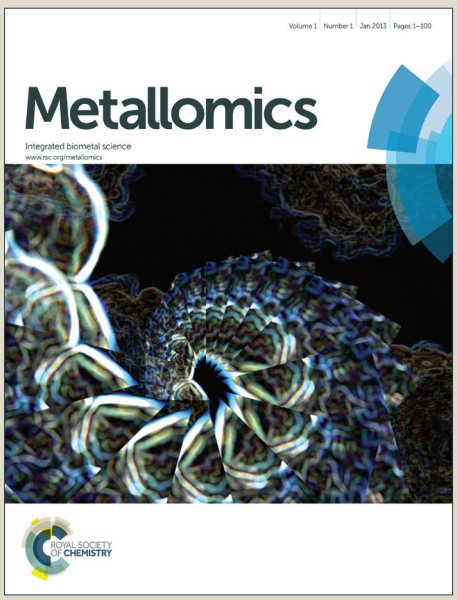
Metallomics

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/metallomics

Mapping the Protein Profile Involved in the Biotransformation of Organoarsenicals Using an Arsenic Metabolizing Bacterium.

John A. Thomas,[a] Peter Chovanec,[a,b] John F. Stolz [b] and Partha Basu*, [a]

[a]Department of Chemistry and Biochemistry, and [b]Biological Sciences

Duquesne University, Pittsburgh, PA 15282

Abstract

Alkaliphilus oremlandii strain OhILAs, a gram-positive bacterium, has been shown to ferment lactate as well as use arsenate and roxarsone as a terminal electron acceptor. This study examines the proteome expressed under four growth conditions to further elucidate the bacterial metabolism of inorganic and organic arsenic. The four growth conditions include, sodium lactate (as fermentative control), sodium lactate with 3-nitro-4-hydroxybenzenearsonic acid (roxarsone), sodium lactate with 3-amino-4-hydroxybenzenearsonic acid (3A4HBAA), and sodium lactate with sodium arsenate. Shotgun proteomics using LC-MS/MS was performed on the soluble cytoplasm as well as solubilized membrane proteins using perfluorooctanoic acid, a surfactant with similar properties to sodium dodecyl sulfate. The MS/MS data were analyzed using the Spectrum Mills proteomics workbench. Positive protein matches were confirmed with protein scores of 20 or greater and the presence of two or more peptides among the three technical replicates. A total of 1357 proteins (out of 2836 predicted) were identified with 791 in sodium lactate, 816 in sodium lactate and roxarsone, 715 in sodium lactate and 3A4HBAA, and 733 in sodium lactate and arsenate. The relative abundances of these proteins were determined using a method called normalized spectral abundance factor (NSAF). Proteins that were identified in both the control and the experimental conditions were compared using the Power Law Global Error Model (PLGEM) to determine proteins that were significantly up or down regulated. All putative proteins were assigned functions and pathways using the COG databases, however a large number of proteins were classified as hypothetical or had unknown function. Using the statistical information and known functionalities of the identified proteins a pathway for the degradation of roxarsone and 3A4HBAA by *A. oremlandii* strain OhILAs is proposed.

Introduction

Many arsonic acid derivatives, particularly those with an As–C bond can pose a significant impact on human health and the environment. Organoarsenicals, in particular, have seen considerable use as antibiotics to treat various diseases in livestock.[1-3] A recent editorial in Science Magazine called for real change in the use of many of these antibiotics.[4] One in particular, 3-nitro-4-hydroxybenzenearsonic acid, also known as roxarsone (structure shown in Figure 1), has been under scrutiny for its extensive use in the production of broiler chickens and as

Insert Figure 1

of June 8th, 2011, has been prohibited by the Food and Drug Administration (FDA) for this use. Roxarsone was used as a chicken feed additive to aid in the prevention of the parasitic infection coccidiosis;[5] as well as having the added benefit of stimulating tissue growth and improvement of tissue pigmentation via increased angiogenesis.[6] The majority of the arsenic additive is not retained by tissue but is excreted into the environment,[7] in a highly soluble form[8] and can be taken up by certain plants.[9] In 2009 the United States Department of Agriculture (USDA) reported that the United States produced 8.6 billion broiler chickens.[10] The average broiler chicken produces 1.46 to 2.67 kilograms of waste in its lifespan, amounting to an estimated 12-23 billion kilograms of chicken litter per year.[11] The contaminated chicken litter can contain from 15 to 48 mg/kg of organoarsenicals.[12] Approximately 90% of the litter is used for land application.[13] The arsenic may seep into ground water and find its way into the air in the form of dust. As a result, arsenic concentration in the soil as well as the presence of arsenicals in house dust were found to be increased in farming communities surrounding concentrated animal feeding operations (CAFOs).[11, 14]

Since the majority of roxarsone is excreted, it would be expected that it is the major arsenic species in fresh litter. However, in composted litter the prevailing arsenic species is inorganic arsenate (As(V)).[12, 15] Inorganic arsenate is not only considered a more toxic arsenic

species but can also readily leach into groundwater supplies.[8, 16] Previous studies have suggested that biological processes are responsible for the transformation of roxarsone to inorganic arsenate.[12, 17, 18] *Alkaliphilus oremlandii* OhILAs is one organism that has demonstrated arsenate-respiring capabilities.[19] It has been described that as roxarsone is degraded by *A. oremlandii*, 3-amino-4-hydroxybenzenearsonic acid (3A4HBAA) (Figure 1) and arsenate are produced.[18] The production of 3A4HBAA is proposed to result from the following equation:[18]

$$2 \text{ roxarsone} + 3 \text{ lactate} \rightarrow 2 \text{ 3A4HBAA} + 3 \text{ acetate} + 3\ CO_2 + H_2O \quad (1)$$

Although the degradation of roxarsone and the release of inorganic arsenic have been demonstrated, the pathway by which it occurs has as yet been elucidated. As a strict anaerobe, *A. oremlandii* does not possess mono or dioxygenases, nor has free oxygen been shown to be involved.[18] Initial examination of its genome did not identify enzymes typically associated with aromatic degradation.[18] This investigation focused on establishing the protein profile of *A. oremlandii* grown under four different growth conditions with arsenicals provided as terminal electron acceptors. Three arsenic species used were: roxarsone, 3A4HBAA, and arsenate. With relative quantitation coupled with statistical analysis we have identified differentially expressed proteins that are statistically significant that provides a clue into the mechanism of arsenical degradation. Our approach has been to use the statistically significant peptides as markers to understand changes in protein profile, which ultimately allows one to predict functions within the cell. This methodology we termed **B**iomarker **I**dentification using **S**tructure-**A**ctivity based **P**roteomics (BISAP) may find a wider appeal in unraveling complex biological processes.

Material and Methods

Bacterial media and culture conditions. Cells of *A. oremlandii* strain OhILAs were grown anaerobically under a nitrogen atmosphere at 30°C in 125 mL Wheaton bottles as

described previously.[18, 19] Base medium was prepared with 0.42% (w/v) sodium bicarbonate, 0.0095% (w/v) magnesium sulfate, 0.1% (w/v) yeast extract (Sigma-Aldrich), 0.01% (v/v) 10 mM sodium selenite, 0.2% (v/v) 500× vitamin mix, and 0.2% (v/v) 500× trace element mix. The pH of the base medium was adjusted to 8.2 using 1N sodium hydroxide. Lactate (provided as the fermentative substrate or electron donor), and electron acceptors were prepared separately as stocks, sterilized by filtration and amended aseptically under anaerobic conditions. Four different growth conditions were used: (1) 10 mM sodium lactate, (2) 10 mM sodium lactate and 1 mM 3-nitro-4-hydroxybenzenearsonic acid (roxarsone), (3) 10 mM sodium lactate and 1 mM 3-amino-4-hydroxybenzenearsonic acid (3A4BAA), and (4) 10 mM sodium lactate and 1 mM sodium arsenate. All growth conditions were done in triplicate.

Cell Growth. To monitor cell yield for each growth condition the total protein content was quantitated using Qubit™ protein assay (supporting information Table S1) following the manufacturer's protocol. The protein concentrations of the cell lysate, soluble (cytoplasmic) fractions, and membrane fractions were also determined the by same method.

Cytoplasmic and membrane sample preparation. Each culture bottle was harvested individually, after 48 h by centrifugation (4200 x *g*). The cells were rinsed once with 700 µL of 100 mM Tris-Cl buffer (pH 7.5), and centrifuged again at 4200 x *g* for 10 min. The pellets were resuspended with 300 µL 100 mM Tris-Cl buffer and transfer to 1.5 mL Eppendorf tubes. The tubes were rinsed twice with 200 µL of 100 mM Tris-Cl buffer. The resulting 900 µL cell mixture was sonicated on ice using a Misonix XL-2000 sonicator. Soluble cytoplasmic and membrane fractions were separated by centrifugation (20817 x *g*), with an Eppendorf 5810 centrifuge for 30 min. The supernatant was decanted and stored in 1.5 mL Eppendorf tubes.

To solubilize the membrane proteins, the membrane fractions were resuspended in 200 µL 100 mM Tris-Cl buffer then centrifuged at 20817 x *g* for 30 min. This was repeated three times to reduce cytoplasmic protein contamination. Membrane factions were resuspended in 250 µL of 0.5% perfluorooctanoic acid (PFOA) buffer (0.6% (w/v) sodium chloride, 0.6% (w/v)

tris, 0.5% PFOA, 25 mL nano-pure water, pH adjusted to 8.0 using sodium hydroxide).[20, 21] The fractions were sonicated on ice and then agitated at 4 °C for 1 h at 4200 x *g*. Solubilized membrane proteins were separated from unsolubilized membrane proteins and debris by centrifugation for fifteen minutes at 20817 x *g*.

For LC-MS/MS analyses, 20 µg of sample was placed into 1.5 mL Eppendorf tubes. Equal volumes of trifluoroethanol (TFE) and sample, and 2.5 µL of 200 mM dithiothreitol (DTT) were added and heated at 90°C for 20 min to reduce and denature the proteins. A 10 µL aliquot of 200 mM iodoacetamide (IAM) was added to the samples, then covered and incubated at room temperature for 60 min. To this 2.5 µL of 200 mM DTT was added, and the samples were again covered and incubated at room temperature for 60 min. Samples were then completely dried using a Forma Speed Vac, and resuspended in 50 μL ammonium bicarbonate buffer (100 mM). To remove the salts and lipids, proteins were precipitated with cold acetone (100 μL) followed by centrifugation at 20817 x *g*. This step was repeated to ensure salts and lipids were removed. The samples were again dried and resuspended with 50 μL of 100 mM ammonium bicarbonate buffer, and digested with trypsin (1:20; v/v trypsin-to-protein) at 30°C for 16 h. Samples were dried and then reconstituted with 20 µL of filtered deionized water and 0.1% formic acid for LC-MS/MS analysis. For the analysis of solubilized membrane proteins, 20 µg of proteins were decanted into 1.5 mL Eppendorf tubes. An aliquot of 40 μL of 100 mM ammonium bicarbonate buffer was added to each sample, followed by trypsin digestion as described above. Samples were dried using a Forma Speed Vac and then reconstituted with 20 µL of filtered water and 0.1% formic acid for LC-MS/MS analysis. A total of 72 samples were analyzed by mass spectrometry.

Shotgun LC-MS/MS Analysis. To analyze peptides generated from the tryptic digestion 2 µL of sample was used for each injection. A 1200 series liquid chromatography (Agilent Technology), HPLC-Chip Cube MS interface (Agilent Technology), 6530 Q-TOF mass

spectrometer (Agilent Technology), and an HPLC-Chip with 160 nL enrichment column was used for this investigation. Peptides were separated using a 150 mm × 75 mm analytical column packed with Zorbax 300 SB-C18 (5 mm particles, Agilent Technology). Samples were loaded onto the enrichment column using a mixture of 97% solvent A (97% water, 3% acetonitrile, 0.1% formic acid) and 3% solvent B (100% acetonitrile, 0.1% formic acid). A 5 min enrichment period was used prior to analysis. The capillary flow pump was set to 4 µL/min using 100% solvent A. Different gradients were used for the cytoplasmic and solubilized membrane samples to elute the peptides. Cytoplasmic and solubilized membrane samples were eluted with different LC methods continued for 110 min with the same flow rate (see supporting information Figure S1). The total runtime, including reconditioning of the column, was 110 min. The column effluent was directly coupled to the 6530 Q-TOF mass spectrometer using the HPLC-Chip Cube nanospray source. The source was set to a capillary voltage gradient in correspondence to the gradient of percent solvent B throughout the runtime. Data was acquired using Mass-Hunter Workstation (Version B2116.30) in positive ionization mode. Gas temperature, drying gas volume, fragmentor voltage, skimmer voltage, and octopole RF were set to 325°C, 5 L/min, 180 V, 65 V, and 750 V. Auto MS/MS was performed with a total cycle time of 2.1 s. In each cycle MS spectra were acquired at 1 Hz (1 spectrum/sec; m/z 250-1500) and the three most-abundant ions (with charge states +2, +3, >+3) exceeding 1000 counts were selected for MS/MS at 3 Hz (3 spectra/sec; m/z 50-3200). Medium isolation (4 m/z) window was used for precursor selection. Collision energy was set to a slope of 3V per 100 Da with a 2 V offset. Precursor exclusion was set to exclude after the first spectra and released after 0.25 min.

Data Processing and Analysis. Collision induced dissociation (CID) data were extracted using the Data Extractor in the Spectrum Mills software (Agilent Technology, v A.03.03.084SR4). Modifications for the data extraction were set to carbamidomethylation, and the MS/MS spectral features were set to $[M+H^+]$ 600.0 to 4000.0 Da, scan time range: 0 to 110 min, merge scans with same precursor m/z: ±10 sec and ± 0.5 m/z with no similarity merging.

Precursor ions required a minimum signal to noise ratio of 25 and were assigned maximum charge of 7. A National Center for Biotechnology Information (NCBI) non-redundant database was created from the *Alkaliphilus oremlandii* strain OhILAs proteome, and was used to reduce search time from the extracted CID data. The database was searched for peptide fragments resulting from tryptic digestions allowing for 2 missed cleavages. The precursor mass tolerance was set to ±2.5 Da and the product mass tolerance was set to ±0.7 Da with a maximum ambiguous precursor charge set to 3 as described in the literature.[22] Reverse database scoring was used to reduce false identification and the search mode was set to "Identity". Autovalidation was performed first for protein details and second by peptide mode. Proteins were filtered by a score of >20.0 and six rules were uses to determine validity shown in Table S2. The forward minus reverse score threshold was set to >1, and the rank 1 minus rank 2 threshold was set to >1. The peptide mode was auto-validated using five rules to determine validity shown in Table S2. The forward minus reverse score threshold was set to >2, and rank 1 minus rank 2 threshold was set to >2. The protein hits were summarized in both a peptide summary and a protein summary using the Spectrum Mills software, recording hits that were scored >5 and percent score based on percent intensity (SPI) >60%. The SPI was calculated by comparing the fragmentation pattern to that of a theoretical fragmentation pattern for the peptide in question. A second MS/MS search was conducted to identify proteins of variable modifications. The parameters for the second search were the same except that the search mode was set to "Variable modification". The reverse database score was calculated as well as dynamic peak thresholding. The precursor mass shift range was set to -18.0 to 177.0 Da. The variable modifications consisted of carboxymethylation, oxidized methionine, pyroglutamic acid, phosphorylated sulfur, phosphorylated tyrosine, and phosphorylated threonine. Again, autovalidation was performed using the same parameters as before.

Label-free Shotgun Quantification. To reduce the number of falsely identifications, positive protein identification required a protein score of 20 or greater and 2 peptides identified

within the three technical replicates for each biological replicate.[23, 24] The Spectrum Mills proteomic workbench utilizes two scoring methods to reduce the number of false identifications. The peptide score is a score based on the fragmentation pattern of the peptide in the MS/MS spectrum. The protein score reflects the added score of all peptides identified for the protein. Scores greater than 15 represent valid protein results using a Q-TOF (Agilent-site manual). The proteins identified were compiled and normalized using the normalized spectral abundance factor (NSAF).[25-27] Cells grown in each of the four conditions were analyzed with three biological replicates and three technical replicates for each biological replicate. Mean and standard deviations using the normalized data were used to validate the proteins identified under each growth condition allowing for a statistically robust comparison of abundance.

Using an R based add-on package, PLGEM (Power Law Global Error Model), proteins that were similarly expressed in control and experimental conditions were used to normalize the data in order to identify proteins that were differentially expressed (e.g., up or down regulated). PLGEM was used to model the variance-versus-mean dependence in the proteomic data.[28]

Results and Discussion

The overall approach of our proteomic study followed previously developed methods,[29, 30] however, improvements in sample preparation and refinement in data analyses resulted in the generation of a much larger dataset. Although a total of 2552 proteins were identified, only 1357 met both requirements of having a score of 20 (or higher) and 2 or more peptides present in the 3 technical replicates. Of the total positively identified proteins 152 proteins were unique to the solubilized membrane fractions, 666 proteins were unique to the cytoplasmic fraction, and 539 were identified in both. Figure 2 depicts the total number of proteins identified in each growth conditions (i.e., in cytoplasmic and solubilized membrane fractions combined). A similar comparison of the cytoplasmic fractions and solubilized membrane fractions as a function under each growth condition is presented in supplementary information (Figure S6).

Previously we have reported that *A. oremlandii* has the ability to respire on environmentally relevant arsenicals to yield inorganic arsenic.[18] A bioinformatic search using databases including MetaCyc, and Joint Genome Institute (JGI) of *A. oremlandii* proteome identified nineteen proteins that can be ascribed to arsenic metabolism. These proteins include transcriptional regulators, arsenical resistance proteins, transporters, arsenical detoxification, ATPases, and reductases (Table 1). These nineteen proteins however may be a small fraction

Insert Table 1

of proteins utilized by this bacterium to metabolize benzene arsonic acid or its derivatives. Of the 2836 ORFs encoded in the genome, 814 are annotated as hypothetical proteins, 115 are protein of unknown function, and 10 have a domain of unknown function. That accounts for 934 proteins whose functions are poorly described or understood. This translates to ~32% of the proteome lacking information related to function. Around 25% of the proteins we identified in this study are annotated as “hypothetical”, “general function”, or “proteins of unknown function”.

Insert Figure 2

The reproducibility of the identified proteins have been compared both in technical and biological replicates. The reproducibility of the technical replicates for both the cytoplasmic and solubilized membrane fractions can be seen in Venn diagrams presented in the supplemental material, Figures S2 and S3. The reproducibility of the biological replicates for the cytoplasmic and solubilized membrane fractions can be seen in Venn diagrams in the supplemental material, Figures S4 and S5. Table 3 shows the percentages of the proteins identified from the four different growth conditions, while the most abundant proteins found calculated using NSAF methodology under these conditions are shown in Table 4.

Insert Tables 2 and 3

Sodium lactate grown cells (i.e., fermentative growth) were used as a control and under this condition 791 proteins were identified meeting the criteria stated above. Under the three experimental conditions i.e., sodium lactate and roxarsone, sodium lactate and 3A4HBAA, and

sodium lactate and arsenate 816, 715, 733 proteins were identified, respectively. In the cytoplasmic fractions 149, 146, 83, and 113 proteins were found to be unique to lactate grown cells, roxarsone and lactate-grown cells, 3A4HBAA with lactate grown cells, and arsenate with lactate grown cells, respectively. There were 578 proteins that were common in the control (i.e., lactate only) and experimental conditions (Figure 2). Of the 578 proteins, 67 proteins were common to lactate, and roxarsone with lactate samples; 46 proteins were unique to lactate, and 3A4HBAA with lactate samples; 52 proteins were unique to lactate, and arsenate and lactate samples; 65 proteins were unique to lactate, roxarsone with lactate, and 3A4HBAA with lactate samples; 33 proteins were unique to lactate, 3A4HBAA with lactate, and arsenate with lactate samples; 49 proteins that were unique to lactate, roxarsone with lactate, and arsenate with lactate samples; and 266 proteins that were common to all conditions. 25 proteins were identified as common in all three experimental conditions; 30 proteins were unique to roxarsone with lactate samples, and 3A4HBAA with lactate samples; 43 proteins were unique to roxarsone with lactate samples, and arsenate with lactate samples; 38 proteins were unique to 3A4HBAA and lactate samples, and arsenate with lactate.

The 539 proteins that were common to both cytoplasmic and solubilized membrane fractions were filtered from the solubilized membrane fraction comparison (Supporting information, Figure S6). In solubilized membrane fractions 20, 44, 35, and 29 proteins were found to be unique to sodium lactate grown cells, roxarsone with lactate grown cells, 3A4HBAA with lactate grown cells, and arsenate with lactate grown cells, respectively. There were 8 proteins that were common in the lactate grown control and experimental conditions. Of the 8 proteins, 3 proteins were unique to lactate, and roxarsone and lactate samples, 1 protein that was unique to lactate, and arsenate and lactate, and 4 proteins that were common to all conditions. There were 16 proteins that were common in all the experimental conditions. Of those, 4 proteins were unique to roxarsone and lactate samples, and 3A4HBAA and lactate samples, 5 proteins that were unique to roxarsone and lactate samples, and arsenate and

lactate samples, and 7 proteins that were unique to 3A4HBAA and lactate samples, and arsenate and lactate. The most expressed proteins identified can be seen in Table 4 based on their relative concentrations calculated.

It has been previously reported that in the *A. oremlandii* genome two specific regions (shown in Figure 3) were of interest.[18] The first is referred to as "Arsenic Island" and the second the aldehyde ferredoxin oxidoreductase operon. The "Arsenic Island", contains genes encoding proteins that may be used in the metabolism and detoxification of arsenic. A few noted proteins in this region are Clos_1096, the catalytic subunit ArrA the respiratory arsenate reductase, Clos_1097 the iron-sulfur containing electron transfer subunit ArrB, Clos_1109 a trans-acting repressor ArsD, and Clos_1110 an arsenite efflux ATP-binding protein, ArsA. The latter two (ArsD and ArsA) are involved in arsenic resistance. Interestingly Clos_1096, ArrA, was found to be expressed under all growth conditions but Clos_1097, ArrB, was only identified in cells grown with roxarsone and arsenate. The second region of interest was the AOR operon, which contains genes encoding two aldehyde ferredoxin oxidoreductases, Clos_2140, and Clos_2143. Previously it was reported that Clos_2140 was found to be expressed higher than Clos_2143 in roxarsone grown cells from 2D-gel electrophoresis and similar results were found here.[29]

Insert Figure 3

In this study a total of 1357 protein were identified with a protein score of 20 or greater and 2 or more peptides identified in the technical replicates. This amounts to nearly 48% of the total proteome of *A. oremlandii* OhILAs as there are 2836 genes that can potentially encode for a protein. It is instructive to point out that some of the genes may be redundant and may have lost function. Some of these genes may also be involved in nucleic acid synthesis. Thus, 48% is a very conservative estimate. When comparing the number of proteins identified in each condition all four conditions resulted in a similar number of identified proteins. There were 791 proteins identified in the sodium lactate grown cells. In the three arsenic experimental conditions 816

proteins were identified in roxarsone grown cells, 719 proteins identified in 3A4HBAA, and 733 proteins identified in arsenate grown cells.

The proteins that were identified were placed into their known Cluster of Orthologous Groups (COG) classes to examine overall shifts in proteome expression in each condition. Proteins were separated into two categories: on/off and up/down. Thus, proteins that uniquely expressed under a condition is termed as 'on' and if a protein is not expressed it is termed as 'off'. In the case of 'up' or 'down', proteins are expressed in control (i.e., lactate) as well as in experimental conditions (i.e., roxarsone, 3A4HBAA, and arsenate). In this case, their abundances are different as judged by their relative NSAF score. The NSAF scores were analyzed using PLGEM, and only those that satisfied the criteria were included in the analysis. Proteins that were similarly expressed in control and experimental were used as a baseline in order to identify proteins that were statistically up and down regulated (Figure 4). The red dashed lines demonstrate the threshold for proteins that were up or down regulated by at least two fold. The orange dash lines separate the significant changes in the protein abundances as evidenced by their p-values. Thus, proteins satisfying the above-mentioned criteria (e.g., exhibiting statistically significant changes) are highlighted with purple circles, and considered for discussion.

Insert Figure 4

We found that the most effective way to analyze this large of a dataset was to divide it into smaller sections. Proteins that were on/off or significantly differentially expressed (i.e., up or down) were further classified into four COG superclasses: metabolism, cellular processes and signaling, information storage, and processing poorly defined. The four superclasses are shown with pie plots (Supporting information, Figure S7). Each superclass is further categorized into subclasses based on their function. For example, proteins in the superclass 'metabolism' are categorized into eight subclasses: amino acid metabolism, carbohydrate metabolism, energy metabolism, inorganic metabolism, lipid metabolism, nucleotide metabolism, secondary

metabolism, and coenzyme metabolism. Similarly proteins in 'cellular processes and signaling' are categorized into seven subclasses: cell cycle, cell motility, cell wall, defense, posttranslational modification, intracellular trafficking, and signaling. Proteins in 'information storage' are subcategorized into replication, transcription, translation, and chromatin. The last superclass of 'poorly defined' proteins is also subcategorized into two categories: function unknown and general function. These subcategories are also shown in the pie chart (Supporting information, Figure S6). Of all four superclasses, more proteins from the metabolism superclass were found to be dominant as only the 'on' proteins in all four conditions. To further investigate the differences in COG functions the specific classes were plotted (Figure 5). From these plots, changes can be noted in a few of the classes. 3A4HBAA showed fewer unique proteins to amino acid metabolism, inorganic metabolism, lipid metabolism, nucleotide metabolism, coenzyme metabolism, cell cycle, cell wall, and defense proteins than other three conditions. Interestingly, 3A4HBAA condition had the most general function proteins than all other conditions. Unique proteins categorized under carbohydrate metabolism were not observed in the control (lactate) or in the As(V) grown cells, and only a few unique proteins of the same category were identified in the roxarsone and 3A4HBAA grown cells. More unique proteins in the energy metabolism class were found in the cells grown under control condition than in the other three conditions. Unique proteins belonging to the inorganic metabolism category were varied in the As(V) grown cells and especially in roxarsone grown cells. As(V) grown cells had twice as many proteins identified in the nucleotide metabolism category than in the control. There were no unique proteins identified in the control for secondary metabolism but were identified in all other conditions with roxarsone grown cells showing the most. Roxarsone grown cells had the fewest unique proteins identified to have transcription function but the more for translation. The 3A4HBAA and As(V) grown cells showed the fewest translation unique proteins. A possible explanation for this is that most bacterial genes are constitutively transcribed unless repressed by metabolites or proteins. The control is more often due to derepression rather than

activation. All three experimental conditions showed unique proteins in cell cycle and cell motility that were not seen in lactate grown cells. Lactate grown cells showed the fewest post-translation modification unique proteins. As(V) grown cells had the fewest number of unique proteins involved in signaling function.

Insert Figure 5

There was more variability in the four superclasses in the up or down regulated proteins (Figure 6). There were a number of up regulated 'information storage' proteins in roxarsone grown cells than any other experimental condition. In roxarsone and arsenate grown cells the largest number of proteins classified under metabolism were down regulated. It is interesting that the bacteria were able to respire arsenate but it seems that there is a shift in the protein profile particularly those belonging to metabolism category. In both roxarsone grown cells and arsenate grown cells, more proteins belonging to metabolism category were down regulated than upregulated, whereas an opposite trend is observed in the case of cells grown with 3A4HBAA. Interestingly, cells grown with roxarsone or 3A4HBAA there is an increase in proteins in translational category, but the opposite is observed in cells grown with arsenate.

Insert Figure 6

One of the aims of this investigation was to identify the protein profile of *Alkaliphilus oremlandii* OhILAs and use protein profile under different growth conditions to propose a pathway for the biologically mediated liberation of arsenic from roxarsone. The proposed scheme is presented in Figure 7. The pathway begins with the arsenate group being reduced by Clos_1096, ArrA. A CoA ligase then binds CoA to the arsenic center, activating the ring by withdrawing electron density. The nitro group is reduced by one equivalent and in the case of 3A4HBAA is oxidized. The hydroxylamine is replaced by a hydroxyl. Several dehydrogenase reactions remove the hydroxyl groups. The system is then protonated to remove one of the double bonds. The ring is then hydroxylated twice. One hydroxyl group is then oxidized and the

ring is cleaved by a dehydrogenase reaction. The final compound can then be further degraded via the fatty acid pathway.

Insert Figure 7

The reaction scheme has seven metal containing proteins, four molybdenum and iron, one molybdenum, and two iron containing proteins. The four molybdenum and iron containing enzymes are molybdopterin oxidoreductase (Clos_1096), aldehyde oxidase and xanthine dehydrogenase molybdopterin binding (Clos_0879), aldehyde oxidase and xanthine dehydrogenase a/b hammerhead (Clos_0377), aldehyde oxidase and xanthine dehydrogenase molybdopterin binding (Clos_0264). All of the iron and molybdenum containing enzymes have oxidoreductase activity. The two iron containing proteins include iron-containing alcohol dehydrogenase (Clos_0794), iron-containing alcohol dehydrogenase (Clos_2306). Both of these proteins have oxidoreductase activity specifically acting on CH-OH groups. The protein containing molybdenum is UBA/THIF-type NAD/FAD binding protein (Clos_1321). The proposed ring cleavage pathway of roxarsone and 3A4HBAA in *Alkaliphilus oremlandii* OhILAs can be seen in Figure 7. As mentioned previously that a large number of poorly defined proteins were detected in this large-scale proteomics investigation. While this information is valuable understanding the protein profile, the findings by itself do not confirm the mechanism of roxarsone degradation. Therefore any proposal like the one proposed in here must be taken as such and await experimental validation. However, it is important to point out that a recent report[31] of ArsI, a non-heme dioxygenase, functioning as As:C lyase in roxarsone degradation is unlikely to be operative as there is no dioxygenase or monooxygenase present in the genome of *A. oremlandii* OhILAs.

Summary

Previous studies have demonstrated that *A. remlandii* has the capability to respire roxarsone and arsenate.[19] We not only have corroborated those observations but also report the growth in

the presence of 3A4HBAA. The identified proteins were divided into proteins that were expressed and not expressed as well as up or down regulated. The proteins were then placed into their COG functional groups to identify potential changes in the protein profile of the bacterium grown on different arsenic compounds. Our study of the proteome of *A. oremlandii* expressed under four different conditions has allowed us to expand our understanding of microbial metabolism of organoarsenicals. We have been able to corroborate previous observations of its growth on lactate fermentatively and respiration with arsenate and roxarsone as terminal electron acceptors, but also report growth in the presence of 3A4HBAA. The improved sample preparation and high throughput shotgun analyses resulted in the detection of 48% of the predicted ORFs (1357 out of 2836). The protein profiles for each condition showed several interesting changes, including up regulation of proteins implicated in roxarsone and arsenate metabolism. Organizing the proteins into their COG functional groups allowed for the assessment of the physiological and metabolic responses to the different arsenic compounds. We propose a pathway for roxarsone degradation that involves several molybdenum enzymes. This study also revealed the need for further improvements to genome annotation as nearly 25% of the detected proteins were annotated as hypothetical, general function, or proteins of unknown function. While it is encouraging that we have been able to reduce the number of hypothetical proteins in the *A. oremlandii* genome and suggest their functionality by using domain prediction software, further work is required to confirm these predictions.

Acknowledgment. We thank Dr. Vadi Bhatt for preliminary experiments and discussions on proteomics. We also thank Prof. Sati Mazumdar for stimulating discussions on different statistical approaches. And finally we would like to thank the Center of Excellence for Mass Spectrometry at Duquesne University supported by NSF DBI 0821401.

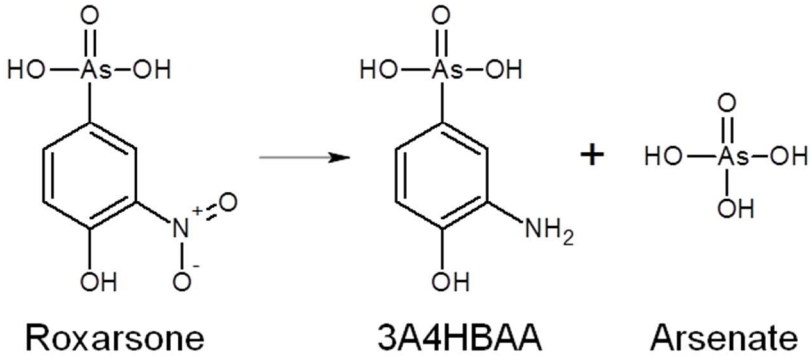


Figure 1: Degradation of roxarsone has been shown to produce 3-amino-4-hydroxybenzenearsonic acid (3A4HBAA) by reduction of the nitro group as well as arsenate formation by an unknown mechanism.

Table 1: Proteins present in *Alkaliphillus oremlandii* OhILAs proteome that may be involved in arsenic metabolism. Proteins highlighted in red are utilized in the arsenate detoxification pathway.

Locus Tag	Gene Proctuct Name	COG Functional Group	KEGG Orthonology Term	Classification
Clos_0632	Arsenate reductase	Arsenate reductase and related proteins, glutaredoxin family	ARSC1, arsC arsenate reductase	EC: 1.20.4.1
Clos_0814	Dicarboxylate carrier protein	Na+/H+ antiporter NhaD and related arsenite permeases		TC: 9.B.48
Clos_0896	Transcriptional regulator, ArsR family	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_1085	Cadmium efflux system accessory protein	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_1086	Cadmium efflux system accessory protein		arsR ArsR family transcription regulator	
Clos_1096	Molybdopterin oxidoreductase (ArrA)	Anaerobic dehydrogenase typically selenocysteine-containing	molybdopterin oxidoreductase	EC: 1.7.2.3
Clos_1097	4Fe-4S ferredoxin iron-sulfur binding domain protein	Fe-S-cluster-containing hydrogenase component 1	4Fe-4S ferredoxin iron-sulfur binding domain-containing protein	EC: 1.97.1.9
Clos_1098	Cytoplasmic chaperone TorD family protein	Uncharacterized component of anaerobic dehydrogenase	Cytoplasmic chaperone TorD family protein	
Clos_1106	Regulatory protein, ArsR	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_1107	Arsenical resistance operon repressor	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_1109	Arsenical resistance operon trans-acting repressor ArsD		Arsenical resistance	
Clos_1110	Arsenical pump-driving ATPase	Oxyanion-translocating ATPase/Septum formation inhibitor-activating ATPase	arsA arsenite-transporting ATPase	EC: 3.6.3.16, TC: 3.A.4
Clos_1112	Arsenical resistance operon repressor	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_1113	Arsenate reductase	Protein-tyrosine-phosphatase	ARSC2, arsC arsenate reductase	EC: 1.20.4.1
Clos_1114	Arsenical-resistance protein ACR3	Arsenite efflux pump ACR3 and related permeases	Arsenite transporter, ACR3 family	TC: 2.A.59
Clos_1121	MerR	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_1590	Purine nucleoside phosphorylase	Purine nucleoside phosphorylase	punA purine-nucleoside phophorylase	EC: 2.4.2.1
Clos_1591	Purine nucleoside phosphorylase	Purine nucleoside phosphorylase	punA purine-nucleoside phophorylase	EC: 2.4.2.1
Clos_2163	Cadmium efflux system accessory protein	Predicted transcriptional regulators	arsR ArsR family transcription regulator	
Clos_2574	Low molecular weight protein tyrosine phosphatase	Protein-tyrosine-phosphatase	Protein-tyrosine phosphatase	EC: 3.1.3.48
Clos_2819	Arsenate reductase	Protein-tyrosine-phosphatase	arsC arsenate reductase	EC: 1.20.4.1
Clos_2820	Arsenical resistance operon repressor	Predicted transcriptional regulators	arsR ArsR family transcription regulator	

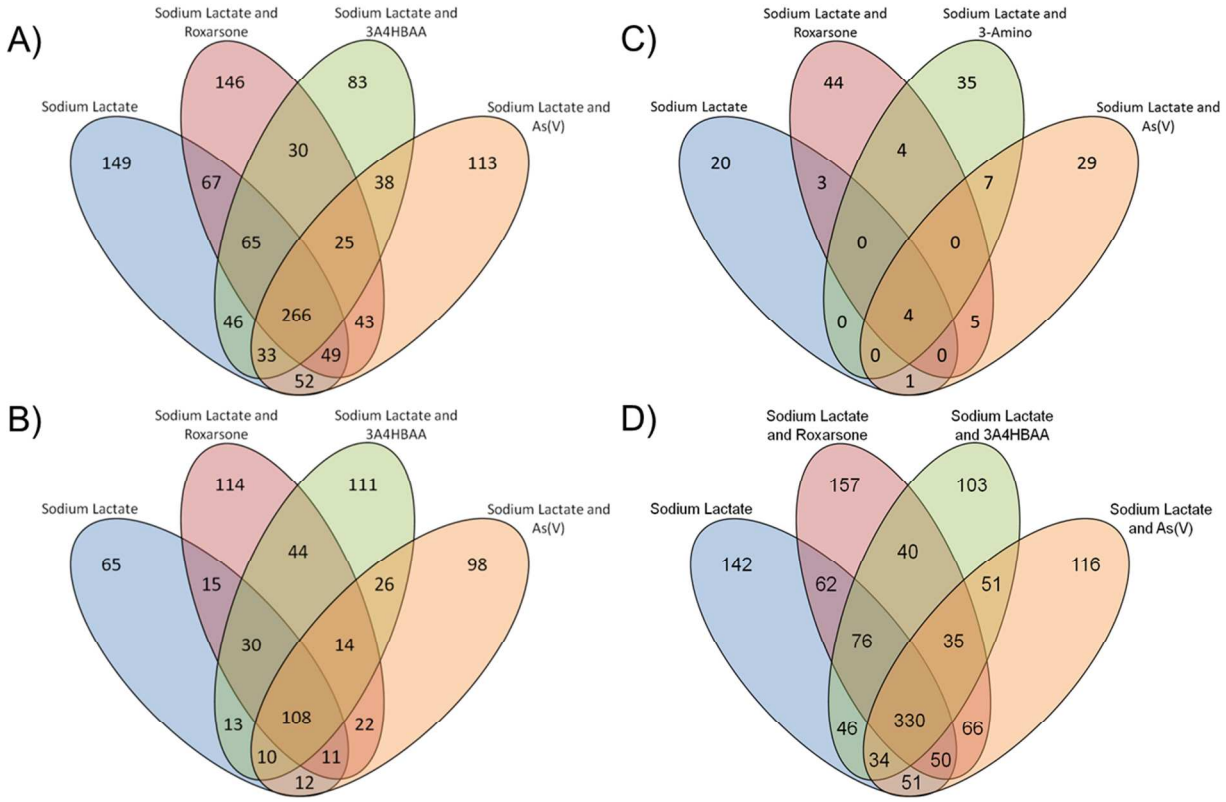


Figure 2: Comparative analysis of proteins identified by LC-MS/MS in cytoplasmic and solubilized membrane fractions isolated from *A. oremlandii* OhiLas. A) Proteins identified in cytoplasmic fractions, B) proteins identified in solubilized membrane fractions, C) proteins identified in only solubilized membrane fraction, D) all protein identified in cytoplasmic and solubilized membrane fractions.

Table 2: A comparison of the proteins identified in the cytoplasmic and solubilized membrane fractions in cells grown with sodium lactate, roxarsone with lactate, 3A4HBAA with lactate, and arsenate with lactate grown cells.

Growth Condition	Sample	Cytoplasmic		Membrane	
		% Identified in two replicates	% Identified in three replicates	% Identified in two replicates	% Identified in three replicates
Lactate Technical Replicates	1	49.2	30.8	59.5	42.3
	2	49.3	32.9	55.3	34.8
	3	48.1	27.1	50.0	33.8
Roxarsone Technical Replicates	1	45.9	32.1	51.2	30.7
	2	47.4	29.7	46.6	32.7
	3	65.8	40.2	50.2	33.3
3A4HBAA Technical Replicates	1	44.2	27.9	44.8	25.3
	2	48.1	34.5	53.6	36.2
	3	47.9	24.5	51.8	34.0
Arsenate Technical Replicates	1	48.14	31.5	51.4	36.5
	2	43.8	26.6	38.6	22.3
	3	39.3	24.6	43.6	28.2

Table 3: A comparison of the proteins identified in the cytoplasmic and soluble membrane fractions in cells grown in sodium lactate, roxarsone with lactate, 3A4HBAA with lactate, and arsenate with lactate grown cells.

Growth Condition	Cytoplasmic		Membrane	
	% Identified in two replicates	% Identified in three replicates	% Identified in two replicates	% Identified in three replicates
Lactate Biological Replicates	45.7	26.3	48.1	30.7
Roxarsone Biological Replicates	47.2	27.1	47.2	30.1
3A4HBAA Biological Replicates	47.4	27.6	38.5	20.8
Arsenate Biological Replicates	48.6	26.3	35.5	19.6

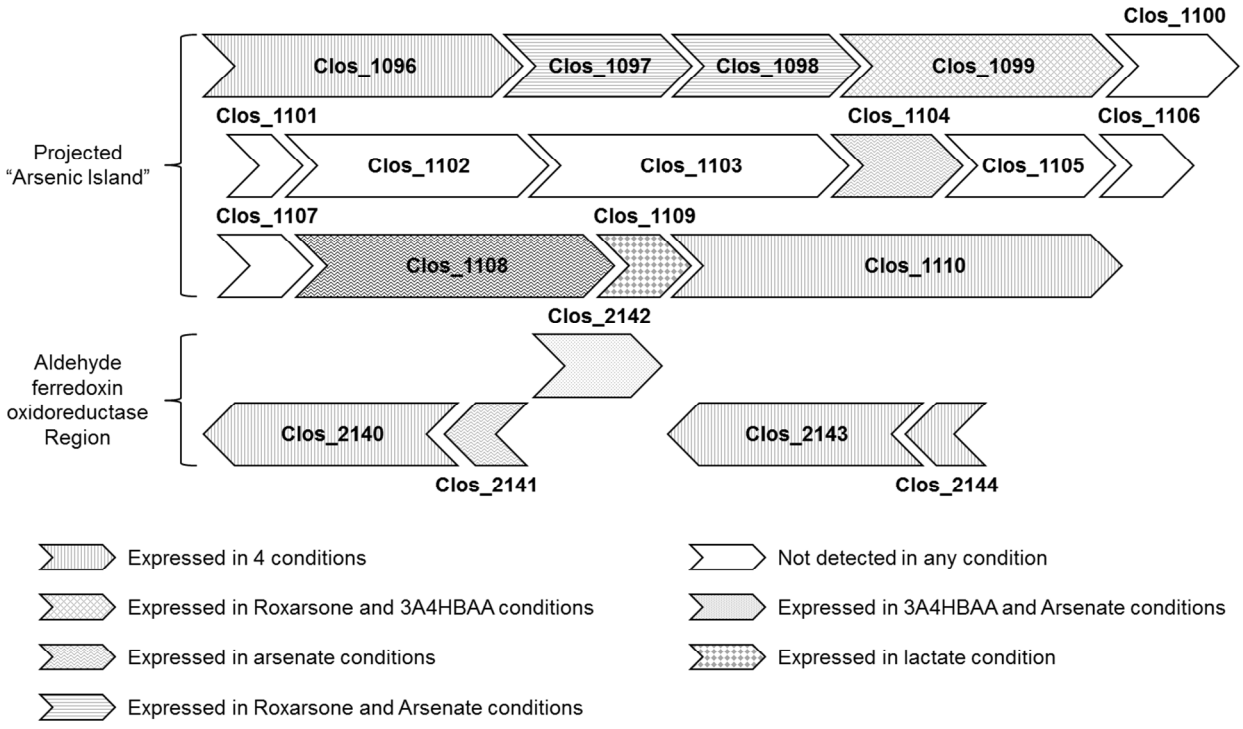


Figure 3: Organization of genes present in the arsenic island and aldehyde ferredoxin oxidoreductase region of *A. oremlandii* OhILAs identified in this study.

Table 4: Most abundant proteins according to normalized spectral abundance factor (NSAF) in sodium lactate, roxarsone with lactate, 3A4HBAA with lactate, and arsenate with lactate grown cells.

GI Number	Locus #	Protein Name	Lactate	Roxarsone	3A4HBAA	Arsenate
158139324	Clos_0066	Glycine cleavage system H protein	0.02204	0.01976	0.03577	0.01102
158139488	Clos_0237	Pyruvate-flavodoxin oxidoreductase	0.01764	0.01399	0.02928	0.01220
158139619	Clos_0369	Microcompartments protein	0.02025	0.00093	0.02278	0.00123
158139726	Clos_0476	Elongation factor Tu	0.03118	0.03011	0.04878	0.02679
158139733	Clos_0483	50S ribosomal protein L7/L12	0.03039	0.04150	0.03995	0.03395
158139741	Clos_0491	30S ribosomal protein S10	0.01489	0.01053	0.01911	0.01616
158139758	Clos_0508	50S ribosomal protein L18	0.01044	0.01062	0.02649	0.00449
158140191	Clos_0956	Sarcosine reductase	0.01054	0.00994	0.01874	0.01685
158140193	Clos_0958	Selenoprotein B, glycine/betaine/sarcosine/D-proline reductase family	0.01352	0.00944	0.02190	0.01742
158140194	Clos_0959	Sarcosine reductase	0.00797	0.00624	0.01332	0.01513
158140688	Clos_1456	Acyl carrier protein	0.02046	0.01465	0.02726	0.01370
158141230	Clos_2005	Cold-shock DNA-binding domain protein	0.01126	0.01314	0.00320	0.00000
158141400	Clos_2176	Cold-shock DNA-binding domain protein	0.01224	0.01644	0.01268	0.00505
158141679	Clos_2459	Glutamate dehydrogenase	0.02285	0.01705	0.04680	0.01798
158141845	Clos_2626	Uncharacterized protein	0.01121	0.00972	0.02623	0.01467
158141846	Clos_2627	S-layer domain protein	0.03341	0.05193	0.07805	0.05890
158141856	Clos_2637	Histone family protein DNA-binding protein	0.11667	0.09453	0.21543	0.08562

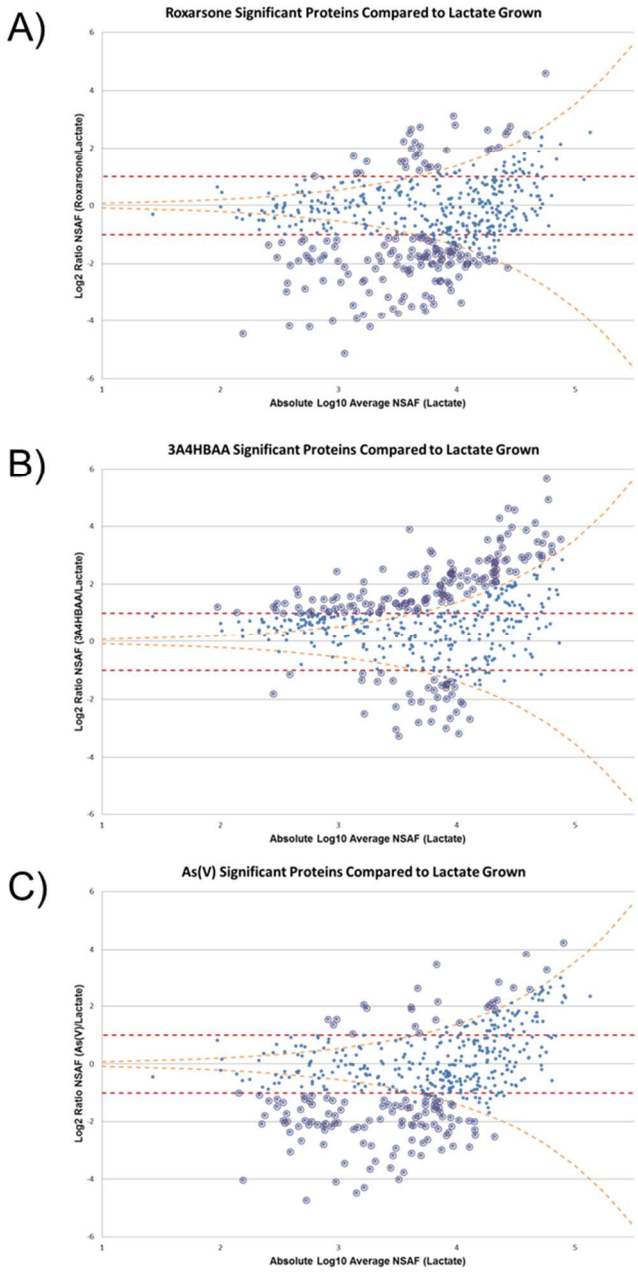


Figure 4: The relative abundances of proteins using normalized spectral abundance factor of common proteins identified in the three experimental condition (i.e. roxarsone with lactate, 3A4HBAA with lactate, and arsenate with lactate) and sodium lactate grown cells are compared using Power Law Global Error Model PLGEM plots to show statistically significant proteins. Proteins above the upper red line demonstrate proteins with >2 fold increase. Proteins below the lower red line show a <0.5 fold decrease when compared to the control. The orange dashed lines separate P-values of 0.05. A) Proteins identified in roxarsone with lactate compared with sodium lactate grown cells, B) proteins identified in 3A4HBAA with lactate compared with sodium lactate grown cells, C) Proteins identified in arsenate with lactate compared with sodium

lactate grown cells.

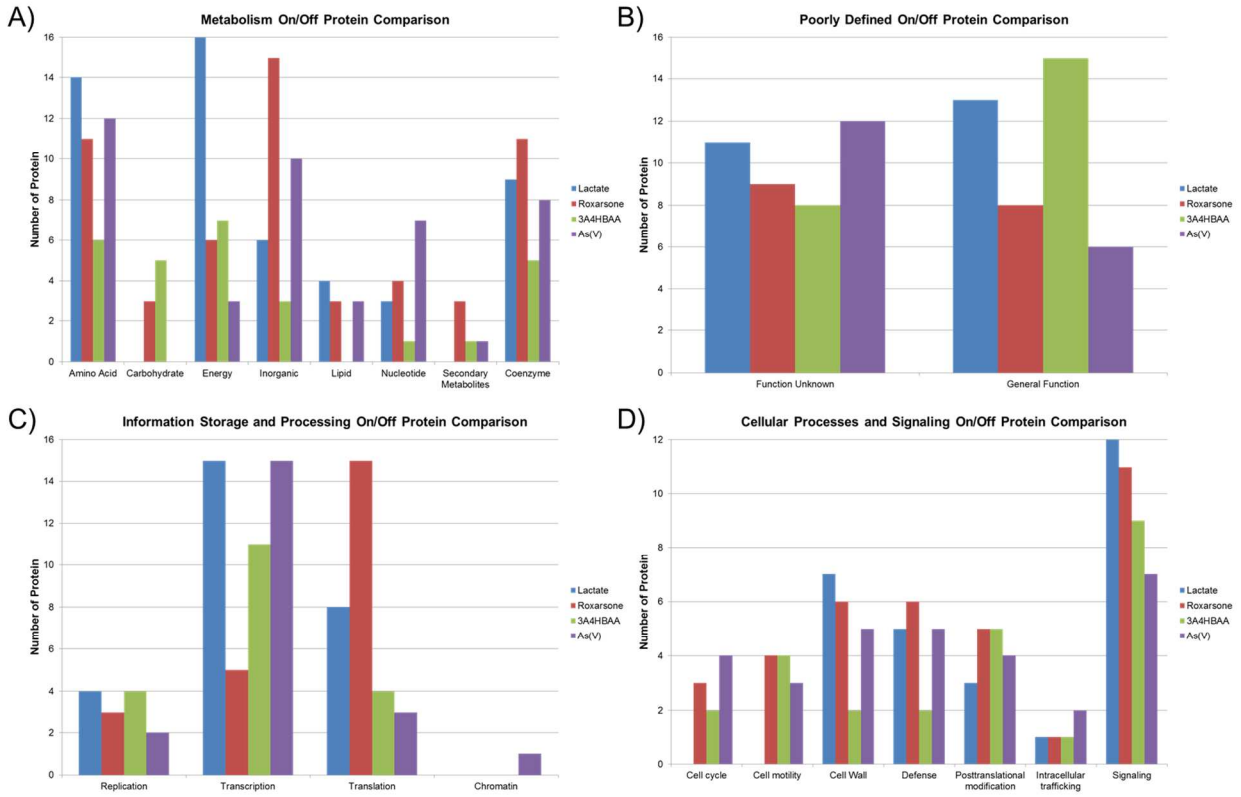


Figure 5: Comparison of on and off proteins in sodium lactate, roxarsone with lactate, and 3A4HBAA with lactate, and arsenate with lactate grown cells; they are categorized according to COG superclasses and subclasses. A) Metabolism proteins, B) poorly defined proteins, C) information storage proteins, D) cellular processes and signaling proteins.

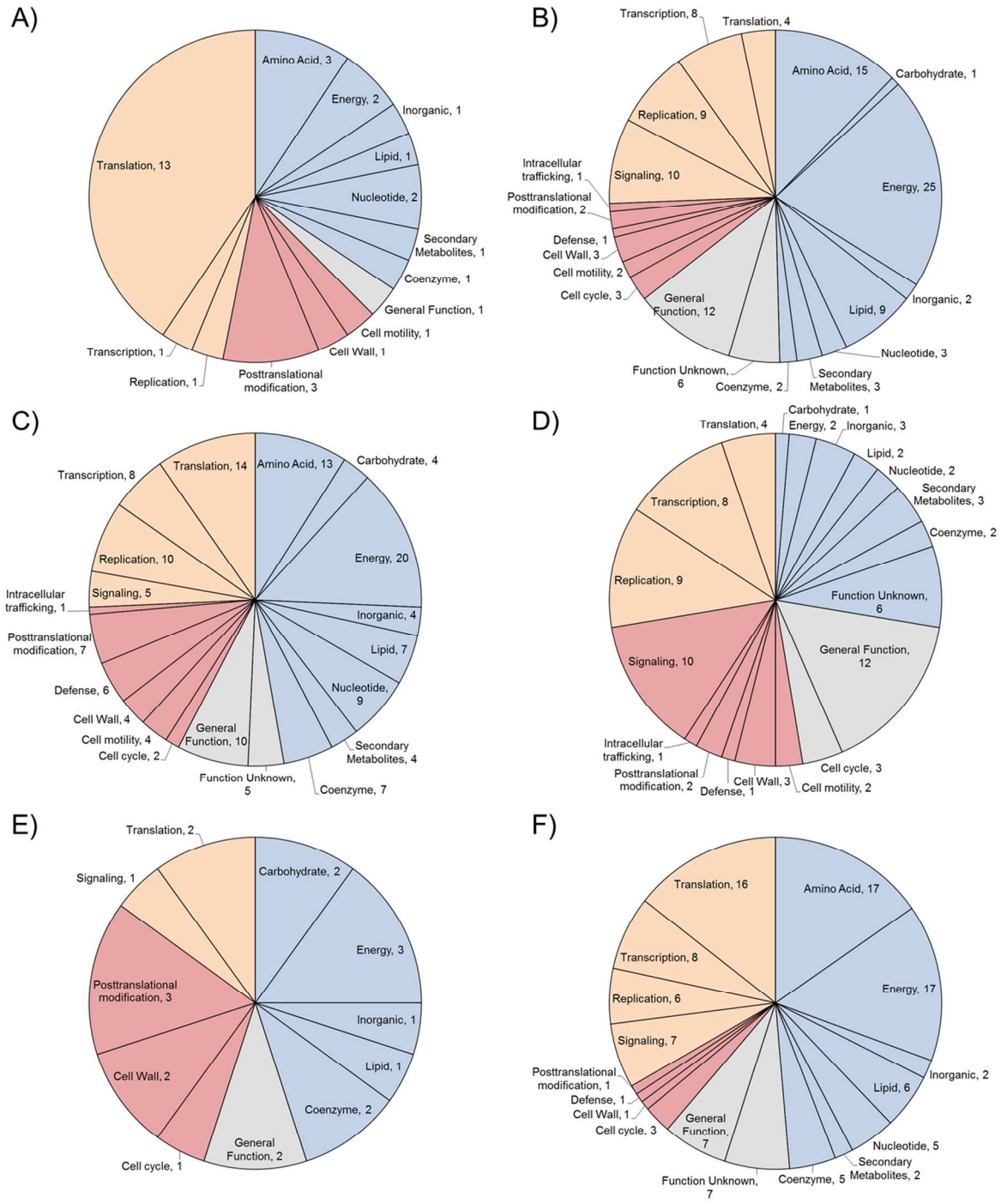
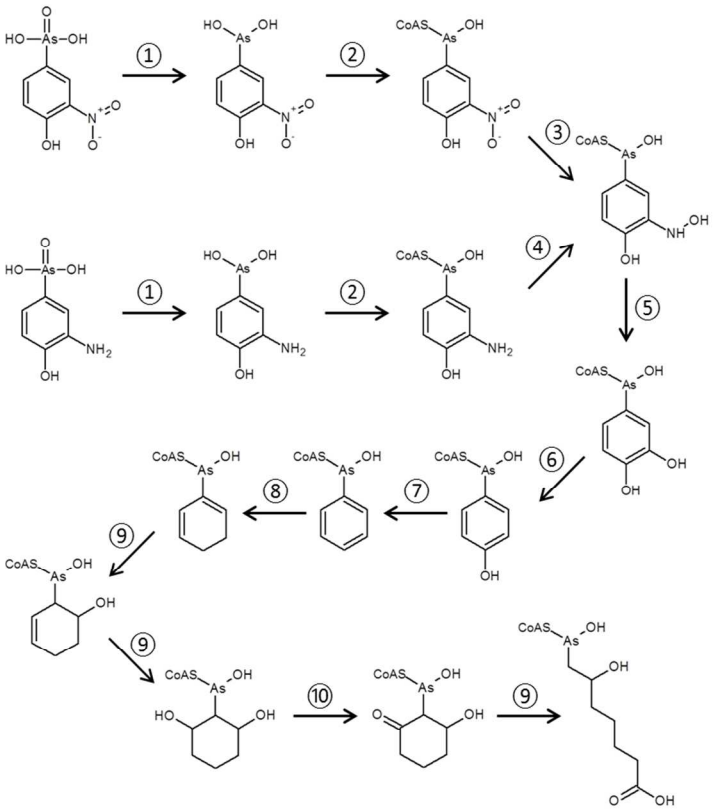


Figure 6: Up/down regulation of proteins divided into their COG superclasses and subclasses compared against lactate grown cells. A) Up, B) down regulated proteins in roxarsone grown cells; C) up, D) down regulated proteins in 3A4HBAA grown cells; E) up, F) down regulated proteins in arsenate grown cells.



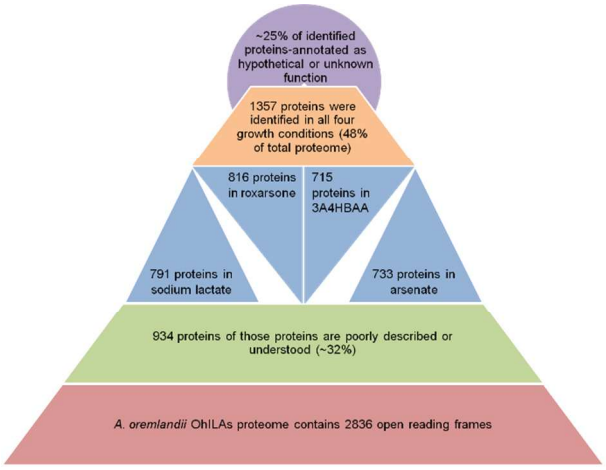
#	Locus Tag	Protein Name
1	Clos_1096	Molybdopterin oxidoreductase
2	Clos_0623	Succinyl-CoA ligase [ADP-forming] subunit beta
	Clos_0624	Succinyl-CoA ligase [ADP-forming] subunit alpha
3	Clos_1321	UBA/THIF-type NAD/FAD binding protein
4	Clos_0359	NADH:flavin oxidoreductase/NADH oxidase
5	Clos_2493	FAD-dependent pyridine nucleotide-disulphide oxidoreductase
6	Clos_0794	Iron-containing alcohol dehydrogenase
	Clos_2306	Iron-containing alcohol dehydrogenase
7	Clos_0264	Aldehyde oxidase and xanthine dehydrogenase molybdopterin binding
	Clos_0377	Aldehyde oxidase and xanthine dehydrogenase a/b hammerhead
	Clos_0379	Adenine deaminase
	Clos_0879	Aldehyde oxidase and xanthine dehydrogenase molybdopterin binding
8	Clos_2150	Putative CoA-substrate-specific enzyme activase
9	Clos_1455	3-oxoacyl-(Acyl-carrier-protein) reductase
10	Clos_0091	Acyl-CoA dehydrogenase domain protein
	Clos_0636	Acyl-CoA dehydrogenase domain protein

Figure 7: Proposed microbial degradation pathway for roxarsone and 3A4HBAA labeled by their locus tag numbers. Metal containing proteins with molybdenum only (blue), iron only (red), and molybdenum and iron (green) are highlighted. The proposed linear fatty acid is the arsenic derivatized to 6,8-dihydroxyoctanoic acid.

References.

1. K. E. Nachman, G. Raber, K. A. Francesconi, A. Navas-Acien and D. C. Love, *Sci. Total Environ.*, 2012, **417-418**, 183-188.
2. K. E. Nachman, J. P. Graham, L. B. Price and E. K. Silbergeld, *Environ. Health Perspect.*, 2005, **113**, 1123-1124.
3. E. K. Silbergeld and K. Nachman, *Ann. N. Y. Acad. Sci.*, 2008, **1140**, 346-357.
4. K. E. Nachman, T. S. Smith and R. P. Martin, *Science*, 2014, **343**, 136.
5. H. D. Chapman and Z. B. Johnson, *Poult. Sci.*, 2002, **81**, 356-364.
6. P. Basu, R. N. Ghosh, L. E. Grove, L. Klei and A. Barchowsky, *Environ. Health Perspect.*, 2008, **116**, 520-523.
7. J. L. Morrison, *J. Agr. Food Chem.*, 1969, **17**, 1288-1290.
8. B. L. Brown, A. D. Slaughter and M. E. Schreiber, *Appl. Geochem.*, 2004, **20**, 123-133.
9. L. X. Huang, L. X. Yao, Z. H. He, C. M. Zhou, G. L. Li, B. M. Yang and Y. F. Li, *Food Addit. Contam., Part A*, 2013, **30**, 1546-1555.
10. D. Harvey, March 29, 2011.
11. K. E. Nachman, J. P. Graham, L. B. Price and E. K. Silbergeld, *Environ. Health Perspect.*, 2005, **113**, 1123-1124.
12. J. R. Garbarino, A. J. Bednar, D. W. Rutherford, R. S. Beyer and R. L. Wershaw, *Environ. Sci. Technol.*, 2003, **37**, 1509-1514.
13. B. P. Jackson, P. M. Bertsch, M. L. Cabrera, J. J. Camberato, J. C. Seaman and C. W. Wood, *J. Environ. Qual.*, 2003, **32**, 535-540.
14. R. O'Connor, M. O'Connor, K. Irgolic, J. Sabrsula, H. Guerleyuek, R. Brunette, C. Howard, J. Garcia, J. Brien, J. Brien and J. Brien, *Environ. Forensics*, 2005, **6**, 83-89.
15. Y. Arai, A. Lanzirotti, S. Sutton, J. A. Davis and D. L. Sparks, *Environ. Sci. Technol.*, 2003, **37**, 4083-4090.
16. D. W. Rutherford, A. J. Bednar, J. R. Garbarino, R. Needham, K. W. Staver and R. L. Wershaw, *Environ. Sci. Technol.*, 2003, **37**, 1515-1520.
17. I. Cortinas, J. A. Field, M. Kopplin, J. R. Garbarino, A. J. Gandolfi and R. Sierra-Alvarez, *Environ. Sci. Technol.*, 2006, **40**, 2951-2957.
18. J. F. Stolz, E. Perera, B. Kilonzo, B. Kail, B. Crable, E. Fisher, M. Ranganathan, L. Wormer and P. Basu, *Environ. Sci. Technol.*, 2007, **41**, 818-823.
19. E. Fisher, A. M. Dawson, G. Polshyna, J. Lisak, B. Crable, E. Perera, M. Ranganathan, M. Thangavelu, P. Basu and J. F. Stolz, *Ann. N. Y. Acad. Sci.*, 2008, **1125**, 230-241.
20. Y. Ishihama, H. Katayama and N. Asakawa, *Anal. Biochem.*, 2000, **287**, 45-54.
21. C. S. R. Kadiyala, S. E. Tomechko and M. Miyagi, *PLoS One*, 2010, **5**, e15332.
22. E. Piruzian, S. Bruskin, A. Ishkin, R. Abdeev, S. Moshkovskii, S. Melnik, Y. Nikolsky and T. Nikolskaya, *BMC Syst. Biol.*, 2010, **4**, No pp. given.
23. J. Grossmann, B. Roschitzki, C. Panse, C. Fortes, S. Barkow-Oesterreicher, D. Rutishauser and R. Schlapbach, *J. Proteomics*, 2010, **73**, 1740-1746.
24. G. Maccarrone, I. Birg, E. Malisch, M. C. Rosenhagen, C. Ditzen, J. A. Chakel, F. Mandel, A. Reimann, C.-C. Doertbudak, K. Haegler, F. Holsboer and C. W. Turck, *Clin. Proteomics*, 2004, **1**, 333-364.
25. H. Choi, D. Fermin and A. I. Nesvizhskii, *Mol. Cell. Proteomics*, 2008, **7**, 2373-2385.
26. M. L. Fournier, A. Paulson, N. Pavelka, A. L. Mosley, K. Gaudenz, W. D. Bradford, E. Glynn, H. Li, M. E. Sardiu, B. Fleharty, C. Seidel, L. Florens and M. P. Washburn, *Mol. Cell. Proteomics*, 2010, **9**, 271-284.
27. N. Pavelka, M. L. Fournier, S. K. Swanson, M. Pelizzola, P. Ricciardi-Castagnoli, L. Florens and M. P. Washburn, *Mol. Cell. Proteomics*, 2008, **7**, 631-644.
28. N. Pavelka, M. Pelizzola, C. Vizzardelli, M. Capozzoli, A. Splendiani, F. G. Ranucci and P. Ricciardi-Castagnoli, *BMC Bioinf.*, 2004, **5**, DOI: 10.1186/1471-2105-5-203.

29. P. Chovanec, J. F. Stolz and P. Basu, *Metallomics*, 2010, **2**, 133-139.
30. P. Chovanec, P. Basu and J. F. Stolz, in *Microbial Metabolism of Metal and Metalloids: Advances and Applications*, eds. S. J.F. and R. S. Oremland, ASM Press, 2011, pp. 247-259
31. M. Yoshinaga and B. P. Rosen, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 7701-7706.



Insight into the organoarsenic metabolism of *Alkaliphilus oremlandii* OhILAs by comprehensive proteomic analysis.