prove the occurrence of recombination, supporting the proteomics-based deduction of recombination between UBA and 5-way CG types. The same genes used for MLST analysis of isolates were amplified from the ABend biofilm and sequenced (Table 1 and Supplementary Information). Results confirmed the inferred recombinant genome structure deduced from proteomic analysis.

The UBA genomic dataset contains a few (<0.1 %) sequencing reads with ~ 92% identity to the UBA genome type. Almost all of these had very high (typically 100%) nucleotide sequence identity to the *Leptospirillum* group II detected in the 5-way CG genomic dataset. In essentially all cases analyzed, both sequenced ends (mate pairs) of these DNA fragments are 5-way CG type over their entire length. This is consistent either with a low abundance of the 5-way CG population type in the UBA community or recombinants with large recombination fragment size. However, in at least one case, a 5-way CG type gene fragment has recombined into about half of the UBA-type genomes sampled (Table S6). Given that 9 of the 15 nucleotide substitutions in this gene correspond to amino acid differences, we infer that the recombination event likely resulted in heterogeneity in the function of this permease within the population. In combination with the example of the RuBisCO-like protein, this observation indicates that modular creation of new protein types is an important source of new functionality in these acidophiles. This is similar to earlier observations in, for example, *E. coli* [16].

A final line of evidence for recombining *Leptospirillum* group II populations derives from a comprehensive analysis of the distribution of single nucleotide polymorphisms

(SNPs) in both the UBA and 5-way (5-way CG) datasets, an approach used for the analysis of recombination patterns in higher organisms [18]. The UBA population has ~1 SNP every 30,000 bases with a fairly random distribution. The SNP frequency in the 5-way CG population is about 10 times higher than in UBA, with only two variants at each variable locus and alternation of tens of Kbp blocks with few or no SNPs with blocks with high SNP concentrations (Fig. 2C). This alternation, as well as linkage of SNPs, suggests that the 5-way CG population is dominated by two very closely related strains that have recombined in large blocks, with heterogeneity preserved in regions not swept in selection events. In combination, the results suggest that adaptation relies, at least in part, on selection for mosaic genome types.

Previously, evidence for extensive recombination was documented in archaeal populations from the Richmond mine biofilms [1] and in other archaeal populations from hypersaline and geothermal environments [19, 20] as well as in bacteria, e.g., between different species of *Thermotoga* [21]. The current study reveals that closely related bacterial types (~95% average amino acid identity) in the Richmond Mine biofilms are also undergoing recombination. However, the recombination block size in *Leptospirillum* group II appears to be unusually large (10's to 100's Kbp) for bacteria, as compared to typical recombination fragment sizes between ~100 bp and 10 Kbp (summarized in [22]). Whether this is due to an unusual recombination makeup of *Leptospirillum* group II is not yet clear. The genomes have many genes required for DNA uptake (competence) and secretion, homologous pairing and recombination. Of principal interest are RecA (homologous pairing), RuvA and B (branch migration), and RuvC (resolvase), proteins that were all confidently detected with high MS coverage. Notably

not detected by MS is the RecB protein of RecBCD presynaptic nuclease, and the absence of antirecombinant genes SbcB, C and D. In comparison with many prokaryote genomes, the overall recombination gene composition is most similar to that of *Geobacter sulforeducens* [23] . Recombination events are apparently infrequent between *Leptospirillum* group II types, as might be expected based on a log-linear dependence of recombination on sequence identity [24]. Less frequent recombination in *Leptospirillum* group II compared to archaea probably accounts for the lower population-level diversity (i.e., near-clonality within populations) compared to the archaea [25].

Proteomic detection of amino acid substitutions has been demonstrated previously for hemoglobin variants related to human disease [26-28]. These prior studies have relied upon the wealth of previous information on hemoglobin variation in human populations, which provided a known starting point for the protein identification. The key methodological advance of our study is the finding that 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 were able to uncover the pattern of recombination around the genome of an uncultivated, uncharacterized microorganism. Proteomic analysis of environmentally-derived samples containing microorganisms for which multiple genomic datasets are now available (e.g., *Shewanella, Prochlorococcus, Burkholderia*) could be used to identify peptide sequences shared with sequenced organisms and to constrain the biochemical function at the time of sampling. A similar high-resolution proteogenomic approach may find application in a much wider variety of biological problems. For example, the method may be employed for medical diagnosis (e.g., strain typing of pathogens), where a

reasonable amount of biomass is available, but where the sample not amenable to the conventional identification methodologies. In addition, variant-specific proteomics could be used to discriminate between closely related proteins involved in human or animal disease (e.g. superoxide dismutase haplotypes linked to amyotrophic lateral sclerosis [29]), and might find important application in the realm of protein biomarkers. Most importantly, this work opens the way for detailed exploration of the microbial ecology of natural consortia.

1. Tyson, G. W. et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428, 37-43 (2004).
2. Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74 (2004).
3. DeLong, E. F. et al. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311, 496-503 (2006).
4. Ram, R. J. et al. Community proteomics of a natural microbial biofilm. Science 308, 1915-20 (2005).
5. Acinas, S. G. et al. Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430, 551-4 (2004).
6. Johnson, Z. I. et al. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311, 1737-40 (2006).
7. DeLong, E. F. Microbial community genomics in the ocean. Nat Rev Microbiol 3, 459-69 (2005).

8. Johnson, D. B. & Hallberg, K. B. The microbiology of acidic mine waters. Res Microbiol 154, 466-73 (2003).

9. Bond, P. L., Smriga, S. P. & Banfield, J. F. Phylogeny of Microorganisms Populating a Thick, Subaerial, Predominantly Lithotrophic Biofilm at an Extreme Acid Mine Drainage Site. Appl. Environ. Microbiol. 66, 3842-3849 (2000).

10. Coram, N. J. & Rawlings, D. E. Molecular Relationship between Two Groups of the Genus *Leptospirillum* and the Finding that *Leptospirillum ferriphilum* sp. nov. Dominates South African Commercial Biooxidation Tanks That Operate at 40 °C. Appl. Environ. Microbiol. 68, 838-845 (2002).

11. Tyson, G. W. et al. Genome-Directed Isolation of the Key Nitrogen Fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an Acidophilic Microbial Community. Appl. Environ. Microbiol. 71, 6319-6324 (2005).

12. Cohan, F. M. What are bacterial species? Annu Rev Microbiol 56, 457-87 (2002).

13. Hu, Q. et al. The Orbitrap: a new mass spectrometer. J Mass Spectrom 40, 430-43 (2005).

14. Maiden, M. C. et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95, 3140-5 (1998).

15. Johnson, D. B. Selective solid media for isolating and enumerating acidophilic bacteria. Journal of Microbiological Methods 23, 205-218 (1995).

16. Mau, B., Glasner, J. D., Darling, A. E. & Perna, N. T. Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. Genome Biol 7, R44 (2006).

17. Baldo, L., Bordenstein, S., Wernegreen, J. J. & Werren, J. H. Widespread Recombination Throughout Wolbachia Genomes. Mol Biol Evol 23, 437-449 (2006).

18. Nielsen, R. et al. Genomic scans for selective sweeps using SNP data. Genome Res 15, 1566-75 (2005).

19. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon Sulfolobus islandicus. Mol Biol Evol 22, 2354-2361 (2005).

20. Papke, R. T., Koenig, J. E., Rodriguez-Valera, F. & Doolittle, W. F. Frequent Recombination in a Saltern Population of *Halorubrum.* Science 306, 1928-1929 (2004).

21. Nesbo, C. L., Dlutek, M. & Doolittle, W. F. Recombination in *Thermotoga*: implications for species concepts and biogeography. Genetics 172, 759-69 (2006).

22. Falush, D. et al. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci U S A 98, 15056-61 (2001).

23. Rocha, E., Cornet, E. & Michel, B. PLoS Genetics 1, 247-59 (2005).

24. Majewski, J. & Cohan, F. M. DNA sequence similarity requirements for interspecific recombination in *Bacillus*. Genetics 153, 1525-33 (1999).

25. Majewski, J. & Cohan, F. M. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics 152, 1459-74 (1999).

26. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2, 930-42 (2001).

27. Manabe, T. Capillary electrophoresis of proteins for proteomic studies. Electrophoresis, 20, 3116-21 (1999).

28. Dalluge, J. J. Mass spectrometry: an emerging alternative to traditional methods for measurement of diagnostic proteins, peptides and amino acids. Curr Protein Pept Sci 3, 181-90 (2002).

29. Ruddy, D. M. et al. Two families with familial amyotrophic lateral sclerosis are linked to a novel locus on chromosome 16q. Am J Hum Genet 73, 390-6 (2003).

30. We thank Mr. T.W. Arman, President, Iron Mountain Mines, Mr. R. Carver, and Dr. R. Sugarek for site access and on-site assistance. We thank Dr. F. Larimer and M. Land of the ORNL Genome Analysis and System Modeling Group for computational resources for proteomic analysis. DNA sequencing was carried out at the DOE Joint Genome Institute. Funding was provided by a DOE Genomics:GTL Program under grant number DE-FG02-05ER64134 (Office of Science), the NSF Biocomplexity Program, and the NASA Astrobiology Institute.

31. All datasets, databases and supplemental files can be found at http://compbio.ornl.gov/biofilm_amd_recombination.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396.

Table 1: A. Multilocus sequence data obtained for 9 genes from the environmental ABend biofilm sample. A single sequence type was recovered, consistent with a single dominant genome type. Red boxes indicate genes that are identical to the 5-way CG type, blue are identical to the UBA type, purple are identical to both 5-way CG and UBA types, grey are cases where sequences were not recovered.

B. Results of MLST analysis comparing gene sequences in the two genomic datasets to those in the isolates of new *Leptospirillum* group II from the Richmond Mine, CA. Gene 26-6 is a CRISPR-associated protein (for other genes, see Table S2). Colors are as for A. and numbers in boxes indicate the number of nucleotide substitutions that distinguish genes distinct from the 5-way CG dataset.

5-WAY CG No.	21_51	3_15	3_65	26_6	13_13	137_1	8_70	174_2	2_52
UBA No.	8241_77	241_307	8241_357	8241_430	8241_652	8692_126	8524_247	8062_371	7931_189
UBA dataset	17	52	53	0	11	54	23	47	55
5-way CG dataset	0	0	0	0	0	0	0	0	0
A. Environmental sample: multilocus sequence analysis									
ABend									
B. Isolates: multilocus sequence typing									
CF-1	0	52	53	0	11	54	23	0	51, short
5w02LII-1	0	0	1		0		0		0
UBABS05-4	0	0	0, short	0	0	0	0	0	0
ABM05-7	0	0	0,v short	0	2	0	0	0	0

Figure Captions:

Figure 1: Ratio of the protein detection coverage (Cov Ratio) when using the UBA vs. the 5-way CG *Leptospirillum* group II protein database (green) across scaffold 7931 of the UBA composite genome (and the corresponding genome regions in the composite 5-way CG genome). In addition, the percentage amino acid identity is plotted for all ortholog pairs (black) and tRNA genes are marked (orange). UUE = expressed unique gene in UBA genome; UE = expressed shared gene only detected using UBA database; UCE = expressed unique gene in 5-way CG genome; CE = expressed shared gene only detected using 5-way CG database.

Figure 2: (A) Ratio of the protein detection coverage (see Fig.1 legend) (green) and percentage amino acid identity between orthologs (black) across scaffold 8241 of the UBA genome. (B) Number of unique 5-way CG (red) and UBA protein variant peptides (blue). (C) Single nucleotide polymorphisms (SNPs) detected in the UBA (blue) and 5-way CG genome datasets (red). Gaps in lines underneath the data indicate absence of orthologs in the respective datasets.

Figure 3: Inferred recombination structure of *Leptospirillum* group II in the ABend sample based on the number of unique 5-way CG (red) and UBA protein variant peptides (blue) detected (arbitrary order of the seven genomic fragments). Percentage amino acid identity between orthologs of the UBA and 5-way CG genomic datasets is indicated on the inner ring, scaled from 50-100% (purple: 100% identity, gap: no ortholog). Locations of genes analyzed by MLST are indicated on the outer ring, colored depending on the

nucleotide sequence type (Table 1; red: 5-way CG type, blue: UBA type, black: no data, purple: 5-way CG = UBA).

Figure 4: (A), (B) Amino acid alignments for orthologs of a *Leptospirillum* group II protein from the region shown in (C) (also see Fig. 2): 5-way CG protein variant sequence on the top line and UBA-type sequence on the bottom line. Peptides detected using the 5-way CG dataset are indicated in red and the UBA database in blue. Red and blue circles indicate positions of amino acid substitutions within detected peptides that distinguish the protein variants (for adjacent peptides, peptide boundaries are not shown). The open circle indicates substitution of isoleucine (I) for leucine (L). (C) Region between UBA_LeptoII_Scaffold_8241_GENE_317 and 8241_330 and corresponding 5-way CG proteins. The gene calls are indicated by boxes, colored yellow when proteins were detected. Positions of amino acid substitutions in the detected peptides that discriminate between the two variants are indicated (colored bars). Dashed lines indicate peptides detected with low confidence (one peptide in one run).

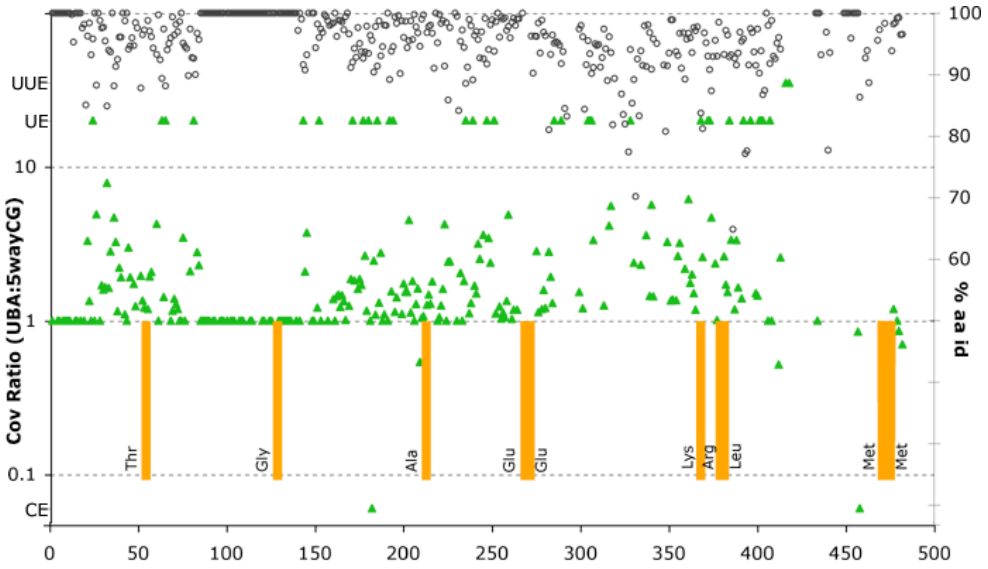
Cov Ratio (UBA:5wayCG)
% aa id
UUE
UE
10
1
0.1
CE
100
90
80
70
60
0
50
100
150
200
250
300
350
400
450
500
Thr
Gly
Ala
Glu
Glu
Lys
Arg
Leu
Met
Met

Figure 1

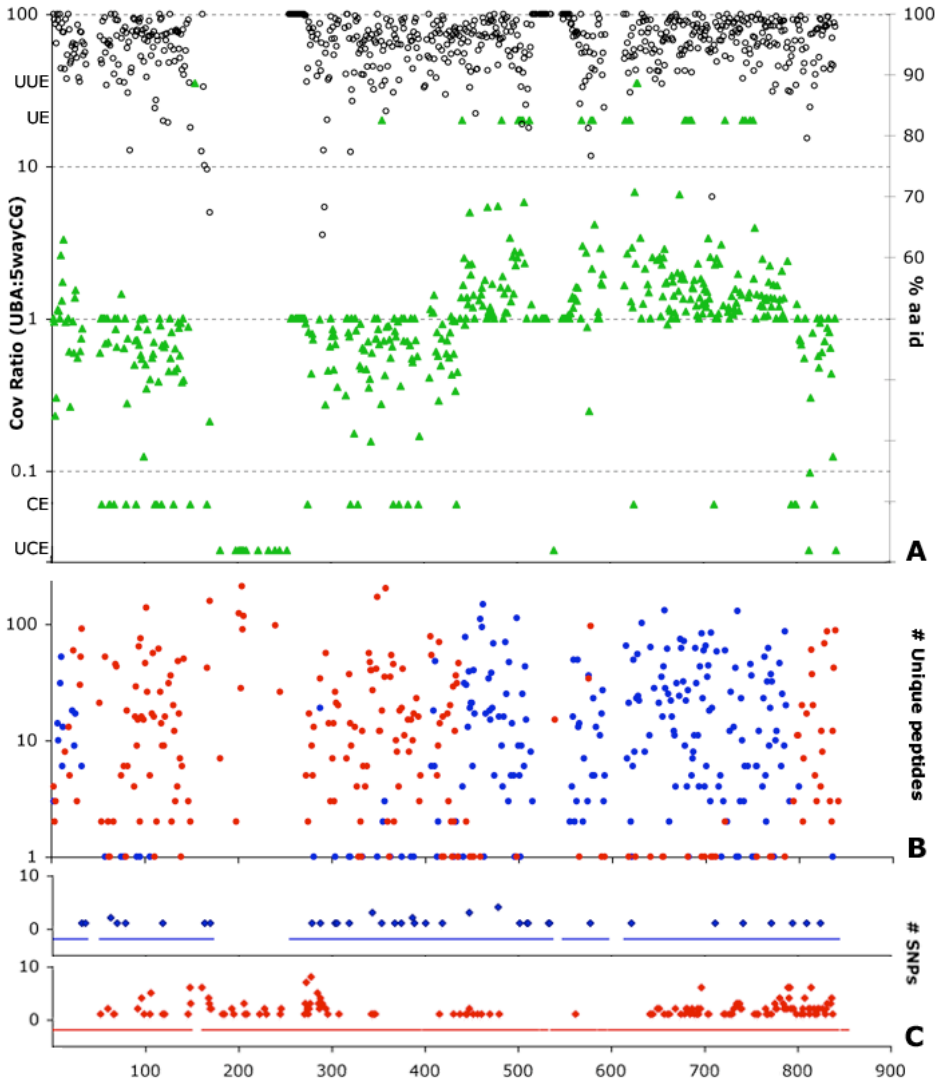
100
UUE
UE
10
1
0.1
CE
UCE
Cov Ratio (UBA:5wayCG)
100
90
80
70
60
% aa id
A
100
10
1
Unique peptides
B
10
0
10
0
SNPs
C
100
200
300
400
500
600
700
800
900

Figure 2

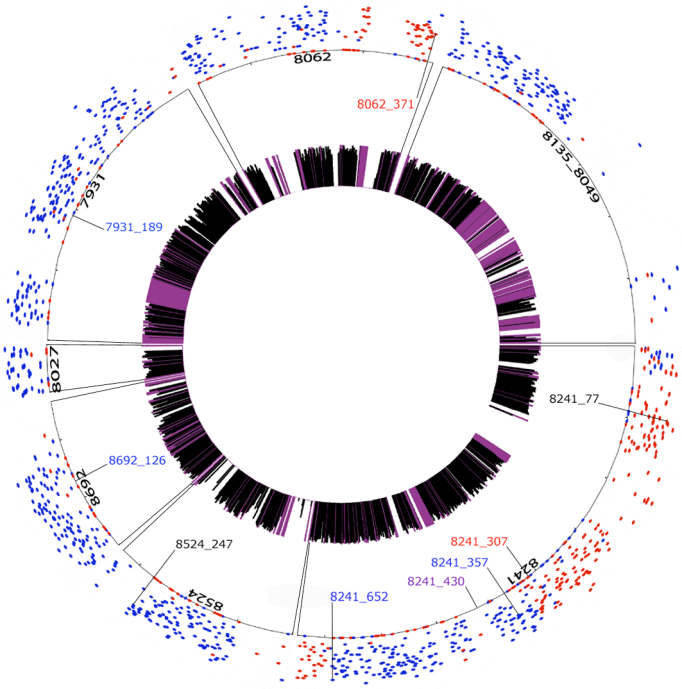
8062
8062_371
8135_8049
7931
7931_189
8027
8241_77
8692_126
8692
8524_247
8241_307
8241_357
8241_430
8241
8524
8241_652

Figure 3

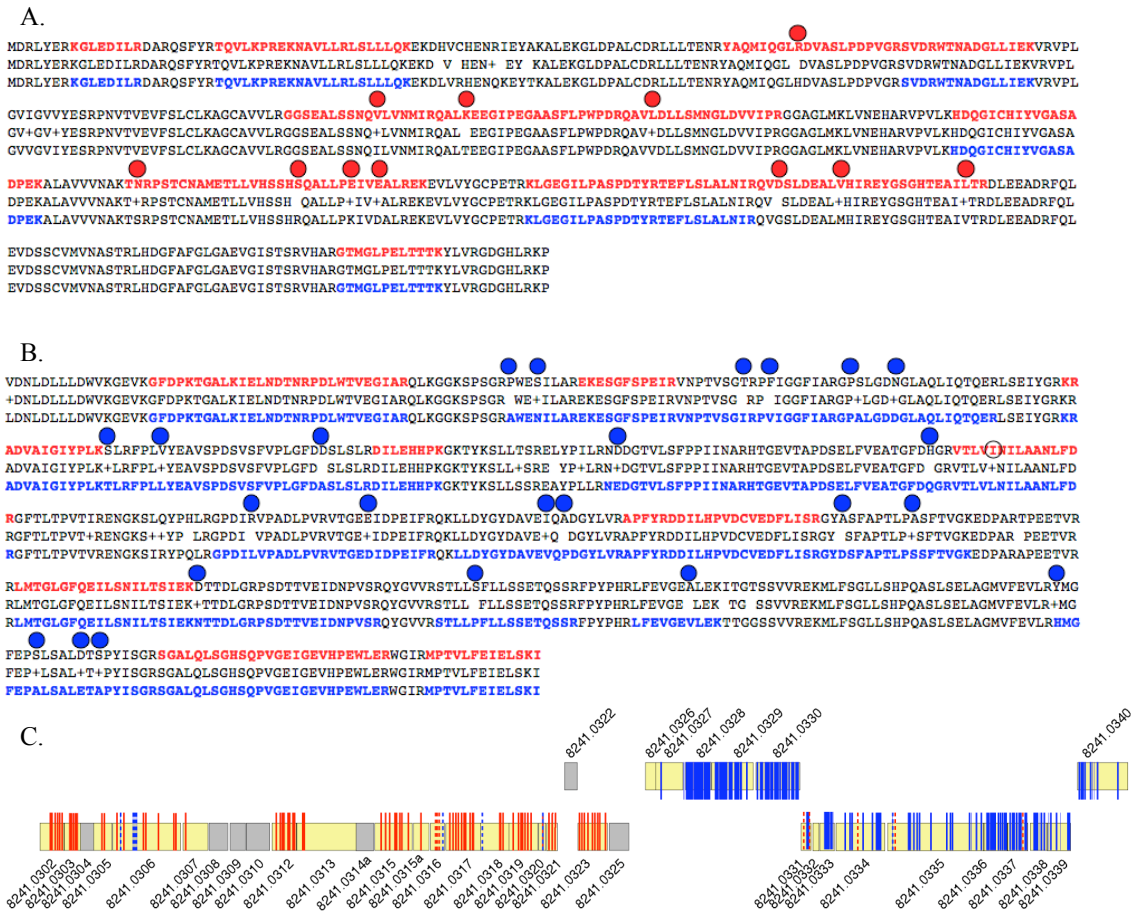
A.
B.
C.
8241.0322
8241.0326
8241.0327
8241.0328
8241.0329
8241.0330
8241.0340
8241.0302
8241.0303
8241.0304
8241.0305
8241.0306
8241.0307
8241.0308
8241.0309
8241.0310
8241.0312
8241.0313
8241.0314a
8241.0315
8241.0315a
8241.0316
8241.0317
8241.0318
8241.0319
8241.0320
8241.0321
8241.0323
8241.0325
8241.0331
8241.0332
8241.0333
8241.0334
8241.0335
8241.0336
8241.0337
8241.0338
8241.0339

Figure 4