Batch cultures of Candidatus Pelagibacter ubique were grown under iron-, organosulfur-, and nitrogen-limiting conditions to understand how this ubiquitous marine bacterium responds to and interacts with environments where growth is limited by the availability of these nutrients. Global gene expression was monitored using microarrays and quantitative mass spectrometry to observe both transcriptional and post-transcriptional responses to nutrient limitation. Iron- and nitrogen-limited cultures were characterized by increased transcription and translation of transporters involved in acquisition of the limiting nutrient, whereas organosulfur-limited cultures were not. Methionine synthesis genes downstream of S-adenosyl methionine riboswitches were up-regulated in mRNA and protein during organosulfur-replete stationary phase. Comparative genomics also revealed Ca. Pelagibacter to be the only genus among the free-living Alphaproteobacteria to lack a PII-mediated nitrogen regulatory pathway – a pathway which may be complemented in Ca. P. ubique by putative riboswitches and a citric acid cycle able to bypass the glutamate precursor 2-oxoglutarate. Overall, the results of this study provide insight into the regulatory and metabolic processes of this ecologically significant organism, and enable better interpretation of meta-transcriptomic and meta-proteomic surveys by identifying *sfuC* and *amtB* as likely biomarkers for iron and nitrogen limitation, respectively, in natural Ca. P. ubique populations.
The Proteomic and Transcriptomic Responses to Iron, Sulfur, and Nitrogen Limitation in the Abundant Marine Bacterium Candidatus Pelagibacter ubique

by
Daniel P. Smith

A DISSERTATION
submitted to
Oregon State University

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Presented November 29, 2011
Commencement June 2012

APPROVED:

______________________________________________________________
Major Professor, representing Molecular and Cellular Biology

______________________________________________________________
Director of the Molecular and Cellular Biology Program

______________________________________________________________
Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

______________________________________________________________
Daniel P. Smith, Author
ACKNOWLEDGEMENTS

The author expresses sincere appreciation to everyone who contributed to the success of this dissertation. Analysis of this organism in axenic cultures was made possible by Mike Rappé, who first isolated the Candidatus Pelagibacter ubique strain used in this study. The ability to limit these cells for a specific nutrient was largely enabled by Paul Carini’s tireless efforts to adapt Ca. P. ubique to artificial seawater. Nutrient mixes progressively refined by Kevin Vergin, Josh Kitner, Jim Tripp, Uli Stingl, Mike Schwalbach, and Paul Carini were essential for attaining the minimum required culture densities. Molecular techniques for preparing messenger RNA samples were patiently imparted to me by Laura Steindler, while the loading and running of microarray chips was expertly handled by Anne-Marie Girard. Sarah Sowell graciously leant her time to bring me up to speed in the proteomics protocols she developed and optimized for the collection of cell pellets from low density cultures. The proteins present in cell pellets were quantitative measured with mass spectrometry by Angela Norbeck, Kristin Burnum, Therese Clauss, Mary Lipton, and Carrie Nicora at Pacific Northwest National Laboratories under the direction of Richard Dick. The majority of data analyses were conducted in Perl – a language that I thank Adam Gustafson for introducing me to. Interpretation of experimental results was aided to a great degree by Jim Tripp’s and Paul Carini’s expansive and detailed understanding of Ca. P. ubique’s metabolic networks, and who are more generous with their time than anyone I know. Research activities were funded by the Gordan and Betty Moore Foundation and logistically supported by Kevin Vergin, Amy Carter, Gail Millimaki, Mary Fulton, Sally Tatala, and Cindy Fisher.

I would like to especially thank Steve Giovannoni. As my major professor, he provided invaluable direction on how to identify and investigate globally relevant scientific questions. Perhaps more importantly, despite numerous setbacks throughout my experiments, his confidence in me never wavered, nor did his hands-off approach for allowing me to learn to independently complete research projects.
The decision to undertake a PhD program in molecular & cellular biology was shaped in large part by two academic mentors. I consider myself fortunate that my chemistry teacher at Chancellor High School was Theresa Pugh, as her genuine passion for molecular interactions not only reversed the preconceived misgivings I had about the subject, but additionally sparked my own interest and prompted me to pursue molecular biology throughout my undergraduate and graduate studies. The integration of molecular biology and computer science into a manageable undergraduate program was made possible by my bioresource research advisor Wanda Crannell who enthusiastically dedicated herself to adapting each of her students’ programs of studies to best realize their individual academic goals. Without the influence of these two women, it is unlikely that my education would have culminated in a graduate degree such as the one embodied by this doctoral dissertation.

I am also grateful to my parents, Sue and Jerry Smith, and my sister, Whitney Smith, for the substantial role they have played in my life and the constant encouragement they have given me throughout my time as a student at Oregon State University. Since my youth, my parents have gone out of their way to provide me with a variety of extracurricular enrichment activities, and instilled in me a sense of value in my classwork. I cannot thank them enough for their financial support of my undergraduate education, without which I may have elected to forego post-secondary education altogether.
CONTRIBUTION OF AUTHORS

Dr. Stephen J. Giovannoni assisted with study design, data interpretation, and manuscript preparation for chapters 2, 3, and 4. Joshua B. Kitner assisted with culturing work in chapter 2. Microarray samples in chapter two were prepared by Michael S. Schwalbach and Laura Steindler. Paul Carini provided the artificial seawater (ASW) recipe and ASW-adapted strains used in chapters 3 and 4, and assisted with deducing the metabolic pathway proposed in chapter 3. J. Cameron Thrash provided ortholog databases and assistance with comparative genomics in chapter 4. Michael S. Schwalbach, Laura Steindler, Paul Carini, and J. Cameron Thrash additionally contributed to editing and revising manuscripts for the studies in which they were involved. Proteomic work in all chapters was done by Angela D. Norbeck, Kristin E. Burnum, Therese R. Clauss, Mary S. Lipton, Carrie D. Nicora, and Richard D. Smith at Pacific Northwest National Laboratories.
# TABLE OF CONTENTS

Chapter 1. General Introduction.................................................................................................1

**MARINE BACTERIA'S INFLUENCE ON BIOGEOCHEMISTRY** ........................................1

**BACTERIOPLANKTON RESPOND TO ENVIRONMENTAL CHANGE** .............................. 3

**THE SAR11 CLADE OF ABUNDANT OCEAN BACTERIA** ............................................. 5

Chapter 2. Transcriptional and translational regulatory responses to iron limitation in the globally distributed marine bacterium *Candidatus Pelagibacter ubique* ..................................................................................................................10

**ABSTRACT** .......................................................................................................................... 11

**INTRODUCTION** .................................................................................................................. 12

**RESULTS** ............................................................................................................................ 15

Reaction to the Siderophore ..................................................................................................... 15

Microarrays .................................................................................................................................. 15

Proteomics ................................................................................................................................... 16

Comparing Changes in mRNA and Protein Abundances ......................................................... 17

**DISCUSSION** ......................................................................................................................... 17

Upregulation of *sfuC* During Iron Limitation .......................................................................... 17

Transcriptome Distinct from Proteome ................................................................................... 18

Cold-shock Proteins Correlated with Iron Stress .................................................................... 18

Summary ....................................................................................................................................... 19

**EXPERIMENTAL PROCEDURES** ................................................................................. 21

Growth Media & Harvesting ....................................................................................................... 21

Messenger RNA Preparation ....................................................................................................... 21

Microarray Clustering ................................................................................................................ 22

Global TFE Protein Preparation ............................................................................................... 23

Capillary LC-MS Analysis .......................................................................................................... 24

Quantitative Proteomics ............................................................................................................. 25

**ACKNOWLEDGEMENTS** ................................................................................................. 27

**TABLES** ............................................................................................................................... 27

**FIGURES** ............................................................................................................................. 29
Chapter 3. Proteomic and transcriptomic analysis of *Candidatus* Pelagibacter ubique describes the first PII-independent response to nitrogen limitation in a free-living alphaproteobacterium ................................33

**ABSTRACT** ................................................................................................................ 34

**INTRODUCTION** ......................................................................................................... 35

**RESULTS AND DISCUSSION** .................................................................................. 38

Nitrogen Assimilation Genes in Ca. P. ubique ................................................................. 38

Absence of the PII Regulatory Network ........................................................................ 38

Regulatory Proteins NtrY/NtrX ..................................................................................... 40

Ammonium Transport and Assimilation ....................................................................... 40

Organic Nitrogen Transport and Assimilation ............................................................... 42

Proposed Regulatory Model ........................................................................................ 43

**EXPERIMENTAL PROCEDURES** ............................................................................ 45

Phylogenetic Analysis/PII Orthologs ........................................................................... 45

Searching Metagenomic Sequences for SAR11 PII Genes ........................................... 45

Growth Media and Harvesting ....................................................................................... 46

Messenger RNA Preparation ........................................................................................ 47

RNA Labeling and Microarray Processing ................................................................... 47

Capillary LC-MS Analysis ............................................................................................ 48

Mass Spectrometry Data Analysis ................................................................................ 50

**TABLES** ..................................................................................................................... 52

**FIGURES** .................................................................................................................... 55

Chapter 4. Regulatory Responses of *Candidatus* Pelagibacter ubique to Sulfur Limitation ................................................................................................................. 59

**ABSTRACT** ................................................................................................................ 60

**INTRODUCTION** ......................................................................................................... 61

**RESULTS** ...................................................................................................................... 63

Correlation between mRNA and Protein ....................................................................... 64

Stationary Phase Differences ......................................................................................... 64
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>SeaWiFS satellite imagery illustrating the global variability in oceanic primary production.</td>
<td>9</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Growth of <em>Candidatus</em> Pelagibacter ubique cells was arrested by iron-sequestering siderophores.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Genes transcribed during iron limitation were different from stationary phase genes.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Protein abundances were largely decoupled from transcript abundances.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Translation of <em>Ca.</em> Pelagibacter ubique’s cold shock and iron-binding genes are influenced by iron availability.</td>
<td>32</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Distribution of nitrogen assimilation genes in alphaproteobacteria.</td>
<td>55</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Proteins for assimilating and metabolizing nitrogenous compounds were up-regulated under nitrogen-limiting conditions.</td>
<td>56</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Putative enzymatic functions of the highly up-regulated protein SAR11_1304.</td>
<td>57</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Growth curves of nitrogen replete (red) and nitrogen limited (blue) batch cultures of <em>Ca.</em> P. ubique.</td>
<td>58</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Effect of stationary phase on methionine metabolism.</td>
<td>87</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Growth curves for batch cultures of <em>Ca.</em> P. ubique grown with either limiting (red) or non-limiting (blue) concentrations of DMSP as the sole sulfur source.</td>
<td>88</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Protein abundance does not depend on mRNA abundance.</td>
<td>89</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Genes initially induced as sulfur became limiting.</td>
<td>90</td>
</tr>
<tr>
<td>Figure 14</td>
<td>The bhmT gene in <em>Candidatus</em> Pelagibacter ubique shows evidence of horizontal gene transfer.</td>
<td>91</td>
</tr>
</tbody>
</table>
Table 1. All 23 *Ca. Pelagibacter ubique* mRNA transcripts that were at least 50 percent more abundant in the iron-limited cultures compared to the control cultures, 24 hours after addition of an iron-chelator.

Table 2. All 17 *Ca. Pelagibacter ubique* proteins that were at least 50 percent more abundant in the iron-limited cultures compared to the iron replete cultures, two and 12 days after addition of an iron chelator.

Table 3: Transcripts and proteins that were more abundant during nitrogen-limited stationary phase.

Table 4: Selected genes related to nitrogen assimilation that either decreased or remained constant in response to nitrogen limitation.

Table 5: Composition of nitrogen-limited artificial seawater media.

Table 6: Effect of sulfur limitation on stationary phase.

Table 7: Post-log phase upregulated mRNAs and proteins.

Table 8: Composition of organosulfur-limited artificial seawater media.

Table 9: Phylogenetic distribution of genes related to sulfur metabolism.
CHAPTER 1. GENERAL INTRODUCTION

MARINE BACTERIA’S INFLUENCE ON BIOGEOCHEMISTRY

Earth’s vast oceans play a crucial role in enabling life to thrive on this planet. The natural ability of water to dissolve the building blocks of life, induce the formation of lipid bilayers, and maintain a temperate environment is believed to have fostered the development of the first single-cell organisms. Today, every cubic millimeter of seawater is colonized by millions of cells that serve as primary producers – microorganisms that collectively assimilate inorganic molecules such as H₂O, CO₂, N₂, SO₄, and PO₄ into biomass on the scale of teragrams per year, removing a significant fraction of greenhouse gases from the atmosphere in the process. These bacterioplankton are the foundation of the marine food web that produces economically valuable seafood including fish, shrimp, and crustaceans. Therefore, the ability of bacterioplankton populations to adapt to deficiencies in specific nutrient pools influences overall ocean productivity and global rates of carbon sequestration.

Primary production is commonly quantified by measuring the green tint to seawater caused by light-harvesting chlorophyll pigments in the chloroplasts of active photosynthetic microorganisms. The Sea-viewing Wide Field-of-view Sensor (SeaWiFS) satellite in low Earth orbit monitors global ocean productivity in this way [1,2], producing maps such as Figure 1 that illustrate the unevenness of productivity across the world’s oceans. Subtropical gyres such as the central South Pacific are characterized by low productivity, whereas coastal regions are highly productive.

This disparity in productivity is caused by unequal distribution of terrestrially-derived essential nutrients. Phosphorous (P) and iron (Fe) are released from mineral deposits through weathering and are transported into marine systems via rivers and dust particles [3–5], creating high nutrient (eutrophic) conditions in coastal regions and low nutrient (oligotrophic) conditions in environments that are geographically distant from
landmasses. The link between nutrient availability and productivity has been experimentally confirmed by studies which observed significant increases in bacterioplankton productivity after fertilizing oligotrophic waters with P [6–8] or Fe [9–11], with the limiting nutrient found to change over time and between sites [12–17].

Nitrogen (N) additions to oligotrophic waters also stimulated increases in productivity [18–27]. These observations may seem counter-intuitive given that nitrogen-fixing diazotrophs are present in bacterioplankton populations and (in lakes) bloom under nitrogen limited conditions [14], thereby replenishing pools of reduced N. However, picomolar concentrations of iron in marine oligotrophic environments [28–30] are thought to severely inhibit nitrogenase, which requires iron as a cofactor [30–33]. This theory is bolstered by studies which found a correlation between rates of nitrogen fixation and ambient iron concentrations [34–36]. Recent measurements of diazotroph activity in oligotrophic regions of the Atlantic Ocean found that the cyanobacterium Trichodesmium [37] accounted for the majority of N\textsubscript{2} fixation, with a maximum observed rate of \(~200\ \mu\text{mol N m}^{-2}\text{ d}^{-1}\) [29] – about equal to the input of NO\textsubscript{3} due to upwelling [38]. While a significant source of N, these inputs are insufficient for bacterioplankton demands [33,39,40], as evidenced by the low concentrations of unincorporated NO\textsubscript{X} (5 nmol L\textsuperscript{-1}) and NH\textsubscript{4} (17 nmol L\textsuperscript{-1}) at the same site [29]. Therefore, nitrogen fixation by diazotrophs in oligotrophic environments may be limited by either P or Fe, and lead to these ecosystems to remain N limited.

The biogeochemistry of oceans is largely controlled by metabolic activities occurring in microbial communities [41]. Depending on nutrient availability and other environmental conditions, marine bacteria may bloom into toxic red tides, produce anoxic conditions lethal to fish and other marine life, or alternately release or absorb compounds from the atmosphere that have been implicated in global climate change (e.g. carbon dioxide, methane, and dimethyl sulfide). Sewage and agricultural fertilizer that makes its way into the ocean has increased the availability of N and P in coastal
environments, and together with industrial processes, has increased the N and P cycles by 100% and 400%, respectively, with significant effects on biological diversity and ecosystem services [42–47]. However, to fully understand how these global changes impact the flow of metabolites through marine ecosystems, we must first determine how the metabolism of bacterial communities changes in response to nutrient conditions.

**Bacterioplankton Respond to Environmental Change**

Characterization of natural bacterioplankton assemblages was first enabled by techniques pioneered in 1985 for directly cloning 16S rDNA sequences from the environment [48,49]. These clone libraries could be sequenced or analyzed though restriction fragment length polymorphisms (RFLP) to quantify the abundance of specific phylogenetic groups in the environment. As advancements in high-throughput sequencing decreased the cost of sequencing by several orders of magnitude, this strategy shifted from sequencing 16S rDNA to sequencing random DNA fragments derived from environmental assemblages [50,51]. “Metagenomic” sequencing of bacterioplankton communities has revealed metabolic potentials correlated with environmental parameters and a high degree of genetic diversity among members of the same species [52–57].

While metagenomics allows an accounting of the major protein families present in a population’s genomes [58], revealing which genes are being transcribed requires the application of metatranscriptomics – the sequencing of RNA present in microbial communities [59–63]. Metatranscriptomic studies have been very successful at identifying nutrient transporters expressed in particular bacterioplankton communities. For instance, transcripts encoding N acquisition genes were abundant in the oligotrophic Pacific Ocean and off the Georgia coast [61,64], and genes for both P and N were abundant in water north of Hawaii [63]. Interestingly, two assemblages were found to encode hypothetical proteins on the most highly expressed transcripts [61,65].
and highlights the fact that meaningful interpretation of –omics results is dependent on experiments which have first characterized the role of proteins in the cell.

Translation of mRNA transcripts into protein products is a carefully regulated process, and as a result, the abundance of proteins in a cell is often unrelated to the abundance of the corresponding transcripts [66–71]. Therefore, since proteins are the effector of most cellular processes, directly analyzing these proteins in an environmental population (metaproteomics) is arguably the best method for investigating changes in cellular physiology in response to environmental conditions. Four studies have explored gene expression in ocean surface bacterioplankton populations using metaproteomic techniques. Transporters were the most abundant proteins in all surveys except for Chesapeake Bay [72], in which proteins of unknown function were most dominant. Among the other sites, protein expression was consistent with P-limitation in the oligotrophic Sargasso Sea [73], N-limitation off the Oregon coast [74] and N-limitation in the South Atlantic open ocean [75].

The high abundance of nutrient transport proteins in metaproteomes [73–75] provides an excellent bioindicator of the limiting nutrient in that environment, but drowns out the signal for the lower-concentration cytoplasmic enzymes that control the fate of nutrients taken into the cell under these conditions. Given that low complexity communities such as those found in acid mine drainages can be sampled to saturation using current technologies [76–80], it is quite possible that future mass spectrometers with greater dynamic range and higher sensitivity will enable comparable analyses of complex bacterioplankton populations in the open ocean.

However, in order to gain insight into biogeochemical processes at work in the oceans today, this dissertation will examine the transcriptome- and proteome-wide responses of a dominant marine bacterioplankton to various nutrient-limiting conditions in axenic batch cultures. The results of these experiments will identify transcripts and proteins that are characteristic biomarkers of particular nutrient limitations in vivo, and
verify if expression of the abundant transporters observed in nature are indeed induced by low concentrations of the nutrient they are annotated as transporting (e.g. if nitrogen-limited cultures up-regulate the N transporter observed by Sowell [74]). Perhaps of broader importance, these experiments will identify metabolic pathways that are active in nutrient-limited cultures and that therefore may be active on an enormous scale in nutrient-limited ocean waters.

**THE SAR11 CLADE OF ABUNDANT OCEAN BACTERIA**

The SAR11 clade of alphaproteobacteria commonly accounts for one-third of marine microbial communities, making it the most abundant free-living aerobic heterotroph in the ocean and a significant influence on global nutrient cycling. This large fraction of bacterioplankton is not culturable using traditional plate-based techniques and was instead discovered in a large-scale phylogenetic profiling of environmental rRNA amplicons from the Sargasso Sea in 1990 [49]. Subsequent community composition surveys found members of this clade to represent a large fraction of bacterioplankton throughout the world’s oceans [57,74,81–88], leading to extrapolated estimates of the global SAR11 population at $2.4 \times 10^{28}$ cells [86]. High-throughput sequencing of environmental SAR11 rRNA has resolved this abundant clade into several distinct subclades which often correspond to ecotypes associated with specific seasons and depths [87,89,90]. A representative “Ia” ecotype adapted to nutrient-scarce summer surface waters was cultured in 2002 [91] and sequenced in 2005 [92]. This isolate, *Candidatus* Pelagibacter ubique HTCC1062, was found to exhibit a number of extreme characteristics that appear to confer an evolutionary advantage in nutrient limited environments.

*Ca. P. ubique* is one of the smallest free-living microorganisms. Electron microscopy defined these cells as curved rods approximately 0.9 μm long, 0.25 μm wide, and enclosing a volume of 0.035 μm$^3$ [91]. Other abundant picoplankton competing in the same habitat have volumes of 0.1 μm$^3$ (*Prochlorococcus*) [93] and 0.8 μm$^3$ (*Synechococcus*) [94]. The smallest microorganism in an oligotrophic environments
has a competitive advantage due to its lower absolute nutrient requirements for growth and division, and because the increased surface-to-volume ratio of smaller cells offers faster transport rates relative to internal metabolite fluxes [27,95,96]. Cryo-electron tomography revealed that 25% of Ca. P. ubique’s volume is periplasmic space in actively growing cells and 35% in stationary phase cells [97]. The large periplasmic space and high proportion of high-affinity ABC transporters detected by mass spectrometry [73,98] are physiological attributes of cells configured to maximize nutrient acquisition [99,100]. Small cell size and a large transporter-rich periplasmic space make Ca. P. ubique very efficient at acquiring nutrients from its environment, but are not the only strategies employed by this bacterium to optimize growth.

In addition to being one of the smallest free-living cells, Ca. P. ubique also contains one of the smallest genomes among free-living bacteria: 1354 genes encoded in 1.3 Mbp [92]. It is hypothesized that scarcity of phosphate and nitrogen in oligotrophic environments selected for cells with highly streamlined genomes which require less of these elements for DNA replication [92,101]. This bacterium appears to further conserve nitrogen by encoding genes using 70% A-T base pairs, which require one less atom of nitrogen compared to a G-C base pair [102]. Though Ca. P. ubique’s small genome is due in part to small intergenic spacers (median = 3 nt) and very few paralogous genes, its size also reflects the surprising absence of genes involved in critical metabolic processes. Ca. P. ubique HTCC1062 is missing pathways for assimilatory sulfate reduction [103], glycolysis [55], and glycine/serine synthesis [104]. Although slightly less dependent on phosphate and nitrogen as a result of genome reduction, this species is much more dependent on ambient availability of organosulfur, glycine, and serine.

Sequencing of additional SAR11 genomes revealed widespread variation in genetic potential among isolates. For instance, the very recently diverged Ca. P. ubique strains HTCC1062 and HTCC1002 were isolated from the same Oregon coast water sample [91], yet the high affinity phosphate uptake system pstABCS present in HTCC1062’s
genome is absent from HTCC1002. A distantly related third strain from the Sargasso Sea, HTCC7211, lacks genes for metabolizing glucose which are present in the isolates from the Oregon coast [55]. Aligning metagenomic DNA fragments from the Sargasso Sea [53] to the HTCC1062 genome identified four genomic regions (~90 kbp) where gene content is highly variable [54]. These hypervariable regions are similar to genomic islands utilized by *Prochlorococcus* and other organisms to encode phosphate, iron, and ammonium transporters, nitrogenase, and other conditionally advantageous genes [105–109]. Thus, members of the SAR11 clade contain a common “core” genome essential for growth as well as a portion of the “flexible” genome that may provide a competitive advantage in e.g. nutrient limited environments.

Despite *Ca. P.*’s metabolic limitations, evidence is mounting that the SAR11 clade has a significant impact on a diverse set of dissolved organic carbon compounds. In the North Atlantic Ocean, SAR11 cells represent 30% of the microbial population yet accounted for 50% of amino acid uptake, 50% of glucose uptake, 30% of 3-dimethylsulphoniopropionate (DMSP) uptake, and 30% of protein uptake [110,111]. Assimilation of DMSP by SAR11 is of particular interest, as degradation of this algal osmolyte by other organisms releases climatically active dimethylsulfide into the atmosphere [112–122]. *Ca. P.*’s ability to assimilate DMSP, glucose, pyruvate, and amino acids into biomass and can derive energy via the oxidation of one-carbon/methylated compounds such as methanol, formaldehyde, glycine betaine, DMSP, methylamine, trimethylamine, and trimethylamine N-oxide [55,103,110,123]. A proteorhodopsin light-driven proton pump also enables *Ca. P.* to regenerate ATP during extended periods of carbon starvation [124], likely allowing these cells to maintain a state of readiness for exploiting brief pulses of nutrient availability.

Although *Ca. P.* is one of the smallest free-living cells, this clade’s combined metabolic activity accounts for a significant portion of the biochemical processes at work in the ocean. To understand the biogeochemistry of this vast ecosystem and how it is impacted by increased anthropological nutrient additions, it is reasonable to start
by examining how nutrients are transformed by this abundant marine bacterium. In
furtherance of this goal, this dissertation describes how Ca. P. ubique utilizes nutrients
under a variety of oligotrophic conditions: Fe-limitation (chapter 2), N-limitation
(chapter 3), and sulfur (S)-limitation (chapter 4). As discussed in detail in chapter 4, S
was chosen in place of P due to the intriguing absence of an assimilatory sulfate
reduction pathway in Ca. P. ubique, meaning that unlike most bacteria, these cells are
unable to assimilate S from the millimolar concentrations of SO$_4^-$ present in seawater.
Results from these experiments identified the primary transporters and metabolic
pathways that are active under each of these conditions, and also revealed previously
undescribed regulatory schemes that enable these cells to thrive with a minimal set of
genes.
Figure 1: SeaWiFS satellite imagery illustrating the global variability in oceanic primary production.

Green-, orange-, and red-shaded coastal regions indicate relatively high seawater concentrations of the light-harvesting chlorophyll pigment synthesized by phytoplankton. In nutrient-poor open ocean gyres, phytoplankton activity is significantly reduced.
CHAPTER 2. TRANSCRIPTIONAL AND TRANSLATIONAL REGULATORY
RESPONSES TO IRON LIMITATION IN THE GLOBALLY DISTRIBUTED
MARINE BACTERIUM CANDIDATUS PELAGIBACTER UBIQUE

Daniel P. Smith, Joshua B. Kitner, Angela D. Norbeck, Therese R. Clauss, Mary S.
Lipton, Michael S. Schwalbach, Laura Steindler, Carrie D. Nicora, Richard D. Smith,
& Stephen J. Giovannoni.

Citation:
Transcriptional and Translational Regulatory Responses to Iron Limitation in the
Globally Distributed Marine Bacterium Candidatus Pelagibacter ubique. PLoS ONE
5: e10487. doi:10.1371/journal.pone.0010487
**ABSTRACT**

Iron is recognized as an important micronutrient that limits microbial plankton productivity over vast regions of the oceans. We investigated the gene expression responses of *Candidatus* Pelagibacter ubique cultures to iron limitation in natural seawater media supplemented with a siderophore to chelate iron. Microarray data indicated transcription of the periplasmic iron binding protein *sfuC* increased by 16-fold, and iron transporter subunits, iron-sulfur center assembly genes, and the putative ferrooxidase ruberythrin transcripts increased to a lesser extent. Quantitative peptide mass spectrometry revealed that *sfuC* protein abundance increased 27-fold, despite an average decrease of 59% across the global proteome. Thus, we propose *sfuC* as a marker gene for indicating iron limitation in marine metatranscriptomic and metaproteomic ecological surveys. The marked proteome reduction was not directly correlated to changes in the transcriptome, implicating post-transcriptional regulatory mechanisms as modulators of protein expression. Two RNA-binding proteins, CspE and CspL, correlated well with iron availability, suggesting that they may contribute to the observed differences between the transcriptome and proteome. We propose a model in which the RNA-binding activity of CspE and CspL selectively enables protein synthesis of the iron acquisition protein SfuC during transient growth-limiting episodes of iron scarcity.
INTRODUCTION

The importance of iron as a nutrient in the oceans was first recognized by Martin [125] and later experiments verified that iron limits primary production over broad regions of the marine environment [4,9,10]. A variety of biological processes such as photosynthesis, N\textsubscript{2} fixation, methanogenesis, respiration, oxygen transport, gene regulation, and DNA synthesis all depend on iron-containing proteins [126]. In pelagic surface waters, planktonic communities must cope with iron concentrations that average just 70 picomolar [28]. The inhibitory effect that this has on growth was most clearly illustrated by a series of iron fertilization experiments in which iron was added to large swaths of the ocean, resulting in a marked increase in nutrient utilization [9–11].

Bacteria commonly have specialized systems for responding to iron limitation. Genes for iron uptake and utilization are primarily regulated by the Fur protein [127,128]. When complexed with Fe(II) cations, Fur binds the "Fur box" recognition sequence, which is made of several GATAAT hexamers [129–133]. In some bacteria, this single transcription factor can directly repress or activate more than 100 genes in response to iron scarcity [128]. Irr is a similar transcription factor that couples intracellular heme levels to expression of many different iron-related pathways [134–140]. Small RNAs [141–144] and mRNA-binding proteins [145,146] can also regulate nonessential iron-utilizing proteins at the post-transcriptional level by selectively targeting their transcripts for degradation. To improve their chances of encountering Fe(III), many bacteria secrete siderophores [147–149]. These chelating agents help dissolve the poorly soluble particles and sequester them in a form that is unusable by competing microorganisms. Due to the spontaneous reactivity of iron ions, cells often encapsulate these atoms inside containers made of ferritin proteins to better modulate redox reactions [150,151].

*Candidatus* Pelagibacter ubique was selected as an iron limitation model for two reasons. First, this alphaproteobacterium is regularly the most numerically abundant
microorganism in surveys of marine microbial diversity. Second, its proteome of just
1,354 genes is possibly the simplest of any free-living heterotrophic organism [92].
_Ca._ Pelagibacter ubique’s genome encodes Fur and Irr, but not ferritin or siderophore-
related proteins, raising questions about how or if this bacterium can cope with iron
stress. Investigating how this organism’s relatively small genetic repertoire produces
thriving populations in the variable ocean environment has been impeded by the lack
of a genetic system able to create knockouts or other genetic modifications. Thus,
oberving how the entire transcriptome or proteome changes in response to growth
conditions has become a primary approach for elucidating metabolic and regulatory
schemes [55,73,98]. A comparison of cultures in exponential and stationary phase did
not reveal a major remodeling of the proteome nor evidence of a global regulatory
mechanism [98], suggesting that this organism may continue to benefit from
temporary nutrient availability regardless of overall cellular activity. That study, along
with a follow-up using a metaproteomics approach on environmental samples [73],
found _Ca._ Pelagibacter ubique’s proteome to be consistently composed of an
unusually high proportion of transport-related proteins.

Arguably the most important characteristic of organisms is their ability to express the
right proteins in the right amounts at the right times. The interplay between stimuli,
sensors, and regulators precisely optimizes the combination of mRNA transcripts and
protein products present in the cell. Several known and putative transcriptional
regulators have been identified in the _Ca._ Pelagibacter ubique genome, as well as cis-
acting riboswitches capable of modulating mRNA translation based on the
concentration of particular metabolites. This method of decoupling production of
mRNA from protein synthesis has been found on glycine metabolism genes in _Ca._
Pelagibacter ubique [104] and sequence motif searches [152] found additional
candidate riboswitches for s-adenosyl methionine [153,154] and thiamine
pyrophosphate [154,155]. Meyer and colleagues also identified homologs to ribosomal
proteins capable of regulating their own translation, as well as regions in the genome
with riboswitch-like characteristics but lacking homologous annotated motifs. One of
these putative structural RNA regions is located immediately upstream of the *sfuA-C* operon, which encodes an iron-acquisition system [156,157]. Post-transcriptional regulation schemes allow the cell to conserve amino acids while still rapidly providing ephemeral enzymes. The success of these characteristics are evidenced by direct observation; *Ca. Pelagibacter ubique* is the most abundant heterotroph in the oceans, accounting for one-third of surface water bacteria [86,110] and consuming up to half of some dissolved organic matter compounds [111].

Combining transcriptomic and proteomic data offers a perspective on cellular activity that cannot be obtained from either method individually. Numerous studies have shown that changes in the transcriptome poorly correlate with changes to the proteome, except for very highly expressed genes [68–71]. Although much of the disparity between these two types of datasets has been attributed to measurement inaccuracy [158] and differences in protein degradation rates [159], some studies have revealed systematic post-transcriptional regulatory schemes. For instance, in the eukaryotic protozoan *Plasmodium falciparum*, mRNAs were often upregulated an entire life phase before the one in which the encoded protein was needed [160]. Additionally, translation in *Escherichia coli* was found to be partially regulated by mRNA secondary structure [161]. Therefore, it is evident that the transcriptome does not necessarily represent the current state of the proteome, but is rather a mixture of transcripts being actively translated and others that are standing by, awaiting activation by post-transcriptional regulatory mechanisms. This study integrates both transcriptomic and proteomic analyses in order to attain a more complete understanding of the cellular response to iron limitation in *Ca. Pelagibacter ubique*. The results strongly suggest that transcription and translation are not always tightly coupled in this bacterium.
RESULTS

Reaction to the Siderophore

Two iron-sequestering siderophores were tested on cultures of *Ca. Pelagibacter ubique* to determine their feasibility for creating iron-limiting conditions. Ferrichrome (Sigma #F8014) and deferoxamine mesylate salt (Sigma #D9533) were both found to arrest batch culture growth within 1/3 of a doubling – an inhibition which could be reversed by addition of iron (Figure 2A). Sufficient bioavailable iron was present in the natural seawater media collected from the Oregon coast to enable cultures to grow when supplemented with 10 nM siderophore, but not when supplemented with 100 nM siderophore.

Microarrays

Six 20 L carboys inoculated with *Ca. Pelagibacter ubique* were grown to near-maximum density, then randomly selected for treatment with either ferrichrome or ferrichrome plus excess iron (Figure 2B). To measure the amount of messenger RNA transcripts present in cells, mRNA from each carboy was hybridized to separate microarray chips containing probes for all *Ca. Pelagibacter ubique* genes. Microarray data was deposited in the NCBI GEO database under accession number GSE20962. Of the three time points where mRNA abundance was measured, the greatest difference in expression of known iron-related genes was observed 24 hours after the siderophore amendment. Table 1 lists the 23 transcripts that were expressed at least 50 percent higher in the iron-limited culture relative to the control. Two-thirds of the genes in this list come from two operons: the first containing iron-sulfur center assembly proteins including *sufA-E* and the second made up of iron uptake proteins such as *sfuA-C*. The four other genes with a known function are: ruberythrin, an iron-binding protein that is postulated to act as a ferroxidase for converting Fe(II) to Fe(III), *hslU* and *hslV* which together form a protease complex, and dimethylglycine dehydrogenase – an enzyme that is necessary for converting betaine to glycine.
A modified radial coordinate visualization plot (Figure 3) of the microarray data shows four distinct clusters of genes: exponential growth, stationary phase, early iron stress, and late iron stress. The early iron stress cluster is dominated by genes from three genomic loci: the \textit{sfu} iron uptake operon, the \textit{suf} iron-sulfur center assembly operon, and the functionally unclear loci SAR11_1157, SAR11_1158, SAR11_1163, and SAR11_1164. The late iron stress cluster also contains different genes that are located in or adjacent to the \textit{sfu} and \textit{suf} operons, but is better characterized by \textit{lexA}, \textit{recA}, and \textit{mucA} – three genes involved in the SOS response. The iron response regulators \textit{fur} and \textit{irr} cluster with stationary phase genes, indicating that the abundances of these two transcripts are more affected by stationary phase than by iron limitation.

\textbf{Proteomics}

Cellular protein fractions from each treatment were isolated and digested before being separated with liquid chromatography and injected into a tandem mass spectrometer. An Accurate Mass and Time Tag library, developed previously [98], was used to make quantitative comparisons of the abundance of individual peptides between samples. This dataset is available at http://omics.pnl.gov/. Of the 216 proteins detected with high certainty in this study, 18 were observed to be at least 50\% more abundant in the iron-limited cultures: four on day 18, and 17 on day 28 (Table 2). The proteins \textit{SfuC}, \textit{CspL}, and \textit{GroES} were higher in the iron-limited cultures at both timepoints. The iron-binding \textit{SfuC} is unique in that it was the only one of these 18 proteins to increase in both protein and mRNA abundance by at least 50\%. \textit{CspL} was originally annotated as a DNA-binding protein, however, similar proteins have been found to modulate the accessibility of mRNA binding sites by selectively melting secondary RNA structures [162]. The third protein, \textit{GroES}, forms a complex with \textit{GroEL} to mediate protein folding. Because the required \textit{GroEL} subunit was much less abundant in iron-limited cultures, and since the three largest \textit{GroES} peptide spectra (out of 9) were less pronounced in the iron-limited cultures, \textit{GroES} may be a false positive. Mass
spectrometry measurements did not reveal a significant change in Fur or Irr abundance between treatments or timepoints.

Iron-limitation had a marked impact on the overall proteome. Two days after addition of an iron-chelator, 181 of the 216 proteins were significantly (P ≤ 0.05) less abundant in the iron-limited cultures relative to the control cultures. Using the same criteria, only 32 of the 216 proteins were found to significantly decrease in the control cultures between days 18 and 28 as Ca. Pelagibacter ubique cells entered stationary phase due to an unknown, non-iron, limitation.

Comparing Changes in mRNA and Protein Abundances
Aside from the highly expressed iron-binding protein SfuC, the abundances of individual proteins appeared to be independent of the amount of mRNA encoding them (Figure 4).

**DISCUSSION**
We are studying keystone microbial plankton species such as Ca. Pelagibacter ubique in culture to provide a basis for interpreting data emerging from molecular ecology studies. In an era of rapid environmental change, metagenomics, and allied technologies such as metaproteomics and metatranscriptomics, are being used to monitor the structure and health of natural ecosystems and to identify ecological processes that impact biogeochemistry. Interpretations of these data depend on understanding how complex cellular systems respond to environmental factors. We focused on a microorganism, Ca. Pelagibacter ubique, that produces the largest signal in most environmental studies of marine macromolecules, and a process, iron limitation, that impacts marine ecology on very large geographical scales.

Upregulation of sfuC During Iron Limitation
The only gene to clearly increase in both mRNA and protein abundance during iron limitation was sfuC. This protein localizes to the periplasmic space and binds
dissolved Fe(III) with high affinity. The SfuC-Fe complex associates with the ATPase (SfuA) and permease (SfuB) components of the tripartite ABC transporter complex to actively transport iron into the cell. The fact that \textit{sfuA} and \textit{sfuB} were not observed to increase in protein abundance is not wholly unexpected – SfuA-SfuB complexes only interact with iron-bound SfuC proteins, which are a very small fraction of the total SfuC pool in an iron-limited environment. Additionally, integral membrane proteins such as SfuB are particularly challenging to recover in proteomic studies because they are not readily soluble. This likely contributed to the complete absence of SfuB peptides in all mass spectrometry studies of \textit{Ca. Pelagibacter ubique} to date.

The identification of \textit{sfuC} expression as a readily quantifiable iron limitation marker is particularly useful for ecological surveys. As its name suggests, \textit{Ca. Pelagibacter ubique}’s genome, transcriptome, and proteome regularly dominate bacterial surveys throughout the pelagic environment. Future oceanographic studies seeking evidence of iron availability limiting bacterioplankton growth may use metatranscriptomic or metaproteomic analyses to assess the expression of \textit{sfuC} in the local \textit{Ca. Pelagibacter ubique} population.

\textit{Transcriptome Distinct from Proteome}

Protein abundance was generally uncorrelated with changes in mRNA abundance, suggesting that post-transcriptional mechanisms might be acting at the RNA level to suppress translation. As reviewed in the introduction, previous studies have shown that disparities between a cell’s transcriptome and proteome are the norm rather than the exception. However, the observation that iron-related genes such as \textit{sufA-E} increased in mRNA but not protein indicates that expression of these proteins are controlled at both the level of transcription and at the level of translation.

\textit{Cold-shock Proteins Correlated with Iron Stress}

CspL was significantly more abundant in iron-limited cultures (Figure 5), leading us to closely examine the biological activity of this protein as well as the inversely
expressed homolog CspE. The first discovered member of the cold-shock protein (CSP) family, *E. coli*’s CspA, is highly upregulated under cold stress; it is believed to associate with and melt double-stranded RNA complexes as a mechanism to prevent spurious stem loop structures from interfering with transcription and translation [163–168]. Despite their homology to CspA, many CSP variants are not cold-inducible, but rather are involved in regulating cellular processes [163,169–171] and can even target their activity to specific RNA sequences [172]. A growing body of literature has described mRNAs which modulate their own expression via temperature- (RNA thermometer) or ligand-sensitive (riboswitch) secondary structures [173,174]. Due to the episodic nature of iron deposition into ocean surface waters [175] and the resulting selective pressure favoring rapid response systems for this limiting nutrient [176], we speculate that *Ca. Pelagibacter ubique* CspE and/or CspL affects a reversible inhibition of translation by facilitating an mRNA secondary structure unfavorable for ribosome processing, thereby maintaining the transcriptome in a state of cell growth readiness during times of stress such as iron limitation.

This is the first report describing the general suppression of translation across the entire transcriptome of the cell. In this case, the apparent adaptive significance of protein synthesis suppression is related to urgent cellular requirements to acquire an essential nutrient. The model we propose to explain this phenomenon incorporates activity previously observed in cold-shock proteins, however, the essence of our model assigns cold-shock proteins a new systemic role in *Ca. Pelagibacter ubique* cells with the apparent result of focusing protein synthesis on transporters that target a missing essential nutrient. The validation of this model is beyond the scope of this study. Future work may more precisely identify interactions between cold-shock proteins with specific RNA motifs.

**Summary**

Census information has left little doubt that *Ca. Pelagibacter ubique* plays its role in biological oceanography on a vast scale. To understand this role, we turned inward,
investigating the mechanisms used by these cells to respond to a common form of nutritional stress. One motivation for this study can be described with a term borrowed from satellite remote sensing: the term “ground truth” was coined to describe the validation, by direct measurements, of remotely sensed observations. Metatranscriptomic and metaproteomic measurements are being widely adopted by microbial ecologists anticipating that these approaches will reveal the metabolic status of cells in microbial communities, providing information that can be extrapolated to interpret broader levels of ecosystem function. Essential to this vision is an understanding of how cells respond to environmental variables. Our findings indicate that the periplasmic iron binding protein $sfuC$ is uniquely suitable for assessing the iron limitation status of $Ca$. Pelagibacter ubique cells. We anticipate that ecologists will use this data for interpreting the nutritional status of $Ca$. Pelagibacter ubique cells in nature.

This study, one of the few to simultaneously examine both transcriptional and translational responses in a bacteria cell, uncovered evidence suggesting that $cspL$ might play a role in the cellular response to iron limitation. We offer the model that this protein controls translation in response to environmental conditions for a specific subset of genes present in the transcriptome. We hypothesize that this activity might serve an emergency function, limiting the synthesis of proteins to those that are critical for survival. This finding is consistent with previous reports of post-transcriptional regulation of the iron stress response in which a protein was found to facilitate the degradation of specific mRNAs which encoded nonessential iron-consuming pathways [145,146].

Not only is $Ca$. Pelagibacter ubique one of the most successful cells known, it is also one of the simplest, giving it value as a model for understanding bacterial cell responses. Indeed, numerous new structural RNAs, some widely distributed among bacteria, have been discovered and described in $Ca$. Pelagibacter ubique [104,152]. It is perhaps hubris to imagine that the concept of systems biology might one day be
extended from the machinery of cells to the machinery of microbial ecosystems at work on the scale of oceans. But, if that vision has a chance, it will be by combining studies that cross scales and disciplines to understand the keystone species of the oceans.

**Experimental Procedures**

*Growth Media & Harvesting*

Seawater was collected on 6/14/08 at the Newport Hydroline station NH5 (44° 39.1' N, 124° 10.6' W) from a depth of 10 m. The water was then filtered through a 0.2 μM filter, autoclaved, and sparged with CO$_2$ for 24 hours followed by air for 24-48 hours as previously described [91,177]. Immediately prior to inoculation with *Ca. Pelagibacter ubique* HTCC1062, the media was amended with 50 μM pyruvate, 50 μM glucose, 10 μM nitrogen, 1 μM methionine, 1 μM glycine, 1 μM phosphate, and vitamins. Cells were grown at 20°C (flasks) or 16°C (carboys) with intermittent light and sparging with air. On day 16, three 20 L control cultures were amended with 100 nM ferrichrome and 1 μM FeCl$_3$, and three 20 L treatment cultures were amended with 100 nM ferrichrome only. On day 18, 8 L from each carboy was harvested. On day 28, the remaining ~10L from each was harvested. Prior to each harvest, and on day 17, three 40 mL samples of culture were removed from each culture for microarrays. Water from the three replicate cultures were then combined and growth was arrested using 0.01g chloramphenical and 0.1 mL protease inhibitor cocktail II (CalBiochem #539132) per liter of culture. Tangential flow filtration, followed by centrifugation produced cell pellets for the mass-spectrometry analysis. All samples were kept at ~80°C until analysis.

*Messenger RNA Preparation*

*Ca. Pelagibacter ubique* cells used in microarray experiments were grown in batch cultures as described above. Cells (40 ml for each biological replicate) were collected via centrifugation, and RNA was extracted using RNeasy Mini kits (Qiagen), followed
by amplification with MessageAmp-II Bacteria RNA amplification kit (Ambion). The resulting aRNA was then screened for length and quality using a Bioanalyzer 2100 (Agilent) and quantified utilizing a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific). 5.5 µg of biotinylated aRNA from each sample was then fractionated and hybridized (45°C) overnight to custom Ca. Pelagibacter ubique Affymetrix GeneChip arrays that contained probes for strains HTCC1002, HTCC1062 and HTCC7211 (Pubiquea520471f) using Affymetrix GeneChip Fluidics Station 450, and Affymetrix GeneChip Hybridization Oven 640. Arrays were then washed as per the manufacturer’s instructions and the resulting images were analyzed using an Affymetrix GeneChip Scanner 3000. Fluorescence measurements were normalized over all 18 microarray chips.

Microarray Clustering
A modified radial coordinate visualization plot was used for illustrating mRNA expression in a manner that accentuated condition-specific preferential transcription. In Figure 3, dimensional anchors (DA) representing each of the six microarray samples were positioned manually around the circumference of a circle such that iron-limited samples are on the left, samples with excess iron are on the right, and the vertical placement corresponds to the culture’s transition from exponential growth (bottom) to stationary phase (top). Each gene is represented by a single point, positioned according to the relative abundances between every sample pair, and sized according to the largest observed change in expression level. $PT_g$ is the point for gene $g$, with attributes $x$, $y$, and $s$ describing its x-axis position, y-axis position, and size, respectively. $S_{i,g}$ is the log base-10 average fluorescence for gene $g$ in sample $i$. $DA_i$ is sample $i$’s dimensional anchor positioned at $(DA_{i,x}, DA_{i,y})$ on the graph.

$$PT_{g,x} = \max \left( S_{i,6,g} \right) - \min \left( S_{i,6,g} \right)$$

$$PT_{g,x} = \sum_{1 \leq i < j \leq 6} \left[ (DA_{i,x} - DA_{j,x}) \times \left( S_{i,g} - S_{j,g} \right) / PT_{g,x} \right]$$
\[ PT_{g,y} = \sum_{1 \leq i < j \leq 6} \left[ (DA_{i,y} - DA_{j,y}) \times \left( S_{i,g} - S_{j,g} \right) / PT_{g,y} \right] \]

This type of graph is ideal for revealing if a given gene’s transcript abundance is changing as a result of iron limitation or as a result of the stationary phase transcriptome remodeling induced by iron limitation. An implementation of this algorithm is available in Appendix 1.

**Global TFE Protein Preparation**

Four samples were prepared using the TFE (2,2,2-Trifluoroethanol) digestion method. The cell pellets were reconstituted in 100 mM NH₄HCO₃, pH 8.4 buffer and transferred to a siliconized 0.6 mL microcentrifuge tube. 0.1 mm Zirconia/Silica Beads were added to the top of the tube and bead beat at maximum speed for 3 minutes and immediately placed on ice. A hole was poked in the base of the 0.6 mL siliconized eppendorf tube and placed in a 1.5 mL siliconized eppendorf tube. The sample was then centrifuged for 5 minutes at 14,000 rpm at 4°C. The cell lysis was mixed to a homogenized state and the volume was determined using a pipette. The sample concentration was determined with a Coomassie protein assay and read on a microplate reader. TFE was added to a concentration of 50%. The sample was then homogenized by sonication for one minute in an ice bath followed by incubation at 60°C for two hours with gentle shaking (300 rpm). Proteins were reduced by adding DTT to a final concentration of 2 mM, sonicated for one minute in an ice bath and incubated at 37°C for one hour with gentle shaking. Samples were then diluted 5-fold with 100 mM NH₄HCO₃ to reduce the salt concentration, and CaCl₂ was added to a final concentration of 1 mM. The sample was digested for 3 hours with Trypsin (Promega, Madison WI) at 37°C at a concentration of 1 unit trypsin/50 units protein. After trypsin incubation, a BCA protein assay was performed on the sample to determine the final concentration and vial for mass spectrometer analysis.
**Capillary LC-MS Analysis**

The custom HPLC system was configured using 65-mL Isco Model 65D syringe pumps (Isco, Inc., Lincoln, NE), 2-position Valco valves (Valco Instruments Co., Houston, TX), and a PAL autosampler (Leap Technologies, Carrboro, NC), allowing for fully automated sample analysis across four separate HPLC columns. Reversed-phase capillary HPLC columns were manufactured in-house by slurry packing 3-µm Jupiter C18 stationary phase (Phenomenex, Torrence, CA) into a 70-cm length of 360 µm o.d. x 75 µm i.d. fused silica capillary tubing (Polymicro Technologies Inc., Phoenix, AZ) that incorporated a 0.5-µm retaining screen in a 1/16” custom laser-bored 75 µm i.d. union (screen and union – Valco Instruments Co., Houston, TX; laser bore - Lenox Laser, Glen Arm, MD). Mobile phases consisted of 0.2% acetic acid and 0.05% TFA in water (A) and 0.1% TFA in 90% acetonitrile/10% water (B). The mobile phase flowed through an in-line Degassex DG4400 degasser (Phenomenex, Torrance, CA). The HPLC system was equilibrated at 10 k psi with 100% mobile phase A. Fifty minutes after sample injection the mobile phase was switched to 100% B, which created a near-exponential gradient as mobile phase B displaced A in a 2.5 mL active mixer. A 30-cm length of 360 µm o.d. x 15 µm i.d. fused silica tubing was used to split ~20 µL/min of flow before it reached the injection valve (5 µL sample loop). The split flow controlled the gradient speed under conditions of constant pressure operation (10 k psi). Flow through the capillary HPLC column when equilibrated to 100% mobile phase A was ~400 nL/min.

MS analysis was performed using a ThermoFinnigan LTQ-Orbitrap mass spectrometer (Thermo Scientific, San Jose, CA) with electrospray ionization (ESI). The HPLC columns were coupled to the mass spectrometer by using an in-house manufactured interface. Chemically etched electrospray emitters, 150 um o.d. x 20 um i.d., were used [178]. The heated capillary temperature and spray voltage were 200ºC and 2.2 kV, respectively. Data was acquired for 100 min, beginning 65 min after sample injection (15 min into gradient). Orbitrap spectra (AGC 1x106) were collected from 400-2000 m/z at a resolution of 100k followed by data dependent ion trap MS/MS spectra (AGC
1x104) of the six most abundant ions using a collision energy of 35%. A dynamic exclusion time of 60 sec was used to discriminate against previously analyzed ions. Three technical replicates were run on the mass spectrometer for each cell pellet.

**Quantitative Proteomics**

Quantitative estimates of peptide abundances, calculated from the area under the isotopic profile, were obtained by using a previously developed accurate mass and time (AMT) tag library [98] to search the mass spectra generated by the 12 runs for the four samples. After deisotoping and calculating monoisotopic mass, mass spectrometric features were matched to database peptides with a mass tolerance window of +/- 6ppm and an elution time window of +/- 0.1% after alignment in both dimensions. Peptide abundances were reported for those which had observations in at least 2 of the 3 technical replicates. Linear regression normalization was used to normalize each set of technical replicates as described elsewhere [179]. Briefly, the abundance of peptide $x$ in sample $i$ was transformed into minus versus average space using the following formulas:

$$m_i = \log_2 \left( \frac{x_i}{\bar{x}} \right)$$

$$a_i = \log_2 \left( \frac{x_i \times \bar{x}}{2} \right)$$

Next, the transformed value was corrected based on a linear regression:

$$m_i' = m_i - \hat{m}_i^*$$

where $m_i^*$ is the value for $m_i$ calculated from the $m$ vs $a$ regression equation. Lastly, the computed values were deconvoluted to yield the normalized abundances:

$$x_i' = 2^{(m_i' + 2a_i)/2}$$
Peptides were excluded from further analysis if the standard deviation exceeded the average measurement value among the three technical replicates for a sample. A final filter was applied to exclude the lowest third of peptides for a given protein, when sorted by the peptides’ maximum PeptideProphet F-Score.

Protein abundance was calculated only if a protein had three or more peptides which passed the above filters. Calculating the difference in protein abundance between two samples was a three step process. First, the three replicate peptide abundance measurements were averaged together. Next, the peptide average from sample 1 was divided by the peptide average from sample 2, then $\log_{10}$ transformed. Finally, all $\log_{10}$ peptide ratios from the same protein were averaged together.

To represent the likelihood that a protein was equally abundant in both samples, the multiple peptide measurements were combined into a single statistic as previously described [180]. Briefly, p-values for individual peptides were calculated using a one-tailed Student’s t-test on the technical replicates’ $x'$ values. A two-tailed Student’s t-test was not used because p-values reflecting a large increase would be indistinguishable from p-values reflecting a large decrease. Instead, peptides which changed in the opposite direction from the protein average were assigned a p-value of 1 for their one-tailed Student’s t-test. All peptide p-values for a single protein were then combined into a single chi-square statistic using Fisher’s method:

$$\chi^2 = -2 \times \sum \ln(P_i)$$

Q-values were calculated via the Benjamini-Hochberg method [285] to estimate the false discovery rate arising from multiple comparisons in the microarray and proteomics analyses. In this procedure, each p-value was multiplied by the number of proteins, $T$, and then divided by that p-value’s rank, $k$, among the set of p-values (e.g. the most significant p-value was divided by 1, the second most significant p-value was divided by 2, etc).
q-value = p-value × T / k

ACKNOWLEDGEMENTS

The authors would like to thank the crew of the Elaka for their help in seawater collection, Zanna Chase for measuring the iron concentration in that water, and Mark Wells for advice on siderophore selection and protocols. Input from Paul Carini greatly helped in the design of Figure 4.

TABLES

Table 1. All 23 Ca. Pelagibacter ubique mRNA transcripts that were at least 50 percent more abundant in the iron-limited cultures compared to the control cultures, 24 hours after addition of an iron-chelator.

Seventy-eight percent of these genes are found in Figure 3’s early and late iron stress clusters. Bullet points in the first column indicate contiguous loci.

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>Gene</th>
<th>Description</th>
<th>Ratio</th>
<th>P-Value</th>
<th>Q-Value</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR11_0144</td>
<td>hsIV</td>
<td>ATP-dependent protease: peptidase</td>
<td>1.54</td>
<td>0.001</td>
<td>0.085</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0333</td>
<td>hslU</td>
<td>ATP-dependent protease: ATP-binding</td>
<td>1.59</td>
<td>0.035</td>
<td>0.290</td>
<td>Stat.</td>
</tr>
<tr>
<td>SAR11_0399</td>
<td>rbr</td>
<td>Ruberythrin, hyp. ferroxidase (Fe²⁺ → Fe³⁺)</td>
<td>1.57</td>
<td>0.025</td>
<td>0.280</td>
<td>Stat.</td>
</tr>
<tr>
<td>SAR11_0738</td>
<td>sufA</td>
<td>Transcriptional regulator</td>
<td>1.76</td>
<td>0.062</td>
<td>0.338</td>
<td>Stat.</td>
</tr>
<tr>
<td>SAR11_0739</td>
<td>sufB</td>
<td>Cysteine desulfurase activator complex</td>
<td>2.00</td>
<td>0.009</td>
<td>0.215</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0740</td>
<td>sufC</td>
<td>FeS assembly ATPase</td>
<td>1.80</td>
<td>0.017</td>
<td>0.255</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0741</td>
<td>sufD</td>
<td>FeS assembly protein</td>
<td>2.29</td>
<td>0.019</td>
<td>0.256</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0742</td>
<td>csdB</td>
<td>Selenocysteine lyase chain A</td>
<td>2.44</td>
<td>0.046</td>
<td>0.300</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0743</td>
<td>sufE</td>
<td>Putative NIU-like protein</td>
<td>2.24</td>
<td>0.004</td>
<td>0.183</td>
<td>Late</td>
</tr>
<tr>
<td>SAR11_0744</td>
<td>paaD</td>
<td>Phenylacetic acid degradation protein</td>
<td>2.00</td>
<td>0.001</td>
<td>0.080</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0745</td>
<td>hesB</td>
<td>HesB protein</td>
<td>1.95</td>
<td>0.000</td>
<td>0.067</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_0785</td>
<td>Conserved hypothetical protein (DUF952)</td>
<td>1.52</td>
<td>0.051</td>
<td>0.305</td>
<td>Late</td>
<td></td>
</tr>
<tr>
<td>SAR11_1233</td>
<td>aziC</td>
<td>AziC protein</td>
<td>2.14</td>
<td>0.016</td>
<td>0.256</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_1235</td>
<td>sfuA</td>
<td>Iron(III) ABC transporter: ATP-binding</td>
<td>4.99</td>
<td>0.000</td>
<td>0.014</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_1237</td>
<td>sfuB</td>
<td>Iron(III) ABC transporter: permease</td>
<td>10.36</td>
<td>0.000</td>
<td>0.023</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_1238</td>
<td>sfuC</td>
<td>Iron(III) ABC transporter: periplasmic</td>
<td>16.00</td>
<td>0.003</td>
<td>0.174</td>
<td>Early</td>
</tr>
<tr>
<td>SAR11_1239</td>
<td>Unknown protein</td>
<td>2.47</td>
<td>0.006</td>
<td>0.200</td>
<td>Late</td>
<td></td>
</tr>
<tr>
<td>SAR11_1240</td>
<td>aceA</td>
<td>Isocitrate lyase</td>
<td>1.58</td>
<td>0.091</td>
<td>0.393</td>
<td>Late</td>
</tr>
<tr>
<td>SAR11_1242</td>
<td>Transcription regulator</td>
<td>1.60</td>
<td>0.054</td>
<td>0.316</td>
<td>Late</td>
<td></td>
</tr>
<tr>
<td>SAR11_1253</td>
<td>dmgdh</td>
<td>Dimethylglycine dehydrogenase</td>
<td>1.54</td>
<td>0.054</td>
<td>0.314</td>
<td>Late</td>
</tr>
<tr>
<td>SAR11_1279</td>
<td>Unknown membrane protein</td>
<td>1.56</td>
<td>0.056</td>
<td>0.319</td>
<td>Late</td>
<td></td>
</tr>
</tbody>
</table>

a Average fluorescence of three replicates, (iron limited culture / iron replete culture).

b Result of a two-tailed Student’s t-test comparing the three biological replicates for each treatment.

^ Computed using the Benjamini-Hochberg method.
Table 2. All 17 *Ca. Pelagibacter ubique* proteins that were at least 50 percent more abundant in the iron-limited cultures compared to the iron replete cultures, two and 12 days after addition of an iron chelator.

<table>
<thead>
<tr>
<th>Day</th>
<th>Locus ID</th>
<th>Gene</th>
<th>Description</th>
<th>Ratio</th>
<th>P-Value</th>
<th>Q-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>SAR11_1238</td>
<td><em>sfuC</em></td>
<td>Iron(III) ABC transporter: periplasmic</td>
<td>11.41</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>18</td>
<td>SAR11_1274</td>
<td><em>cspL</em></td>
<td>DNA-binding cold shock protein</td>
<td>6.20</td>
<td>0.000</td>
<td>0.015</td>
</tr>
<tr>
<td>18</td>
<td>SAR11_0161</td>
<td><em>groES</em></td>
<td>Protein-folding chaperonin</td>
<td>2.01</td>
<td>0.028</td>
<td>0.049</td>
</tr>
<tr>
<td>18</td>
<td>SAR11_1062</td>
<td><em>dapA</em></td>
<td>Dihydricdipicolinate synthase</td>
<td>1.53</td>
<td>0.849</td>
<td>0.949</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_1238</td>
<td><em>sfuC</em></td>
<td>Iron(III) ABC transporter: periplasmic</td>
<td>26.96</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_1161</td>
<td><em>sbcC</em></td>
<td>ATPase involved in DNA repair</td>
<td>4.59</td>
<td>0.011</td>
<td>0.031</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0161</td>
<td><em>groES</em></td>
<td>Protein-folding chaperonin</td>
<td>3.59</td>
<td>0.008</td>
<td>0.024</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0601</td>
<td><em>ftsH</em></td>
<td>Metalloprotease</td>
<td>3.28</td>
<td>0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_1124</td>
<td><em>rplL</em></td>
<td>50S ribosomal protein L31</td>
<td>3.16</td>
<td>0.006</td>
<td>0.020</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0430</td>
<td><em>aceF</em></td>
<td>Dihydrolipoamide S-acetyltransferase</td>
<td>3.02</td>
<td>0.094</td>
<td>0.192</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0171</td>
<td><em>ftsH</em></td>
<td>Rhodanese-related sulfuryltransferase</td>
<td>2.74</td>
<td>0.002</td>
<td>0.010</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0235</td>
<td><em>cspL</em></td>
<td>DNA-binding cold shock protein</td>
<td>2.27</td>
<td>0.437</td>
<td>0.558</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0791</td>
<td><em>cspL</em></td>
<td>DNA-binding cold shock protein</td>
<td>2.26</td>
<td>0.035</td>
<td>0.085</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0054</td>
<td><em>pilA</em></td>
<td>Pilin protein</td>
<td>2.16</td>
<td>0.020</td>
<td>0.052</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0727</td>
<td><em>accB</em></td>
<td>Acetyl-CoA carboxylase</td>
<td>2.03</td>
<td>0.301</td>
<td>0.421</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0987</td>
<td><em>ppIB</em></td>
<td>Peptidylprolyl isomerase</td>
<td>1.99</td>
<td>0.291</td>
<td>0.412</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0793</td>
<td><em>ppIB</em></td>
<td>Peptidylprolyl isomerase</td>
<td>1.74</td>
<td>0.128</td>
<td>0.226</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0599</td>
<td><em>ppIB</em></td>
<td>Peptidylprolyl isomerase</td>
<td>1.70</td>
<td>0.623</td>
<td>0.778</td>
</tr>
<tr>
<td>28</td>
<td>SAR11_0708</td>
<td><em>acpP</em></td>
<td>Acyl carrier protein</td>
<td>1.55</td>
<td>0.198</td>
<td>0.304</td>
</tr>
</tbody>
</table>

* Average spectra height of at least three peptides, (iron limited culture / iron replete culture).

** Combined one-tailed Student's t-test comparing the three technical replicates for each treatment.

^ Computed using the Benjamini-Hochberg method.
Figure 2: Growth of *Candidatus* Pelagibacter ubique cells was arrested by iron-sequestering siderophores.

(A) Cell densities observed during a pilot experiment to test the effect of the two siderophores ferrichrome and deferoxamine mesylate salt at varying concentrations on the growth of *Candidatus* Pelagibacter ubique HTCC1062. The first arrow indicates the introduction of siderophore/iron as described by the legend. The second arrow indicates the delayed 1 μM iron additions parenthetically noted in the legend. (B) Cultures for harvesting were grown in six 20 L carboys. The first arrow indicates the introduction of siderophore/iron as described by the legend. Proteins and mRNA were analyzed on the dates indicated by the unfilled arrows: microarray samples were taken from cultures on days 17, 18, and 28; proteomic samples were taken on days 18 and 28.
Figure 3: Genes transcribed during iron limitation were different from stationary phase genes.

The four clusters indicate up-regulation of similar condition-specific mRNA. Symbols for each microarray sample (open circles) were manually positioned on a circle according to each sample’s iron availability and growth rate. Genes were “attracted” to the samples in which they were most abundant. Larger points indicate genes with larger condition-to-condition variation; a key for the 10 largest points in each cluster is provided.
Figure 4: Protein abundances were largely decoupled from transcript abundances.

The change in protein abundance versus the change in mRNA abundance was plotted for all Ca. Pelagibacter ubique genes that showed a significant ($P \leq 0.05$) change in either measurement. Each color represents a different comparison between treatments or timepoints, with $R^2$ values of 0.11, 0.08, 0.09, and 0.02 respective to the legend’s ordering. Large ellipses indicate clusters of the same colored points. Histograms on the low end of each axis further define the distribution of points. Points represented by a diamond are discussed at length in the text.
Figure 5: Translation of *Ca. Pelagibacter ubique*’s cold shock and iron-binding genes are influenced by iron availability.

The abundance of two *Ca. Pelagibacter ubique* cold shock proteins, CspE and CspL, and the iron-binding protein SfuC, appear to be correlated with iron availability (p-value of .02, .08, and 3e-79, respectively).
CHAPTER 3. PROTEOMIC AND TRANSCRIPTOMIC ANALYSIS OF 
*CANDIDATUS PELAGIBACTER UBIQUE* DESCRIBES THE FIRST PII-
INDEPENDENT RESPONSE TO NITROGEN LIMITATION IN A FREE-
LIVING ALPHAPROTEOBACTERIUM

Daniel P. Smith, J. Cameron Thrash, Carrie D. Nicora, Mary S. Lipton, Kristin E.
Burnum, Paul Carini, Richard D. Smith, & Stephen J. Giovannoni

In preparation for submission to: PLoS ONE
ABSTRACT
Nitrogen is one of the major nutrients limiting microbial productivity in the ocean, and as a result marine microorganisms have evolved specialized systems for responding to nitrogen stress. The highly abundant alphaproteobacterium Candidatus Pelagibacter ubique lacks the canonical GlnB, GlnD, and NtrB/NtrC genes for regulating nitrogen assimilation. A survey of 127 Alphaproteobacteria genomes found these genes to be highly represented in free-living and pathogenic organisms with large genomes and only missing in a subset of obligate intracellular organisms and other SAR11 strains. We examined global differences in mRNA and protein expression in Ca. P. ubique strain HTCC1062 during nitrogen-limited and nitrogen-replete stationary phase to understand how this thriving organism responds to nitrogen limitation. Transporters for ammonium (AmtB), taurine (TauA), amino acids (YhdW), and opines (OccT) were all elevated in nitrogen-limited cells, indicating they devote increased resources to the assimilation of nitrogenous compounds. Enzymes for assimilating amine into glutamine (GlnA) and glutamate (AspC, GltBD) were similarly up-regulated. Differential regulation of the transcriptional regulator NtrX in the two-component signaling system NtrY/NtrX was also observed, implicating it in the control of the nitrogen starvation response. Comparisons of the transcriptome and proteome suggest that Amt is post-transcriptionally repressed during nitrogen limitation, supporting previous studies that computationally identified a novel cis-acting riboswitch upstream of this gene. These observations support the conclusion that Ca. P. ubique has an unusually simple regulatory system that enables it to increase its capacity for the uptake of nitrogenous compounds in response to nitrogen limitation.
**INTRODUCTION**

Identifying nutrient factors that limit microbial productivity in the oceans has been one of the key missions of biological oceanographers for over half of a century. Using a combination of nutrient fertilization and direct measurement of dissolved nutrient concentrations, studies have alternately found nitrogen [18–27], phosphate [6–8], iron [11,181–183], or silica [17,184–186] to limit the productivity in seawater. Meta-analyses that coalesced experimental results across hundreds of studies found that anthropogenic contamination and geographical features influenced the limiting nutrient, with nitrogen more often limiting in pelagic marine environments and in polluted coastal waters [16,17]. Recently, individual studies and meta-analyses have begun to conclude that more than one nutrient – often nitrogen and phosphate – are co-limiting [12,15,16], due to shifts in overall N:P stoichiometry of bacterial communities according to nutrient availability [187,188].

Biosynthesis of nitrogenous compounds such as DNA, RNA, and proteins is dependent on maintaining intracellular pools of glutamine and glutamate. In nearly all bacteria, these two compounds are synthesized by glutamine synthetase (GS) and glutamate synthetase (glutamine-2-oxoglutarate-amido transferase; GOGAT). These two enzymes work in concert to first condense ammonia and glutamate via GS to form glutamine, followed by the GOGAT-mediated transfer of an amine group from glutamine onto 2-oxoglutarate to yield two molecules of glutamate [189–192]. The activity of these enzymes in many Alphaproteobacteria is regulated by the P_II protein GlnB, which is alternatively uridylylated/de-uridylylated by GlnD based on the 2-oxoglutarate/glutamine ratio within the cell [189–192]. The two-component signaling system NtrB/NtrC transduces the uridylylation state of GlnB into transcriptional inhibition/activation of GS and other nitrogen assimilation genes [193]. Un-uridylylated GlnB also adenylylates GS, thereby inhibiting GS activity when glutamine is sufficient [194,195]. Uridylylated GlnB catalyzes the de-adenylylation of GS to restore its activity. A second P_II protein, GlnK, is commonly co-transcribed with the ammonium transporter amtB, and post-translationally regulates AmtB’s activity.
based on the uridylylation state of GlnB [190,196–198]. Altogether, this post-translational signaling cascade is believed to enable the cell to quickly inhibit ammonia uptake and glutamine synthesis when exposed to pulses of high concentrations of ammonia [199–201], thereby preventing toxic build-up of intracellular ammonia and depletion of the TCA cycle intermediate 2-oxoglutarate.

Despite the ecological importance bacterioplankton communities’ interaction with nutrient limited environments, very little work has been done to date to determine how the abundant clade of SAR11 organisms survives in oceans where bioavailable nitrogen is commonly a scarce resource. When the first representative species of the SAR11 clade was sequenced in 2005, only two genes for regulating the assimilation of nitrogen were identified: \textit{ntrY} and \textit{ntrX} [92]. This two-component signaling system, although ubiquitous throughout alphaproteobacteria, is still poorly understood. These two genes bear a high level of sequence similarity to \textit{ntrB} and \textit{ntrC}, a well described two-component system for regulating the transcription of nitrogen assimilation genes [189–192]. Many organisms have both pairs of genes, and due to cross-talk between them, it has been speculated that they work in concert by detecting different stimuli relating to nitrogen availability. Evidence suggests that unlike NtrB/NtrC which responds to fluctuations in intracellular glutamine, NtrY/NtrX may be involved in sensing the concentration of extracellular nitrate [202] and has been postulated to connect nitrogen control to the redox state of the cell through interactions with the RegB/RegA two-component system [203].

The elemental composition of microorganisms is shaped in large part by nutrient availability [204–210]. A survey of metagenomic sequences and the genomes of marine bacteria, including SAR11, concluded that competition for nitrogen in the marine environment has selected for genomes high in AT and proteomes low in nitrogenous amino acids [27]. Additionally, microorganisms with the lowest mass in this environment have a distinct advantage – not only because their absolute nutrient requirements for growth and division are minimal, but also because the increased
surface-to-volume ratio of small cells favors nutrient acquisition [27,95]. Nitrogen and phosphate limitation on evolutionary timescales are hypothesized to drive genome streamlining, in which loss of non-essential DNA results in an increase in fitness due to decreased costs of genome replication [92,102,211]. Oligotrophic marine environments are numerically dominated by members of Prochlorococcus and SAR11 that are well adapted to surviving in nitrogen-limited environments through the aforementioned adaptations [27,92,212]. A cultured representative of the SAR11 clade, Candidatus Pelagibacter ubique HTCC1062, was found to have one of the shortest genomes (1.3 Mbp), lowest GC content (33%), and smallest cell size (0.025 – 0.045 \( \mu \text{m}^3 \)) among free-living heterotrophs [92,97]. Assuming that these characteristics are the result of extreme selective pressure to optimize nitrogen utilization, we are led to question if the regulation, acquisition, or assimilation of nitrogen has also been radically remodeled in this bacterium.

The broad objective of this study was to understand the regulatory responses of SAR11 cells to nitrogen limitation. Our experimental strategy was to measure both the transcriptome and the proteome. In a previous study of iron limitation in HTCC1062, we observed that the response of the proteome and transcriptome are largely uncoupled, an apparent consequence of an integrated regulatory response that includes both transcriptional and post-transcriptional mechanisms of regulation [213]. Previous work has also shown that SAR11 cells encode more riboswitches than any other organism, including a number of novel riboswitches of uncertain function [104,152,214]. The data we present demonstrate that SAR11 has a robust response to nitrogen limitation that enables it to increase its capacity to assimilate amine groups from nitrogenous compounds in response to nitrogen limitation. This regulatory response is simpler than any described previously in alphaproteobacteria, and includes a novel riboswitch that we propose regulates translation of the Amt transcript.
RESULTS AND DISCUSSION

Nitrogen Assimilation Genes in Ca. P. ubique.
We analyzed published data from SAR11 genomes and metagenomic data in detail to provide a basis for interpreting experimental data. As reported previously, the genome of Ca. P. ubique HTCC1062 encodes the requisite transporters and enzymes necessary for assimilating nitrogen from ammonium and organic molecules, but lacks the pathways for reducing nitrite, nitrate, and di-nitrogen [92,215]. Ammonium is likely to be imported by one or more of this bacterium’s four putative Amt transporters. The genome encodes one high efficiency GS-I class enzyme (GlnA) and two low efficiency GS-III class enzymes (GlnT) for condensing ammonium with glutamate to synthesize glutamine. The large (GltB) and small (GltD) subunits of GOGAT are also present, and complete the metabolic pathway for regenerating glutamate from glutamine and 2-oxoglutarate. Dissolved organic matter (DOM) is an important source of reduced nitrogen in the marine environment, and Ca. P. ubique encodes several genes which may be involved in scavenging amine moieties from DOM compounds including amino acid transporters, deaminases, and aminotransferases.

Absence of the P II Regulatory Network
The canonical P II nitrogen assimilation regulatory network present in all other free-living Alphaproteobacteria that we examined is absent from Ca. P. ubique. The absence of glnB, glnD, glnK, ntrB, and ntrC is not unique to strain HTCC1062; examination of all seven available SAR11 genomes (one draft – HIMB114; six finished – HTCC1062, HTCC7211, HTCC1002, HTCC9565, HIMB5, and HIMB59) suggests that the absence of these five genes is a common characteristic of species in the Ca. Pelagibacter genus, although they are present in the highly divergent SAR11 Group V organism, HIMB59 [216]. The distribution of these genes across the Alphaproteobacteria confirmed that only a subset of obligate intracellular organisms was similarly deficient in P II regulatory genes (Figure 6).
The missing P$_{II}$ genes $glnB$, $glnD$, $glnK$, $ntrB$, and $ntrC$ are often observed in other bacteria within widely conserved gene neighborhoods: $glnB$-$glnA$-$ntrB$-$ntrC$, $map$-$glnD$-$dapA$, and $glnK$-$amtB$ [192]. We therefore examined the gene neighborhoods of the SAR11 genes $glnA$, $map$, $dapA$, and $amtB$ in each of the available SAR11 genomes and in environmental reads as a complementary method to confirm the absence of the P$_{II}$ genes from this clade. Manual inspection of whole genome alignments constructed with progressiveMauve [217] revealed that gene order around $glnA$, $map$, $dapA$, and $amtB$ was conserved among the seven available SAR11 genomes (except HIMB59), and that $glnB$, $glnD$, $glnK$, $ntrB$, and $ntrC$ were indeed absent from these loci.

HIMB59, which possessed $glnB$, $glnD$, and $glnK$, exhibited the typical $glnB$-$glnA$ and $glnK$-$amtB$ arrangements, however, $glnD$ was not encoded adjacent to $map$ or $dapA$.

To assess the abundance of the P$_{II}$ genes $glnB$, $glnD$, $glnK$, $ntrB$, and $ntrC$ in natural populations of SAR11, the global ocean sampling (GOS) environmental dataset [57] was queried for known P$_{II}$ operons [192] – $glnB$-$glnA$-$ntrB$-$ntrC$, $map$-$glnD$-$dapA$, or $glnK$-$amtB$ – in which $glnA$, $map$, $dapA$, or $amtB$ showed greatest similarity to a SAR11 strain. Among reads contain two or more genes in the GOS database, 287 encoded a SAR11-like $amtB$, and of those, 3 encoded both $glnK$-$amtB$ (3/287). Using $glnA$, $map$, and $dapA$ found no P$_{II}$ genes: 0/137, 0/15, and 0/74, respectively. The widespread absence of the P$_{II}$ regulatory network in this bacterial clade suggests that Ca. P. ubique uses an alternative mechanism to modulate the expression and activity of nitrogen assimilation genes – likely one which does not rely on the relative concentrations of glutamine and 2-oxoglutarate.

Ca. P. ubique responded to nitrogen limitation by increasing the abundance of proteins responsible for transporting nitrogenous molecules into the cell and for transferring amine groups to 2-oxoglutarate or glutamate to form glutamate or glutamine, respectively (Table 3; Figure 7). The two most highly up-regulated genes – an ammonium transporter and a sarcosine oxidase – appear to be expressed solely during nitrogen-limiting conditions. A diverse collection of organic molecule transporters, as
well as aminotransferases, were also observed at elevated abundance in nitrogen-limited cells, apparently enabling the cell to utilize a broad range of nitrogenous compounds to satisfy its nitrogen requirements. Interestingly, many genes predicted to be involved in nitrogen metabolism – including three of four putative ammonium transporters – were down-regulated or unchanging under nitrogen-limiting conditions (Table 4).

**Regulatory Proteins NtrY/NtrX**

Nitrogen limitation in *Ca. P. ubique* did not have any effect on transcription or translation of the membrane sensor NtrY. The protein concentration of the transcriptional regulator NtrX, however, was much lower during nitrogen-limiting conditions (Table 4). Therefore, it is likely that in *Ca. P. ubique*, the two-component signaling system NtrY/NtrX is responsive to nitrogen limitation. However, future studies are needed to determine which, if any, proteins are regulated by NtrX. Fortunately, *Ca. P. ubique* is an excellent model organism for studying the activity of the NtrY/NtrX system due to the absence of NtrB/NtrC, GlnB, and GlnD, which influence the activity of NtrY/NtrX in other organisms [202,218,219].

**Ammonium Transport and Assimilation**

Inorganic nitrogen in the form of ammonium (NH₄⁺) is an important source of nitrogen in the marine environment. Although *Ca. P. ubique* is predicted to encode four ammonium transporters in its genome – *SAR11_0049, SAR11_0050, SAR11_0818*, and *SAR11_1310* – only one, *SAR11_0818* (*amtB*), was observed to significantly increase in response to nitrogen limitation. *AmtB*’s transcripts were nearly three times more abundant, and its protein product was only detected, under nitrogen-limiting conditions. A second putative ammonium transporter, *SAR11_1310* (*amt*), was equally transcribed under both conditions but its protein product was never detected in the nitrogen-limited treatment, indicating that the Amt protein is unlikely to be an ammonium importer. This is the first evidence that three of the predicted ammonium
transporters in Ca. P. ubique (SAR11_0049, SAR11_0050, and SAR11_1310) are not responsible for importing ammonium, and may transport another substrate.

The expression pattern of amt is consistent with riboswitch-mediated translation inhibition. Evidence for the involvement of structural RNA elements in nitrogen regulation has been increasing in recent years: a glutamine-sensing riboswitch has been described in cyanobacteria and marine metagenomic sequences [220], a transcriptional attenuator was found to regulate expression of the PII gene for nitrogen fixation in the deltaproteobacterium Geobacter sulfurreducens [221], and putative riboswitches have been computationally identified upstream of both amt and amtB in Ca. P. ubique [152]. In light of Ca. P. ubique’s utilization of riboswitches for regulating a diverse array of metabolic processes [104,152,214], and the absence of the usual AmtB regulators (NtrC and GlnK) in this genome, we speculate that Ca. P. ubique may have acquired a novel riboswitch system for regulating Amt and AmtB protein synthesis.

Ammonium transported into the cell via the AmtB protein is condensed with glutamate to form glutamine – a reaction catalyzed by glutamine synthetase (GlnA). As with amtB, Ca. P. ubique’s genome encodes multiple copies of glutamine synthetase; only one of which (SAR11_0747) is up-regulated in response to nitrogen limitation. Protein products from the other two glutamine synthetase genes (SAR11_1305 and SAR11_1316) were observed to be less abundant during nitrogen limitation. This pattern of regulation for Ca. P. ubique’s glutamine synthetase proteins appears related to the specific subclass each glutamate synthetase gene belongs to. SAR11_0747 is a class I (glnA), while SAR11_1305 and SAR11_1316 are class III (glnT). Understanding the role of each Ca. P. ubique glutamine synthetase gene may bring greater insight into the fundamental differences between the classes of the widespread glutamine synthetase protein family.
**Organic Nitrogen Transport and Assimilation**

Several organic molecule transporters were observed to be more abundant in cells limited by nitrogen availability. As illustrated in Figure 7, these include the periplasmic components of ABC transporters which recognize a diverse collection of nitrogenous molecules including amino acids (YhdW), taurine (TauA), and opines (OccT). The outer membrane porin protein OmpU was also more abundant in nitrogen-limited cultures, and may serve to increase the diffusion rate of both ammonium and nitrogenous organic molecules into the periplasmic space.

Although a variety of transporters are used by HTCC1061 to import nitrogenous compounds, only two enzymes appear to be responsible for scavenging their amine moieties. The first of these, aspartate transaminase (AspC; EC 2.6.1.1), catalyzes the transfer of primary amine groups onto 2-oxoglutarate to form glutamate. The second up-regulated enzyme (SAR11_1304) is tentatively annotated as a monomeric sarcosine oxidase, as it bears sequence and structural similarities to both deaminases and amine demethylases. As an amine demethylase, this enzyme would function to convert secondary amines (R-NH-CH$_3$) and tertiary amines (R-N(CH$_3$)$_2$) to primary amines (R-NH$_2$). A deaminase, on the other hand, would liberate ammonia from compounds from primary amine groups. In order to metabolize the broad range of nitrogenous compounds transported into the cell, these two enzymes would have to exhibit a high degree of substrate promiscuity – an enzymatic property which has recently been the focus of increased interest [222,223] and could allow Ca. P. ubique to scavenge nitrogen from a wide range of compounds with minimal investment in protein synthesis.

Nitrogenous compounds such as taurine and opines, for which increased transporter expression was observed, are not known substrates for AspC or sarcosine oxidase, suggesting that in Ca. P. ubique either or both of these enzymes may be more promiscuous than previously known. Together, a pair of broad specificity aminotransferases and deaminases/amine-demethylases could be capable of efficiently
scavenging nitrogen from a wide variety of organic compounds. Given that ArgC and SAR11_1304 are among the few enzymes that were up-regulated during nitrogen limitation, the promiscuous enzyme model we propose is quite possibly Ca. P. ubique’s strategy for acquiring nitrogen. However, additional studies are necessary to definitively characterize the activities of these two proteins.

Proposed Regulatory Model
The PII regulatory system performs three important functions: 1) regulate GS/GOGAT activity to prevent the depletion of the TCA cycle intermediate 2-oxoglutarate, 2) control ammonia transport to guard against toxic pulses, and 3) inhibit/stimulate transcription of genes related to nitrogen assimilation. Based on the observations in this study, as well as those from previous experiments with Ca. P. ubique, we propose the following model for how this organism regulates these three processes without the usual collection of PII genes.

In most bacteria, maintaining an adequate pool of 2-oxoglutarate is essential to ensure proper operation of the TCA cycle, of which 2-oxoglutarate is an intermediate. However, the TCA cycle in Ca. P. ubique can operate in a 2-oxoglutarate independent manner through the AceA-mediated cleavage of isocitrate to succinate and glyoxylate; succinate re-enters the TCA cycle directly, and glyoxylate, after condensation with acetyl-CoA, re-enters as malate [104]. This “glyoxylate bypass” bypasses the steps in the TCA cycle that produce and consume 2-oxoglutarate. Therefore, the unchecked consumption of 2-oxoglutarate by GS/GOGAT may be tolerated by this bacterium. This theory is supported by our observation that the SAR11 strain HIMB59 which encodes a PII network is lacking the genes for the glyoxylate bypass.

We propose that two ammonia transporters control the intracellular concentration of ammonia: AmtB functioning as an ammonia importer and SAR11_1310 as an ammonia exporter serving to limit intracellular spikes in ammonia concentration. This theory is derived from the observation that AmtB protein was observed only in
nitrogen-limited cultures, whereas SAR11_1310 was observed only in nitrogen-replete cultures. Additionally, the evidence for post-transcriptional regulation of SAR11_1310 suggests that transcripts for this putative ammonia exporter can be translated rapidly to produce protein in response to an increase in substrate concentration. We speculate that toxic levels of ammonia interact with a cis-acting structural RNA [152] to activate translation of the downstream SAR11_1310 gene on the mRNA transcript, thereby producing the protein product to export ammonia from the cell.

The nitrogen-responsive transcriptional regulator in Ca. P. ubique is proposed to be NtrX. This protein has been previously noted for its similarity to NtrC, the transcriptional regulator in the typical PII regulatory system, and can recognize DNA promoter regions [218,219]. Nitrogen limitation was observed to have an effect on NtrX abundance in this experiment, with nitrogen-limited cells containing less than half as much NtrX as nitrogen-replete cells. This difference in concentration, as well as potential post-translational modifications through the NtrY/NtrX two-component system, may affect transcription of genes involved in nitrogen assimilation in a manner similar to NtrC in other alphaproteobacteria.

SAR11 cells are proving to be fascinating subjects for experimental study because of their genomic and metabolic simplicity and their extraordinary success in competitive pelagic environments which are frequently nutrient-limited. The data and analyses we report here demonstrate that Ca. Pelagibacter possesses a robust response to N-limitation that includes both transcriptional and post-transcriptional control elements. This response involves a reduced set of genes relative to other free-living alphaproteobacteria, including the absence of the otherwise highly conserved PII system. We propose a model in which a novel riboswitch and enzyme promiscuity play important roles in governing the response of HTCC1062 and other SAR11 to periodic limitation in the availability of nitrogenous substrates.
**Experimental Procedures**

*Phylogenetic Analysis/P* II Orthologs*

The Hal software package [224] was used to generate clusters of orthologous proteins from 127 representative *Alphaproteobacteria* and six outgroups, including *Escherichia coli*, using all vs. all BLASTP followed by Markov clustering (MCL) at thirteen inflation parameters. Clusters generated by the conservative inflation parameter of 3.0 were used to identify orthologs to relevant nitrogen assimilation genes in each of the species analyzed. The pictured phylogenetic tree was constructed using 33 protein sequences conserved among these species, corresponding to the Alphaproteobacterial tree with 20% allowed missing data and liberal GBlocks settings as previously described [216].

*Searching Metagenomic Sequences for SAR11 P* II Genes*

To investigate the frequency of P* II genes in natural populations of SAR11, we searched the global ocean sampling (GOS) dataset [57] for P* II genes present on the same read as SAR11 genes. In this search, P* II genes absent from most SAR11 genomes (P* II genes) are defined as *glnB, glnD, glnK, ntrB,* and *ntrC.* Genes present in SAR11 which are often neighbors to P* II genes in other species (neighbors) are defined as *glnA, map, dapA, amtB, ntrX,* and *ntrY.* TBLASTN was used to search the GOS read dataset for neighbors using an expect score cutoff of 1e⁻⁵ and a limit of 10,000 matches per neighbor. Open reading frames (ORFs) were translated from reads returned by this search, and then searched against the NCBI non-redundant database using BLASTP to locate their top match. Reads were retained for further analysis if they contained at least two ORFs, where one ORF was a neighbor with a top match to a SAR11 strain, and a second ORF had a top match to a P* II gene from any species. The three reads which met these criteria were examined manually and determined to have non-spurious top matches; all top matches on these reads had amino acid identities ranging from 63 – 67 % over 108 – 140 residues, and expect scores of 2e⁻³³ – 2e⁻⁴³.
Growth Media and Harvesting

Artificial seawater (ASW) media was made using previously established protocols [225]. Water, salts, and metals were added to eight 20-L polycarbonate carboys as detailed in Table 5, then autoclaved for ten hours. After cooling to room temperature, carboys were sparged with CO₂ for 20 hours and brought up to 20 L using sterile water to compensate for evaporation due to autoclaving. Vitamins, nutrients, and ammonium were added to each carboy from a filter-sterilized stock solution and then sparged overnight on air while cooling to 16 °C. Final nutrient concentrations were as follows: 10 µM DMSP, 500 µM oxaloacetate, 500 µM pyruvate, and (for the nitrogen-replete cultures only) 20 µM ammonium. Nitrogen-limited cultures received no ammonium amendment, but were expected to grow to a lower density on trace contaminants containing nitrogen. The media was then inoculated with an ASW-adapted culture of Candidatus Pelagibacter ubique HTCC1062 [225] growing exponentially in nitrogen-limited media. Following inoculation, cultures were incubated at 16 °C with constant air sparging.

Culture growth was tracked daily by staining cells with SYBR Green and counting on a Guava EasyCyte flow cytometer. Two cultures, one from each treatment, failed to attain the density necessary for analysis and were not harvested. Samples for microarray and proteomic analysis were taken from the six remaining cultures at two time points: exponential growth and stationary phase. At these time points samples were collected using the identical protocol from chapter 3: 5×10¹⁰ cells were removed to a separate vessel and amended with 10 mg chloramphenicol, 100 µL 500 mM EDTA, and 100 µL 100X Halt Protease Inhibitor Cocktail (Thermo Scientific #78438) per liter of culture. Tangential flow filtration against a Pellicon 2 Mini 30K membrane (Millipore #P2C030C01), reduced the volume of culture to less than 150 mL which was subsequently centrifuged for 1 hour at 20,000 rpm and 0 °C. The pelleted cells were resuspended 1 mL TE and centrifuged again in a single 1.5 mL tube for 40 minutes at 40,000 rpm and 10 °C. After decanting the supernatant, the pellet was stored at -80 °C until proteomic analysis. Microarray samples, which require less
biomass, were harvested through a separate protocol simultaneously. Eighty milliliters of culture was collected and centrifuged for 1 hour at 20,000 rpm and 0 °C. Cells were resuspended in 1 mL of RNAprotect Bacteria Reagent (Qiagen #76506) and then centrifuged again in a 1.5 mL tube for 40 minutes at 40,000 rpm and 0 °C. After decanting the supernatant, microarray samples were placed at -80 °C until analysis.

**Messenger RNA Preparation**

Each cell pellet was resuspended in 100 µL TE (1 mM Tris; 1 mM EDTA; pH 8) containing 40 µg lysozyme and incubated at room temperature for 5 minutes. Total RNA was extracted using an RNeasy MinElute Cleanup Kit (Qiagen #74204) according to the manufacturer's instructions, with the exception of an additional wash step with 700 µL Buffer RW1 (Qiagen #1053394) immediately prior to the prescribed washes with Buffer RPE. Eluted RNA was quantified using a NanoDrop spectrophotometer (Thermo Scientific ND-1000) and found to vary over the range of 243 – 837 ng. One-hundred nanograms of total RNA from each sample was amplified and labeled using a MessageAmp II-Bacteria RNA Amplification Kit (Ambion #AM1790) and biotin-11-UTP (Ambion #AM8451) according to the MessageAmp "Improved Protocol" Handbook. Polyadenylation, reverse transcription, second strand synthesis and in-vitro transcription reactions were performed in the recommended volume and incubated for 19 hours rather than 14. NanoDrop measurements of amplified RNA showed recovery of 93 – 147 µg per sample.

**RNA Labeling and Microarray Processing**

Protocols for microarray analysis were identical to those in chapter 3: RNA integrity screening, Probe synthesis, hybridization and scanning were conducted by the CGRB Core Laboratories at Oregon State University, Corvallis OR. Five micrograms of total RNA was used to generate biotinylated complementary RNA (cRNA) for each treatment group using the One-Cycle Target Labeling protocol (Affymetrix, Santa Clara, CA) from the GeneChip Expression Analysis Technical Manual (701021 Rev. 5). In short, isolated total RNA was checked for integrity and concentration using the
RNA 6000 Nano LabChip kit on the Agilent Bioanalyzer 2100 (Agilent Technologies, Inc., Palo Alto, CA). Poly-A RNA control kit RNA and Ca. P. ubique RNA were reverse transcribed using a T7-(dT) 24 primer and Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA) and double stranded cDNA was synthesized and purified with GeneChip® Sample Cleanup Modules (Affymetrix, Santa Clara, CA). Biotinylated cRNA was synthesized from the double stranded cDNA using T7 RNA polymerase and a biotin-conjugated pseudouridine containing nucleotide mixture provided in the IVT Labeling Kit (Affymetrix, Santa Clara, CA). Prior to hybridization, the cRNA was purified with GeneChip Sample Cleanup Modules (Affymetrix, Santa Clara, CA), and fragmented. Ten micrograms from each experimental sample along with Affymetrix prokaryotic hybridization controls were hybridized for 16 hours to Ca. P. ubique genome arrays (pubiquea) in an Affymetrix GeneChip® Hybridization Oven 640. Affymetrix GeneChip® Fluidics Station 450 was used to wash and stain the arrays with streptavidin-phycoerythrin (Molecular Probes, Eugene, OR), biotinylated anti-streptavidin (Vector Laboratories, Burlingame, CA) according to the standard antibody amplification protocol for prokaryotic targets. Arrays were scanned with an Affymetrix GeneChip Scanner 3000 at 570nm. The Affymetrix prokaryotic hybridization control kit and Poly-A RNA control kit were used to ensure efficiency of hybridization and cRNA amplification. All cRNA was synthesized at the same time. Hybridizations were conducted with one replicate of all times and treatments concurrently. Each array image was visually screened to discount for signal artifacts, scratches or debris. Data is pending submission to the NCBI GEO database under platform number GPL7330.

**Capillary LC-MS Analysis**

Quantitative mass spectrometry measurements were obtained using the identical protocol as in chapter 3: the HPLC system consisted of a custom configuration of 65-mL Isco Model 65D syringe pumps (Isco, Inc., Lincoln, NE), 2-position Valco valves (Valco Instruments Co., Houston, TX), and a PAL autosampler (Leap Technologies, Carrboro, NC), allowing for fully automated sample analysis across four separate
HPLC columns [226]. Reversed-phase capillary HPLC columns were manufactured in-house by slurry packing 3-µm Jupiter C$_{18}$ stationary phase (Phenomenex, Torrence, CA) into a 60-cm length of 360 µm o.d. x 75 µm i.d. fused silica capillary tubing (Polymicro Technologies Inc., Phoenix, AZ) using a 1-cm sol-gel frit (unpublished PNNL variation of [227]) for retention of the packing material. Mobile phase consisted of 0.1% formic acid in water (A) and 0.1% formic acid acetonitrile (B). The mobile phase was degassed by using an in-line Degassex Model DG4400 vacuum degasser (Phenomenex, Torrence, CA). The HPLC system was equilibrated at 10 kpsi with 100% mobile phase A, and then a mobile phase selection valve was switched 50 min after injection, which created a near-exponential gradient as mobile phase B displaced A in a 2.5 mL active mixer. A 40-cm length of 360 µm o.d. x 15 µm i.d. fused silica tubing was used to split ~17 µL/min of flow before it reached the injection valve (5 µL sample loop). The split flow controlled the gradient speed under conditions of constant pressure operation (10 kpsi). Flow through the capillary HPLC column when equilibrated to 100% mobile phase A was ~500 nL/min.

MS analysis was performed using a LTQ Orbitrap Velos ETD mass spectrometer (Thermo Scientific, San Jose, CA) outfitted with a custom electrospray ionization interface. Electrospray emitters were custom made using 150 um o.d. x 20 um i.d. chemically etched fused silica [178]. The heated capillary temperature and spray voltage were 250 ºC and 2.2 kV, respectively. Data was acquired for 100 min, beginning 65 min after sample injection (15 min into gradient). Orbitrap spectra (AGC 1×10$^6$) were collected from 400-2000 m/z at a resolution of 100k followed by data dependent ion trap CID MS/MS (collision energy 35%, AGC 1×10$^4$) and orbitrap ETD MS/MS (activation time 100 ms, AGC 2×10$^5$) of the six most abundant ions. A dynamic exclusion time of 60 sec was used to discriminate against previously analyzed ions.
Mass Spectrometry Data Analysis
Identification and quantification of the detected peptide peaks was performed utilizing
the Accurate Mass and Time (AMT) Tag approach [228]. Briefly, multiple in-house
developed/publicly available informatics tools were used to process LC-MS data and
correlate the resulting LC-MS features to an AMT tag database containing accurate
mass and LC separation elution time information for peptide tags generated from
Pelagibacter ubique proteins. Among the tools used were algorithms for peak-picking
and for determining isotopic distributions and charge states [229]. Further
downstream data analysis incorporated all of the possible detected peptides into a
visualization program VIPER [230] to correlate LC-MS features to the peptide
identifications in the AMT tag database. VIPER provided an intensity report for all
detected features, normalized LC elution times via alignment to the database, and
feature identification. In DAnTE software, peptide peak intensity values were
converted to a log2 scale, normalized with a Central Tendency algorithm, and assessed
at a protein level using Rrollup parameters [231].

Each of the 12 biological samples were measured with quantitative mass spectrometry
in triplicate. The process used to compare samples in chapter 3 was also employed
here: calculating the difference in protein abundance between two conditions was a
three step process. First, an average peptide abundance in each condition was
calculated by averaging together the values for individual peptides across all samples
assigned to a particular condition. Peptides with fewer than three observations in a
condition were marked as "not observed" in that condition. Next, the peptide average
from condition 1 was divided by the peptide average from condition 2, then log_{10}
transformed. Finally, all log_{10} peptide ratios from the same protein were averaged
together.

To represent the likelihood that a protein was equally abundant in both samples, the
multiple peptide measurements were combined into a single statistic as previously
described [180,213]. Briefly, p-values for individual peptides were calculated using a
one-tailed Student’s t-test. A two-tailed Student’s t-test was not used because p-values reflecting a large increase would be indistinguishable from p-values reflecting a large decrease. Instead, peptides which changed in the opposite direction from the protein average were assigned a p-value of 1 for their one-tailed Student’s t-test. All peptide p-values for a single protein were then combined into a single chi-square statistic by using Fisher’s method with a Bonferroni correction:

$$\chi^2 = -2 \times \sum_{i=1}^{n} \ln \left( \left\{ \begin{array}{ll} n \times P_i, & n \times P_i < 1 \\ 1, & n \times P_i \geq 1 \end{array} \right. \right)$$

Where $n$ is the total number of peptides for a given protein and $P_i$ is an individual peptide's p-value as determined by a Student’s t-test. The $\chi^2$ value is then transformed into a p-value using a chi-square probability table and $2n$ degrees of freedom. A Perl script for performing this computation is available as Appendix 2.

Q-values were calculated via the Benjamini-Hochberg method [285] to estimate the false discovery rate arising from multiple comparisons in the microarray and proteomics analyses. In this procedure, each p-value was multiplied by the number of proteins, $T$, and then divided by that p-value’s rank, $k$, among the set of p-values (e.g. the most significant p-value was divided by 1, the second most significant p-value was divided by 2, etc).

$$q-value = \frac{p-value \times T}{k}$$
### Table 3: Transcripts and proteins that were more abundant during nitrogen-limited stationary phase.

The 32 genes in this table showed a minimum fold change of 1.5 supported by a p-value of less than 0.05 in either mRNA or protein. Numbers in bold indicate a change in expression supported by a p-value of less than 0.05. +INF (positive infinity) is used to denote instances where the protein was detected solely in the nitrogen-limited treatment. *Q-value $\leq 0.05$. **Q-value $\leq 1\times10^{-5}$.

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>Gene</th>
<th>Description</th>
<th>Fold Change</th>
<th>mRNA</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR11_0818</td>
<td>amtB</td>
<td>Ammonium transporter</td>
<td>2.72</td>
<td>+INF</td>
<td></td>
</tr>
<tr>
<td>SAR11_1304</td>
<td>sox</td>
<td>Monomeric sarcosine oxidase</td>
<td>2.31</td>
<td>+INF</td>
<td></td>
</tr>
<tr>
<td>SAR11_0747</td>
<td>glnA</td>
<td>Glutamine synthetase</td>
<td>1.91</td>
<td>5.16*</td>
<td>2.35**</td>
</tr>
<tr>
<td>SAR11_0460</td>
<td>acuC</td>
<td>Histone deacetylase-like</td>
<td>1.08</td>
<td>2.65*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0953</td>
<td>yhdW</td>
<td>Amino acid transporter: periplasmic</td>
<td>1.19</td>
<td>2.35**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0279</td>
<td>aspC</td>
<td>Aspartate transaminase</td>
<td>1.42</td>
<td>2.04**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0433</td>
<td>gltD</td>
<td>Glutamate synthase, small subunit</td>
<td>0.44</td>
<td>2.03**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0032</td>
<td>ftsZ</td>
<td>Cell division protein</td>
<td>1.48</td>
<td>1.84**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1365</td>
<td>msrB</td>
<td>Peptide methionine sulfoxide reductase</td>
<td>1.15</td>
<td>1.77**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0434</td>
<td>gltB</td>
<td>Glutamate synthase, large subunit</td>
<td>0.90</td>
<td>1.74**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1210</td>
<td>occT</td>
<td>Octopine/nopaline transporter: periplasmic</td>
<td>1.35</td>
<td>1.61**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0061</td>
<td>pilP</td>
<td>Type 4 fimbrial biogenesis protein</td>
<td>1.23</td>
<td>1.58**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1166</td>
<td></td>
<td>Protein of unknown function</td>
<td>1.12</td>
<td>1.57**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0359</td>
<td>ompU</td>
<td>Outer membrane porin</td>
<td>0.67</td>
<td>1.55**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0733</td>
<td>prfB</td>
<td>Bacterial peptide chain release factor 2 (RF-2)</td>
<td>1.04</td>
<td>1.55*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0321</td>
<td>phhB</td>
<td>Pterin-4-alpha-carbinolamine dehydratase</td>
<td>0.91</td>
<td>1.52**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0807</td>
<td>tauA</td>
<td>Sulfonate/nitrate/taurine transporter: periplasmic</td>
<td>1.30</td>
<td>1.52**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0134</td>
<td>cox1</td>
<td>Cytochrome-c oxidase</td>
<td>1.21</td>
<td>1.51*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0183</td>
<td>hflK</td>
<td>Integral membrane proteinase</td>
<td>1.59</td>
<td>1.45</td>
<td></td>
</tr>
<tr>
<td>SAR11_0162</td>
<td>groEL</td>
<td>Chaperonin GroEL</td>
<td>2.55</td>
<td>1.42**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1285</td>
<td>fhs</td>
<td>Formate-tetrahydrofolate ligase</td>
<td>1.80</td>
<td>1.20*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0267</td>
<td>msmX</td>
<td>Multiple sugar transporter: ATP-binding</td>
<td>1.54</td>
<td>1.19**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0008</td>
<td>hflK</td>
<td>Integral membrane proteinase</td>
<td>2.36</td>
<td>1.13*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0334</td>
<td>hslU</td>
<td>ATP-dependent protease</td>
<td>1.53</td>
<td>1.08*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0077</td>
<td>trxB</td>
<td>Thioredoxin reductase</td>
<td>3.00</td>
<td>0.83**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0142</td>
<td></td>
<td>Pirin; protein of unknown function</td>
<td>1.69</td>
<td>0.82*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0304</td>
<td></td>
<td>Short chain dehydrogenase</td>
<td>1.88</td>
<td>0.79*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0750</td>
<td>mmuM</td>
<td>Homocysteine S-methyltransferase</td>
<td>1.62</td>
<td>0.76**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0504</td>
<td>cycM</td>
<td>Cytochrome c</td>
<td>2.30</td>
<td>0.74*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0680</td>
<td>ldsB</td>
<td>Formate dehydrogenase</td>
<td>1.58</td>
<td>0.64**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1143</td>
<td>griA</td>
<td>Glutaredoxin</td>
<td>1.82</td>
<td>0.61**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0171</td>
<td></td>
<td>Rhodanese-related sulfurtransferase</td>
<td>1.62</td>
<td>0.57**</td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Selected genes related to nitrogen assimilation that either decreased or remained constant in response to nitrogen limitation.

Fold change values in bold indicate that differential expression is supported by a p-value of 0.05 or less. †Peptides for SAR11_1310 were detected in control cultures, but not in nitrogen-limited cultures. *Q-value ≤ 0.05. **Q-value ≤ 1e-5.

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>Gene</th>
<th>Description</th>
<th>Fold Change</th>
<th>mRNA</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR11_1310</td>
<td>amt</td>
<td>Ammonium transporter</td>
<td>1.01</td>
<td>0.00†</td>
<td></td>
</tr>
<tr>
<td>SAR11_0948</td>
<td>ntrX</td>
<td>Nitrogen assimilation regulatory protein</td>
<td>0.91</td>
<td>0.43**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0216</td>
<td>hisC</td>
<td>Histidinol-phosphate aminotransferase</td>
<td>1.04</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td>SAR11_0086</td>
<td>livE</td>
<td>Branched-chain amino acid aminotransferase</td>
<td>0.71</td>
<td>0.56**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1187</td>
<td>gdhA</td>
<td>Glutamate dehydrogenase</td>
<td>0.53</td>
<td>0.57**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0080</td>
<td>aatA</td>
<td>Aspartate transaminase</td>
<td>1.04</td>
<td>0.60**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0829</td>
<td>metC</td>
<td>Cystathionine beta-lyase</td>
<td>1.31</td>
<td>0.60**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0080</td>
<td>aatA</td>
<td>Aspartate transaminase</td>
<td>1.04</td>
<td>0.60**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0534</td>
<td></td>
<td>Aminotransferase</td>
<td>1.09</td>
<td>0.64*</td>
<td></td>
</tr>
<tr>
<td>SAR11_1361</td>
<td>livJ2</td>
<td>Leucine/isoleucine/valine transporter: periplasmic</td>
<td>1.41</td>
<td>0.65**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1308</td>
<td>gilB2</td>
<td>Glutamate synthase, large subunit</td>
<td>1.29</td>
<td>0.66**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1305</td>
<td>glnT</td>
<td>Glutamine synthetase</td>
<td>0.94</td>
<td>0.67**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1151</td>
<td>glmS</td>
<td>Glutamine-fructose-6-phosphate transaminase</td>
<td>1.10</td>
<td>0.70*</td>
<td></td>
</tr>
<tr>
<td>SAR11_1313</td>
<td>gixB</td>
<td>Glutamine amidotransferase</td>
<td>1.08</td>
<td>0.70**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1316</td>
<td>glnT</td>
<td>Glutamine synthetase</td>
<td>0.87</td>
<td>0.72**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1368</td>
<td>dadA</td>
<td>D-amino-acid dehydrogenase, small chain</td>
<td>1.16</td>
<td>0.73**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1317</td>
<td>ygeX</td>
<td>Diaminopropionate ammonia-lyase</td>
<td>0.78</td>
<td>0.73**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1346</td>
<td>livJ</td>
<td>Leucine/isoleucine/valine transporter: periplasmic</td>
<td>0.76</td>
<td>0.79**</td>
<td></td>
</tr>
<tr>
<td>SAR11_0809</td>
<td>ald</td>
<td>Alanine dehydrogenase</td>
<td>1.03</td>
<td>0.83**</td>
<td></td>
</tr>
<tr>
<td>SAR11_1315</td>
<td>gixD</td>
<td>Glutamate synthase, large subunit</td>
<td>0.82</td>
<td>0.83*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0050</td>
<td>amt</td>
<td>Ammonium transporter</td>
<td>1.05</td>
<td>0.88*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0049</td>
<td>amt</td>
<td>Ammonium transporter</td>
<td>1.21</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>SAR11_0502</td>
<td>argD</td>
<td>Acetylornithine aminotransferase</td>
<td>1.03</td>
<td>0.99*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0655</td>
<td>braC</td>
<td>Leucine/isoleucine/valine transporter: periplasmic</td>
<td>1.08</td>
<td>1.02*</td>
<td></td>
</tr>
<tr>
<td>SAR11_0946</td>
<td>ntrY</td>
<td>Nitrogen regulation protein</td>
<td>1.00</td>
<td>1.04</td>
<td></td>
</tr>
<tr>
<td>SAR11_0040</td>
<td>carA</td>
<td>Carbamoylphosphate synthase, small subunit</td>
<td>1.02</td>
<td>1.09</td>
<td></td>
</tr>
<tr>
<td>SAR11_0699</td>
<td>purF</td>
<td>Amidophosphoribosyltransferase</td>
<td>1.32</td>
<td>1.11*</td>
<td></td>
</tr>
</tbody>
</table>
Table 5: Composition of nitrogen-limited artificial seawater media.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Molecular Weight</th>
<th>Nitrogen Limited</th>
<th>Nitrogen Replete</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ammonium</td>
<td>NH$_4$Cl</td>
<td>53.5</td>
<td>None Added</td>
<td>20 µM</td>
</tr>
<tr>
<td>Oxaloacetate</td>
<td>C$_4$H$_8$O$_5$</td>
<td>132.1</td>
<td>500 µM</td>
<td>500 µM</td>
</tr>
<tr>
<td>DMSP</td>
<td>C$_6$H$_5$O$_2$S•HBr</td>
<td>214.1</td>
<td>10 µM</td>
<td>10 µM</td>
</tr>
<tr>
<td>Pyruvate</td>
<td>C$_3$H$_3$NaO$_3$</td>
<td>110.0</td>
<td>500 µM</td>
<td>500 µM</td>
</tr>
<tr>
<td>Sigma #S7653 NaCl</td>
<td></td>
<td>58.4</td>
<td>460 nM</td>
<td>460 nM</td>
</tr>
<tr>
<td>Sigma #M5921 MgSO$_4$•7H$_2$O</td>
<td></td>
<td>246.5</td>
<td>28 mM</td>
<td>28 mM</td>
</tr>
<tr>
<td>Sigma #M2670 MgCl$_2$•6H$_2$O</td>
<td></td>
<td>203.3</td>
<td>27 mM</td>
<td>27 mM</td>
</tr>
<tr>
<td>Sigma #C5080 CaCl$_2$•2H$_2$O</td>
<td></td>
<td>147.0</td>
<td>10 mM</td>
<td>10 mM</td>
</tr>
<tr>
<td>Sigma #P9333 KCl</td>
<td></td>
<td>74.6</td>
<td>9 mM</td>
<td>9 mM</td>
</tr>
<tr>
<td>Sigma #S6297 NaHCO$_3$</td>
<td></td>
<td>84.0</td>
<td>2 mM</td>
<td>2 mM</td>
</tr>
<tr>
<td>Sigma #P9881 KBr</td>
<td></td>
<td>119.0</td>
<td>800 µM</td>
<td>800 µM</td>
</tr>
<tr>
<td>Baker 0084-01 H$_2$BO$_3$</td>
<td></td>
<td>61.8</td>
<td>400 µM</td>
<td>400 µM</td>
</tr>
<tr>
<td>Sigma #S0390 SrCl$_2$•6H$_2$O</td>
<td></td>
<td>266.6</td>
<td>91 µM</td>
<td>91 µM</td>
</tr>
<tr>
<td>Sigma #S1504 NaF</td>
<td></td>
<td>42.0</td>
<td>68 µM</td>
<td>68 µM</td>
</tr>
<tr>
<td>Phosphate</td>
<td>NaH$_2$PO$_4$</td>
<td>120.0</td>
<td>50 µM</td>
<td>50 µM</td>
</tr>
<tr>
<td>myo-Inositol</td>
<td>C$<em>9$H$</em>{12}$O$_6$</td>
<td>180.2</td>
<td>5 µM</td>
<td>5 µM</td>
</tr>
<tr>
<td>Thiamine</td>
<td>C$<em>{12}$H$</em>{17}$N$_4$OS•ClHCl</td>
<td>337.3</td>
<td>5 µM</td>
<td>5 µM</td>
</tr>
<tr>
<td>Niacin</td>
<td>C$_5$H$_6$NO$_2$</td>
<td>123.1</td>
<td>1 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>Pantothenate</td>
<td>(C$<em>3$H$</em>{18}$NO$_5$)$_2$Ca</td>
<td>476.5</td>
<td>1 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>Pyridoxine</td>
<td>C$<em>9$H$</em>{12}$NO$_2$•HCl</td>
<td>205.6</td>
<td>1 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>PABA</td>
<td>C$_7$H$_8$NO$_2$</td>
<td>137.1</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>d-Biotin</td>
<td>C$<em>9$H$</em>{18}$N$_2$O$_5$S</td>
<td>244.3</td>
<td>5 nM</td>
<td>5 nM</td>
</tr>
<tr>
<td>Folic Acid</td>
<td>C$<em>{16}$H$</em>{19}$N$_4$O$_6$</td>
<td>441.4</td>
<td>5 nM</td>
<td>5 nM</td>
</tr>
<tr>
<td>B12</td>
<td>C$<em>{36}$H$</em>{88}$CoN$<em>{14}$O$</em>{14}$P</td>
<td>1355.4</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Iron</td>
<td>FeCl$_2$•6H$_2$O</td>
<td>270.3</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Zinc</td>
<td>ZnSO$_4$•7H$_2$O</td>
<td>287.5</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Molybdenum</td>
<td>Na$_2$MoO$_4$•2H$_2$O</td>
<td>241.9</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Copper</td>
<td>CuSO$_4$•5H$_2$O</td>
<td>249.7</td>
<td>10 nM</td>
<td>10 nM</td>
</tr>
<tr>
<td>Manganese</td>
<td>MnCl$_2$•4H$_2$O</td>
<td>197.9</td>
<td>10 nM</td>
<td>10 nM</td>
</tr>
<tr>
<td>Selenite</td>
<td>Na$_2$SeO$_3$</td>
<td>172.9</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Nickel</td>
<td>NiCl$_2$•6H$_2$O</td>
<td>237.7</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Aluminum</td>
<td>AlK(SO$_4$)$_2$•12H$_2$O</td>
<td>474.4</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Cobalt</td>
<td>CoCl$_2$•6H$_2$O</td>
<td>237.9</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Cadmium</td>
<td>CdCl$_2$•2½H$_2$O</td>
<td>228.3</td>
<td>500 pM</td>
<td>500 pM</td>
</tr>
</tbody>
</table>
Figure 6: Distribution of nitrogen assimilation genes in alphaproteobacteria.

"+" – at least one homolog present in genome, "-" – no homolog detected. Organisms highlighted in red (free-living) and blue (obligate intracellular) are missing both PII genes, *glnB* and *glnK*. Monophyletic groups with the same presence/absence distribution were collapsed, except for SAR11. Bootstrap values are at the nodes, scale bar indicates 0.2 changes per amino acid position.
Figure 7: Proteins for assimilating and metabolizing nitrogenous compounds were up-regulated under nitrogen-limiting conditions.

All proteins illustrated here were more abundant during nitrogen-limiting conditions, compared to a control treatment. Abbreviations: OM, outer membrane; 1304, protein product of gene SAR11_1304 that may catalyze neither, one, or both of the reactions indicated by dashed lines.
Figure 8: Putative enzymatic functions of the highly up-regulated protein SAR11_1304.
Figure 9: Growth curves of nitrogen replete (red) and nitrogen limited (blue) batch cultures of Ca. P. ubique.

Artificial seawater media contained 500 µM oxaloacetate, 10 µM DMSP, 500 µM pyruvate, and trace nutrients (see Table 5). Arrows indicate time points where samples for proteomic and microarray analysis were collected.
CHAPTER 4. REGULATORY RESPONSES OF *CANDIDATUS PELAGIBACTER UBIQUE* TO SULFUR LIMITATION

Daniel P. Smith, Carrie D. Nicora, Paul Carini, Mary S. Lipton, Angela D. Norbeck, Richard D. Smith, & Stephen J. Giovannoni

In preparation for submission to PLoS ONE
ABSTRACT

The alphaproteobacterium *Candidatus* Pelagibacter ubiquique lacks genes for assimilatory sulfate reduction, making it dependent on organosulfur compounds that occur naturally in dissolved organic matter. To investigate how these cells adapt to sulfur limitation, batch cultures were grown in defined media containing either limiting or non-limiting amounts of dimethylsulfoniopropionate (DMSP) as the sole sulfur source. Protein and mRNA expression were measured during exponential growth, immediately prior to stationary phase, and in late stationary phase. Enzymes for acquiring and assimilating sulfur, particularly proteins regulated by S-adenosyl methionine riboswitches (BhmT, MmuM, and MetY), were less abundant in the sulfur-limited cultures. These three methionine synthesis genes increased in all cultures at the onset of stationary phase, then decreased in sulfur-limited cultures. Sulfur-limited stationary phase was also characterized by a 4-8 fold increase in transcription of the heme c shuttle *ccmC* and three small genes of unknown function unique to *Ca. P. ubiquique* (*SAR11_0259*, *SAR11_1163*, and *SAR11_1164*). No membrane transporter genes were up-regulated in response to sulfur limitation, suggesting that this bacterium’s strategy for coping with sulfur stress is intracellular redistribution, rather than increased import, of organosulfur.
INTRODUCTION

*Candidatus* Pelagibacter ubique is one of the few aerobic bacteria unable to incorporate sulfur from dissolved sulfate ($\text{SO}_4^{2-}$) and instead depends on reduced organic compounds, including methionine and dimethylsulfoniopropionate (DMSP), for sulfur [103]. Genome streamlining is believed to be a driving force behind gene loss in this organism, and may have contributed to the absence of an assimilatory sulfate reduction pathway [27,92]. The tradeoff – an increased dependence on organosulfur compounds produced by other members of the plankton community – suggested that these cells might occasionally become sulfur limited.

Large quantities of DMSP are synthesized by marine algae, which use this compound for its antioxidant, osmotic, and predator deterrent properties [232–237]. Lysis of algal cells maintains a DMSP concentration of 1 – 100 nM in the euphotic zone [234,238–240], with a turnover rate of 2 – 28 hours [238,241]. Although not as abundant as $\text{SO}_4$, the surface seawater concentrations of DMSP are theoretically more than sufficient to meet the sulfur requirements of marine microorganisms [103,241–244]. Assimilation of DMSP is common among heterotrophic bacterioplankton and preferred over $\text{SO}_4$ [238,245], though species smaller than 1 µm in diameter (including *Ca. P*. ubique) account for 66 – 100 % of DMSP consumption [241,246,247]. Uptake studies in the natural environment revealed members of the *Ca. Pelagibacter* [110] and *Roseobacter* [112,238] genera to be the primary consumers of DMSP. *Ca. Pelagibacter* isolates are known to degrade DMSP to methylthioacryloyl-CoA, and degradation to methanethiol is theorized [122,248,249]. Genera such as *Roseobacter*, which are not dependent on DMSP sulfur, have been found to discard the sulfur from this compound into the environment in the form of dimethylsulfide (DMS) [112–117]. Because atmospheric DMS originating from oceanic DMSP has been implicated in global climate change [118–121], understanding the fate of marine DMSP through metabolic processes has become increasingly important.
Sulfur limitation has been studied in a variety of bacteria, including *Bacillus* [250,251], *Brevibacterium* [252], *Pseudomonas* [253,254], and *Synechocystis* [255]. All of these species respond to sulfur limitation by up-regulating sulfur import systems and cysteine synthesis pathways. The primary sulfur assimilation strategy in these species is to acquire sulfate and sulfonates from the environment and incorporate them into cysteine. From cysteine, transsulfuration is employed to generate homocysteine, which in turn is methylated by MetH or MetE to produce methionine. Most of the genes up-regulated by these species in response to sulfur limitation are absent from *Ca. P. ubique*’s genome, including both *metH* and *metE*.

Recent studies of DMSP catabolism have led to a greater understanding of how this abundant form of organic carbon and sulfur is utilized by marine bacterioplankton [115,119,248,249,256–258]. The metabolic pathway used by *Ca. P. ubique* to assimilate sulfur from DMSP into biomass begins with demethylation of DMSP by DmdA to yield methylmercaptopyruvate (MMPA). The next three steps in this pathway are catalyzed by DmdB, DmdC, and DmdD to sequentially yield 3-methylmercaptopyruvate-CoA (MMPA-CoA), methylthioacryloyl-CoA (MTA-CoA), and acetaldehydes + methanethiol (Figure 10). In *Ca. P. ubique*, the gene encoding DmdD is currently unknown. The MetY enzyme completes the incorporation of DMSP-sulfur into biomass by condensing methanethiol with O-acetyl homoserine to produce acetate and methionine [259]. The metabolic reconstruction of this organism from genome sequence data [215] indicates that *Ca. P. ubique* employs two additional pathways for methionine biosynthesis in which methyl groups are transferred from glycine betaine to homocysteine by BhmT or from SAM to homocysteine by MmuM, in both cases forming methionine [260,261].

In a wide variety of bacterial species, metabolism of sulfur is regulated by riboswitches [262,263]. In this scheme, S-adenosyl methionine (SAM) binding domains present in the 5' UTR of messenger RNA inhibit expression of downstream genes involved in sulfur metabolism when SAM concentrations are sufficient to meet
cellular demands [264]. In Ca. P. ubique, SAM-V riboswitches that repress translation have been identified in the leader regions of *bhmT*, *mmuM*, and *MetY* [152] and biochemically verified *in vitro* for *bhmT* and *metY* [214]. Another class of SAM riboswitches which effect transcription of downstream genes, SAM-II [265,266], is present in the 5' UTR of *bhmT* and *metX*. A total of 16 characterized riboswitches and 27 putative riboswitches have been computationally identified in *Ca. P. ubique* [152]. The response of these genes to environmental stimuli, however, has yet to be determined with *in vivo* experiments.

The unusual pathways of sulfur metabolism and proliferation of SAM-activated riboswitches in *Ca. P. ubique* prompted us to study changes in mRNA and protein expression in *Ca. P. ubique* strain HTCC1062 in response to sulfur limitation. The observations we report support the conclusion that *Ca. P. ubique* does not activate additional transporter genes for organosulfur compound acquisition when it becomes sulfur-limited. Instead, enzymes for methionine synthesis are differentially expressed in a manner that adapts organosulfur metabolism to the pool of available substrates.

**RESULTS**

Ten cultures of *Ca. P. ubique* were grown in parallel and randomly selected for amendment with either 100 nM DMSP (sulfur-limiting) or 1 µM DMSP (sulfur-replete; control) as the sole sulfur source (Figure 11). On average, sulfur-limited cultures grew to a maximum cell density of 1.2×10^7 cells/mL. Control cultures containing 10x more DMSP grew 3x times higher, with an average maximum density of 3.5×10^7 cells/mL. One control culture was excluded from analysis because its growth rate and maximum density were 1/2 and 1/10 that of the other control cultures. Each of the remaining nine cultures was harvested at three time points: i) exponential/logarithmic growth phase, ii) transitioning from exponential to stationary, and iii) during late stationary phase. One control culture was harvested at two additional pre-stationary phase time points to enable a more detailed time course
analysis. Proteomic and microarray samples were collected simultaneously at each of these 29 harvests.

After analyzing ribosomal protein mRNA expression patterns (see methods), the growth states of two exponential phase samples were re-classified as transitioning and one transitioning sample was re-classified as stationary phase, as indicated in Figure 11B. In the event that a single culture yielded multiple exponential or stationary phase samples, only the first exponential phase or last stationary phase sample was used as a biological replicate for their respective group.

**Correlation between mRNA and Protein**

Examining the relative expression of mRNA and protein from 872 genes across five time points from a single sulfur-replete culture revealed no overarching correlation between mRNA abundance and protein abundance (Figure 12). Sample correlation coefficients for each gene were distributed remarkably evenly from -1.0 (inversely correlated) to 1.0 (directly correlated), indicating that expression values were neither random (clustered near 0) nor inter-dependent (clustered near -1 or 1). This observation is in agreement with pairwise comparisons of samples (e.g. Figure 13B) in which collective mRNA:protein ratios do not form a linear trend.

**Stationary Phase Differences**

Both sulfur-limited (n=5) and control (n=4) cultures were sampled after one week in stationary phase; the statistically significant differences in mRNA and protein abundance between treatments are presented in Table 6. Unexpectedly, proteins for acquiring and assimilating sulfur were less abundant in the sulfur-limited cultures (Table 6). Methyltransferases able to synthesize methionine from homocysteine (bhmT and mmuM) were approximately three times less abundant, and the enzyme for incorporating DMSP-derived sulfur into methionine (metY) was about half as abundant. Messenger RNA transcripts for sulfonate and DMSP transporters (tauA and opuAC) and the adenylyl-sulfate reductase aprA were also 3-4 times less abundant.
Expression of transporters for many non-sulfurous compounds followed a similar trend; mannitol, ammonium, amino acid, and spermidine/putrescine transporters were all less abundant in the sulfur-limited treatment as well. In contrast, mRNA transcripts encoding phosphate acquisition and metabolism genes (pstS, gap, pgk) were three times higher in sulfur-limited cells. Higher expression of mRNA and protein for the DNA repair enzyme RecA was also observed.

The top four transcripts that were more abundant in sulfur-limited stationary phase were ccmC, SAR11_0259, SAR11_1163, and SAR11_1164. Interestingly, none of these genes were detected in the proteome by mass spectrometry. CcmC has been well characterized as catalyzing the transfer of heme c groups to cytochrome c [267,268]. Because heme c molecules contain two sulfur atoms, they are likely to be less abundant in the cell during sulfur-limited conditions. It appears that the five-fold higher expression of ccmC transcripts is a mechanism to compensate for this deficiency and maintain a constant supply of this essential cofactor to cytochrome c. The absence of protein detections for CcmC is not unusual, as insoluble integral membrane proteins such as CcmC are rarely detectable with current mass spectrometry sample preparation techniques.

Three genes encoding hypothetical proteins of unknown function – SAR11_0259, SAR11_1163 and SAR11_1164 – were transcribed four to eight times higher during sulfur-limited stationary phase (Table 7). The proteins products for these transcripts were never detected by mass spectrometry, indicating that these genes may be translationally inhibited or resistant to mass spectrometry preparation techniques. SAR11_1164 has previously been annotated as a putative lipoprotein due to a predicted trans-membrane domain [54,92]. On the chromosome, SAR11_1163 and SAR11_1164 are set apart from neighboring genes by 248 nucleotides downstream and 445 nucleotides upstream, including a putative riboswitch [152]; intergenic distances of this size are particularly conspicuous given that Ca. P. ubique’s median intergenic spacer is only 3 nucleotides. Searching for homologs to these genes using both amino
acid alignment and structural alignment techniques revealed these genes to be unique to *Ca. P. ubique* and therefore may represent a previously unknown class of proteins used to relieve sulfur stress – two of which are also potentially regulated by a novel sulfur-related riboswitch.

**Oxidoreductase: OrdL**

The oxidoreductase *ordL* stands out in this study for increasing 17-fold in transcript abundance while abundance of its protein product did not change significantly. Although changes in the transcriptome are commonly not mirrored in the proteome, a discrepancy of this magnitude is highly unusual and therefore suggestive of a post-transcriptional control mechanism. Such a regulatory element, a cis-acting riboswitch, has previously been predicted in *ordL*'s upstream intergenic region [152]. This study offers the first experimental evidence that metabolite concentrations may affect the expression of *ordL* via the activity of this heretofore un-described riboswitch class.

OrdL belongs to a family of deaminating oxidoreductases which includes PuuB – an *Escherichia coli* enzyme that deaminates \( \gamma \)-L-glutamylputrescine to \( \gamma \)-glutamyl-\( \gamma \)-aminobutyraldehyde in the putrescine degradation pathway [269]. However, the absence of other putrescine degradation pathway enzymes from the *Ca. P. ubique* genome suggests that OrdL may be responsible for catalyzing a different reaction in this organism. Because *ordL* is encoded immediately downstream of *bhmT* and *osmC*, the biological roles of these three proteins may be related. For instance, the BhmT reaction product dimethylglycine may be a substrate for OrdL. Under sulfur-limiting conditions in this speculative model, DMG would be produced at a decreased rate and less OrdL would be required for metabolizing DMG; a DMG-sensing riboswitch could precisely regulate the translation of OrdL in response to fluctuations in DMG concentration, and account for the discrepancy in OrdL expression observed in this study.
Methionine Synthesis

MmuM, MetY, and BhmT—three enzymes in *Ca. P. ubique* responsible for synthesizing the sulfur-containing amino acid methionine—were among the most up-regulated transcripts and proteins throughout this experiment. This observation is in line with previous findings that each of these genes is preceded by a cis-acting riboswitch that couples protein translation to the intracellular concentration of SAM, an organosulfur compound (see introduction). As detailed below however, these three genes exhibited patterns of expression which were much more closely tied to a general stationary phase response than to organosulfur limitation. In sulfur-replete exponential phase the abundance of mRNA and protein for these genes was lowest, and is used as the base level to which all other conditions are compared; the terms “increase” and “decrease” in the following paragraphs refer to the fold change in abundance for a gene’s mRNA or protein product relative to its abundance in sulfur-replete exponential phase. The most significant increases across the entire transcriptome and proteome are summarized in Table 7. Increases in protein abundance for MmuM, MetY, and BhmT are additionally illustrated in Figure 10.

Of these three methionine synthesis genes, *bhmT* increased the most in both transcript and protein abundance as cells neared stationary phase and adapted to non-growth. As illustrated in Figure 10, BhmT catalyzes the transfer of a methyl group from glycine betaine onto homocysteine to form methionine. Assuming equal reaction rates for BhmT, MmuM, and MetY, the preferential expression of BhmT may indicate a shift towards glycine betaine utilization for methionine synthesis during stationary phase when grown in the glycine betaine-rich media used in this experiment (Table 8). A detailed time course of *bhmT* expression under sulfur-replete conditions is presented in Figure 12B, and reveals a sharp increase (29x) in mRNA abundance during late log phase which persists throughout stationary phase. This is in contrast to *bhmT* transcripts in sulfur-limited cultures, which similarly increased towards the end of log phase, but subsequently decreased three-fold as stationary phase progressed. As a result, sulfur-replete cultures exhibited a significantly higher abundance of *bhmT*
mRNA transcripts than sulfur-limited cultures in stationary phase (Table 6). Increases in BhmT protein abundance correlated well with increases in bhmT mRNA transcript abundance (Table 7), indicating that the translation-inhibiting SAM-V riboswitch upstream of bhmT did not noticeably affect differential protein expression over the course of this experiment.

The expression of MetY was similar to BhmT; increases in mRNA abundance correlated well with increases in protein abundance, metY mRNA transcript abundance increased sharply just prior to stationary phase, and MetY protein was twice as high in sulfur-replete stationary phase than in sulfur-limited stationary phase. Examining the five-point time course for MetY expression in the presence of excess sulfur (Figure 12) revealed that this gene has the highest correlation between mRNA and protein (Pearson correlation coefficient = 0.96) among the three methionine synthesis genes. Unlike BhmT and MmuM, which continue to increase in abundance throughout sulfur-replete stationary phase while mRNA abundance remains constant, abundance of protein and mRNA for metY plateaus concomitantly. As illustrated in Figure 10, MetY assimilates DMSP-derived sulfur into methionine. The up-regulation of MetY and BhmT in sulfur-replete stationary phase cultures suggests that catabolism of DMSP and glycine betaine may be preferred sources of energy over pyruvate and glycine, which were also present in the culturing medium (Table 8).

MmuM, the third methionine synthesis gene, also increased suddenly in mRNA abundance at the onset of stationary phase (Figure 12) followed by a subsequent decrease only in the sulfur-limited cultures (Table 7). However, unlike BhmT and MetY, protein expression of MmuM exhibited evidence of post-transcriptional regulation. Sulfur-limited cultures entering stationary phase were characterized by an increase in mmuM mRNA transcript abundance without a concomitant increase in MmuM protein. This is in contrast to sulfur-replete stationary phase, which showed increases in both mmuM mRNA and MmuM protein (Table 7), and was the only condition in which MmuM protein was observed to increase in abundance. This
protein catalyzes the transfer of the S-adenosyl moiety from SAM onto homocysteine to yield methionine and S-adenosyl homocysteine (Figure 10), thereby decreasing the intracellular concentration of SAM and leading to the activation of genes regulated by SAM-II and SAM-V riboswitches (i.e. bhmT, metY, and mmuM). Without additional regulatory mechanisms, this positive feedback system would lead to continual expression of methionine synthesis proteins during sulfur-limited conditions. Therefore, it appears that Ca. P. ubique is able to adapt to an organosulfur deplete environment, and does so in part by inhibiting additional expression of MmuM protein as a strategy for conserving SAM, which is an essential methyl donor in a multitude of diverse reactions.

**Sulfur Content of Proteome Not Reduced**

A study of Pseudomonas putida's response to sulfur starvation identified the expression of low-sulfur-content proteins as a survival mechanism, resulting in a five-fold reduction in cellular thiol [254]. This strategy was not observed in Ca. P. ubique. Methionine and cysteine average 2.7% of amino acid residues encoded by Ca. P. ubique’s genome (3.1% in P. putida); the six proteins observed to be up-regulated during sulfur-limiting conditions in this experiment (positive value in column eight or nine in Table 7) averaged 3.2% met+cys content – a higher-than-average sulfur content.

**DISCUSSION**

Previous experimental work with Ca. P. ubique has revealed that transporters for iron and nitrogen are significantly up-regulated in environments where these nutrients are limiting (chapters 2 & 3; [213]). Sulfur limitation in a variety of other bacteria elicits a similar response dominated by the expression of sulfur acquisition proteins [251–253,255]. Therefore, it was unexpected that proteins for transporting organosulfur compounds were not observed at a higher abundance under sulfur-limiting conditions, particularly the periplasmic binding components of the glycine betaine ABC
transporters (ProX and OpuAC) which are also known to bind DMSP for transport [270].

The results of this experiment suggest a bi-phasic response to sulfur stress. Cells initially entering stationary phase due to sulfur limitation exhibited increased mRNA abundance of *mmuM*, *metY*, *ordL*, *osmC*, and *BhmT*. These transcripts appear to be part of a general stress response, as they are similarly up-regulated during sulfur-replete stationary phase. However, MmuM and OrdL protein abundance does not concomitantly increase with the higher levels of these transcripts present in sulfur-limited cultures, indicating that post-transcriptional regulation may be inhibiting the translation of these mRNA transcripts when organosulfur is scarce. In contrast, all five of these proteins are expressed at a higher abundance in sulfur-replete stationary phase cultures (Figure 10). The second phase of the sulfur limitation response, observed a week after exponential growth ended, was decreased transcript and protein abundances of the five aforementioned genes, and increased abundances of RecA protein and transcripts encoding *recA*, *ccmC*, *SAR11_0259*, *SAR11_1163*, and *SAR11_1164*. This study was unable to determine whether the abundance of CcmC and the hypothetical proteins was constant or varying between conditions due to a complete absence of peptide detections by mass spectrometry, possibly due to translation inhibition or protein insolubility.

**Methionine Synthesis and Stationary Phase**

Up-regulation of the methionine synthesis proteins MetY, MmuM, and BhmT during stationary phase in *Ca. P. ubique* was first observed by Sowell et al [98]. In that study, *Ca. P. ubique* was grown in filtered and sterilized Oregon coast seawater amended with DMSP, NH₄, and PO₄. Similar to the control cultures of this experiment, Sowell’s cultures reached stationary phase well below the density at which DMSP was calculated to become exhausted, confirming our finding that these proteins are expressed in sulfur-replete stationary phase. Furthermore, the expression of these proteins in both natural seawater media and defined artificial seawater media indicates
that this pattern of expression is conserved between very different growth mediums. However, stationary phase cultures from chapters 2 and 3 of this dissertation were not characterized by higher abundances of MetY, MmuM, or BhmT, showing that expression of these genes is not constitutive in stationary phase, but is instead dependent on specific environmental stimuli. Examining the nutrient additions to these four experiments does not readily identify the causative agent for this disparity in expression of methionine synthesis proteins.

**Riboswitches**

The four most up-regulated genes upon entry into stationary phase — *bhmT*, *metY*, and *mmuM* — are downstream of SAM-V class riboswitches that inhibit translation of mRNA into protein by occluding the ribosome binding site when the concentration of SAM is plentiful: $K_D = 15 \, \mu\text{M}$ for *metY* and $K_D = 120 \, \mu\text{M}$ for *bhmT* [214]. *BhmT* is also under the control of a SAM-II riboswitch – previously described as a regulator of methionine and cysteine metabolism in *Bacillus subtilis* and other Gram-positive bacteria [262,263] – which terminates transcription upstream of *bhmT* when SAM rises above a threshold: $K_D = 1.2 \, \mu\text{M}$ [214]. Although tandem riboswitches are not uncommon, the SAM-II/SAM-IV pairing is unique to *Ca. P. ubique* among currently sequenced organisms and relatively rare in the global ocean survey metagenomic dataset [152]. Meyer and colleagues additionally identified putative riboswitches in front of two other genes which we observed to increase in mRNA abundance under sulfur-limiting conditions: *ordL* and *SAR11_1164* [152]. Due to the close association between SAM riboswitches, mRNA expression, and protein expression, it is apparent that functional RNAs play a central role in this organism's adaptive response to a low sulfur environment.

The genes regulated by SAM-sensing riboswitches in *Ca. P. ubique* control the interconversion of organosulfur compounds: methanethiol, methionine, SAM, homocysteine, and S-adenosyl-homocysteine (Figure 10). Cells entering stationary phase up-regulated MetY, MmuM, and BhmT. Examining Table 7, however, reveals
that expression of MmuM may be post-transcriptionally inhibited under sulfur-limited stationary phase as a means to conserve SAM. Also in sulfur-limited stationary phase, transcription of \textit{bhmT} and its downstream neighbors \textit{osmC} and \textit{ordL} appear to be inhibited – possibly by the SAM-II transcriptional attenuator identified in the upstream region of \textit{bhmT}. Overall, these observations lend support to the \textit{in vitro} analyses of SAM-II and SAM-V riboswitches in \textit{Ca. P. ubique} and establish their role for adapting the expression of methionine synthesis genes in stationary phase according to metabolite availability.

\textbf{Organic Hydroperoxidase: OsmC}

\textit{Ca. P. ubique}'s OsmC protein is structurally similar to OsmC in \textit{Escherichia coli}, which has been described as a peroxidase that favors organic hydroperoxides, but also acts on inorganic hydrogen peroxide [271]. As its name implies, \textit{osmC} is induced by osmotic stress in \textit{E. coli} and other species [272]. The up-regulation of this protein in stationary phase has previously been described as a strategy for defending against oxidative damage that occurs from exposure to visible and ultraviolet light in oxic surface waters [98]. \textit{Ca. P. ubique}’s genome also encodes the peroxidase rubrerythrin (\textit{rbr}), but since Rbr relies on an iron-sulfur center [273–275], OsmC may be better suited to low sulfur conditions.

\textbf{Cysteine Synthesis}

In contrast to genes for methionine biosynthesis, genes dedicated to cysteine biosynthesis such as \textit{cysK}, \textit{cysE}, and \textit{metC} were not up-regulated in response to sulfur stress. MetY and MetC in \textit{Ca. P. ubique} have high sequence similarity (e-value = 6e-50 and 2e-45) to Rv1079, a gene in \textit{Mycobacterium tuberculosis} that has been suggested to act as a cystathionine gamma-lyase (e.c. 4.4.1.1) to catalyze the reversible reaction from cystathionine to cysteine [276]. Therefore, we hypothesize that MetY, in addition to synthesizing methionine, may also synthesize cysteine. However, the pathway by which cystathionine is generated remains unclear. An alternative explanation for not observing differential expression of cysteine synthesis
proteins is that these genes may be present at a constitutive level, and are reliant on \textit{bhmT}, \textit{metY}, and \textit{mmuM} for maintaining an adequate supply of methionine to be converted into cysteine.

**Hypervariable Genomic Island**

The loci encoding \textit{SAR11}_1163, \textit{SAR11}_1164, \textit{bhmT}, \textit{osmC}, and \textit{ordL} have previously been identified as part of the hypervariable region HVR3 in \textit{Ca. P. ubique} [54], a region of the genome subjected to a particularly high rate of recombination in an organism already known for its high rate of gene transfer and genome rearrangement [277]. Genomic islands such as HVR3 commonly encode genes conferring increased fitness in a particular environment, and often also contain a higher concentration of novel hypothetical proteins [278]. The dominant marine phototroph \textit{Prochlorococcus} utilizes genomic islands to encode phosphate uptake genes in strains inhabiting phosphate-limited waters [105,106]. Dobrindt et al reviews the use of hypervariable genomic islands by a broad range of species for acquiring genes involved in diverse processes such as iron and sucrose uptake, toxin and antibiotic resistance, and nitrogen fixation [107]. Hypervariable regions are vital for adapting to the particular stresses present in an environment, and in the case of HVR3, for responding to changes in the availability of organic sulfur compounds.

Previous studies of metabolic variability encoded in HVR microbial plankton genomes have reported that genes for phosphorus and nitrogen assimilation are located in these regions and vary among closely related strains (e.g. \textit{Prochlorococcus} [105,106,108,109]). These observations belie the general theories of plankton productivity which emphasizes the importance of competition for macronutrients and their significance in controlling natural populations. The results suggest that during evolution, cellular strategies for obtaining essential nutrients are flexible. To understand whether similar variability among the genes for the acquisition of reduced sulfur compounds exist among SAR11 strains, we surveyed seven SAR11 genomes (one draft – HIMB114; six finished – HTCC1062, HTCC7211, HTCC1002,
HTCC9565, HIMB5, and HIMB59) for genes involved in sulfur metabolism consistent with the observations reported above that the genes for assimilation are in HVR3. We observed considerable variability between strains (Table 9). There was no evidence in any of the strains of complete operons for assimilatory sulfate reduction. A sulfide reductase that functions bi-directionally, either catalyzing the first step of sulfur reduction or alternatively serving as a thiosulfate reductase (AprA) [279] is encoded in the Ca. P. ubique genome.

Studies of other abundant bacterium such as the cyanobacterium Prochlorococcus have reported that genes for assimilation of phosphorus and nitrogen are variable among closely related strains and are frequently found in hypervariable regions. This is particularly noteworthy in marine systems where competition for phosphorus and nitrogen are predicted to be key determinants of success. In the light of this information, the variability of genes for organosulfur assimilation among SAR11 strains does not seem unusual. However, it raises the question of why organisms would be so variable in properties that are so important to their fitness. The distribution of phosphorus and nitrogen uptake genes in Prochlorococcus and phosphorus uptake genes in SAR11 are correlated with environmental variation in macronutrient compounds. If this model also holds true for organosulfur acquisition and metabolism in SAR11, then SAR11 ecotypes may have different organosulfur acquisition strategies depending on the specific habitat they occupy within the marine water column environment.

Comparing the genome of the coastal Oregon isolate Ca. P. ubique HTCC1062 used in this study to an isolate collected from the Sargasso Sea, Ca. P. ubique HTCC7211, revealed the extent of the variability at this genomic island. Interestingly, the bhmT gene from HTCC1062 is more closely related to orthologs in actinobacteria and firmicutes than to bhmT in HTCC7211 (Figure 14). More importantly, the Sargasso Sea strain of Ca. P. ubique is completely missing osmC, SAR11_1163, and SAR11_1164 – the peroxidase and two novel proteins of unknown function. The
absence of osmC, which was induced in stationary phase cells in this experiment, may indicate that the culturing methodology used in this study is not representative of the Sargasso Sea where OsmC is not necessary for adapting to stationary phase. A metaproteomic study of the Sargasso Sea which found a predominance of phosphate transporters in the natural population of Ca. P. ubique [73] supports the theory that nutrients other than organosulfur compounds are that ecosystem's limiting nutrient, and could enable genes such as osmC which are specific to other environmental stressors to be lost through the genome streamlining process [92].

**EXPERIMENTAL PROCEDURES**

**Growth Media and Harvesting**

Artificial seawater (ASW) media was made using previously established protocols [225]. Water, salts, and metals were added to ten 20-L polycarbonate carboys as detailed in Table 8, then autoclaved for ten hours. After cooling to room temperature, carboys were sparged with CO₂ for 20 hours and brought up to 20 L using sterile water to compensate for evaporation due to autoclaving. Vitamins and nutrients were added to each carboy from a stock solution containing either high or low amounts of dimethylsulfoniopropionate and then sparged overnight on air while cooling to 16 °C. Final nutrient concentrations were as follows: 10 µM glycine, 500 nM glycine betaine, 500 µM pyruvate, and 100 nM or 1 µM DMSP. The media was then inoculated with an ASW-adapted culture of Candidatus Pelagibacter ubique HTCC1062 [225] growing exponentially in low-sulfur media. Two carboys, one from each treatment, were inoculated with twice as much inoculum to aid in predicting the growth behavior of the other carboys. However, differences in culture growth rates made this strategy infeasible and these carboys were used instead as normal replicates in their respective treatments. Following inoculation, cultures were incubated at 16 °C with constant air sparging.
Culture growth was tracked daily by staining cells with SYBR Green and counting on a Guava EasyCyte flow cytometer. Samples for microarray and proteomic analysis were taken from each culture at three time points: exponential growth, exponential to stationary transition, and stationary phase. At these time points, $5 \times 10^{10}$ cells were removed to a separate vessel and amended with 10 mg chloramphenicol, 100 µL 500 mM EDTA, and 100 µL 100X Halt Protease Inhibitor Cocktail (Thermo Scientific #78438) per liter of culture. Tangential flow filtration against a Pellicon 2 Mini 30K membrane (Millipore #P2C030C01), reduced the volume of culture to less than 150 mL which was subsequently centrifuged for 1 hour at 20,000 rpm and 0 °C. The pelleted cells were resuspended 1 mL TE and centrifuged again in a single 1.5 mL tube for 40 minutes at 40,000 rpm and 10 °C. After decanting the supernatant, the pellet was stored at -80 °C until proteomic analysis. Microarray samples, which require less biomass, were harvested through a separate protocol simultaneously. Eighty milliliters of culture was collected and centrifuged for 1 hour at 20,000 rpm and 0 °C. Cells were resuspended in 1 mL of RNAprotect Bacteria Reagent (Qiagen #76506) and then centrifuged again in a 1.5 mL tube for 40 minutes at 40,000 rpm and 0 °C. After decanting the supernatant, microarray samples were placed at -80 °C until analysis.

**Messenger RNA Preparation**

Each cell pellet was resuspended in 100 µL TE (1 mM Tris; 1 mM EDTA; pH 8) containing 40 µg lysozyme and incubated at room temperature for 5 minutes. Total RNA was extracted using an RNeasy MinElute Cleanup Kit (Qiagen #74204) according to the manufacturer's instructions, with the exception of an additional wash step with 700 µL Buffer RW1 (Qiagen #1053394) immediately prior to the prescribed washes with Buffer RPE. Eluted RNA was quantified using a NanoDrop spectrophotometer (Thermo Scientific ND-1000) and found to vary over the range of 103 – 794 ng. Thirty-seven nanograms of total RNA from each sample was amplified and labeled using a MessageAmp II-Bacteria RNA Amplification Kit (Ambion #AM1790) and biotin-11-UTP (Ambion #AM8451) according to the MessageAmp
Improved Protocol" Handbook. Polyadenylation, reverse transcription, and second strand synthesis reactions were carried out in half the recommended volumes; the in-vitro transcription reaction was performed in the recommended volume and incubated for 19 hours rather than 14. NanoDrop measurements of amplified RNA showed recovery of 21 – 91 µg per sample.

RNA Labeling and Microarray Processing
RNA integrity screening, Probe synthesis, hybridization and scanning were conducted by the CGRB Core Laboratories at Oregon State University, Corvallis OR. Five micrograms of total RNA was used to generate biotinylated complementary RNA (cRNA) for each treatment group using the One-Cycle Target Labeling protocol (Affymetrix, Santa Clara, CA) from the GeneChip Expression Analysis Technical Manual (701021 Rev. 5). In short, isolated total RNA was checked for integrity and concentration using the RNA 6000 Nano LabChip kit on the Agilent Bioanalyzer 2100 (Agilent Technologies, Inc., Palo Alto, CA). Poly-A RNA control kit RNA and Ca. P. ubique RNA were reverse transcribed using a T7-(dT) 24 primer and Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA) and double stranded cDNA was synthesized and purified with GeneChip® Sample Cleanup Modules (Affymetrix, Santa Clara, CA). Biotinylated cRNA was synthesized from the double stranded cDNA using T7 RNA polymerase and a biotin-conjugated pseudouridine containing nucleotide mixture provided in the IVT Labeling Kit (Affymetrix, Santa Clara, CA). Prior to hybridization, the cRNA was purified with GeneChip Sample Cleanup Modules (Affymetrix, Santa Clara, CA), and fragmented. Ten micrograms from each experimental sample along with Affymetrix prokaryotic hybridization controls were hybridized for 16 hours to Ca. P. ubique genome arrays (pubiquea) in an Affymetrix GeneChip® Hybridization Oven 640. Affymetrix GeneChip® Fluidics Station 450 was used to wash and stain the arrays with streptavidin-phycoerythrin (Molecular Probes, Eugene, OR), biotinylated anti-streptavidin (Vector Laboratories, Burlingame, CA) according to the standard antibody amplification protocol for prokaryotic targets. Arrays were scanned with an Affymetrix GeneChip Scanner 3000 at 570nm. The
Affymetrix prokaryotic hybridization control kit and Poly-A RNA control kit were used to ensure efficiency of hybridization and cRNA amplification. All cRNA was synthesized at the same time. Hybridizations were conducted with one replicate of all times and treatments concurrently. Each array image was visually screened to discount for signal artifacts, scratches or debris. All data was deposited in the NCBI GEO database under accession number GSE31630.

**Time Point Classification**
Harvest time points were selected to provide a minimum of 2×10^{10} cells for mass spectrometry. As a result, the sulfur-limited exponential phase samples were harvested close to stationary phase. In order to ensure that samples categorized as exponential phase are biologically accurate, the expression of *rpsCEGHJLNS*, *rplBCDEFNOPRVWX*, and *fusA* mRNA were taken in account. These 21 genes were selected because they are a near-contiguous loci of ribosomal proteins which all significantly decreased in mRNA expression (p < 0.01) between the exponential phase samples (SL-1, SL-4, SL-7, SL-14, and SL-18; all ≤ 50% maximum cell density) and stationary phase samples (SL-26, SL-27, SL-28, and SL-29) in the control treatment. Transcript RMA expression values from the microarray chips for each gene were first log2 transformed, then normalized to the range 0.0 – 1.0. A sample’s "growth state" was then calculated by averaging all 21 genes' normalized expression value in that sample. E.g. a growth state of 1.0 would indicate that all 21 genes were expressed at each of their maximum observed abundances. The samples from this study clustered into three distinct growth states: 0.91 – 0.80, 0.58 – 0.39, and 0.29 -0.08, which corresponded well to the exponential, transitioning, and stationary labels. This qualitative assessment reclassified two sulfur-limited exponential phase samples as transitioning.

**Mass Spectrometry Sample Preparation**
Each pellet was brought up to 100 µl with 8 M urea (Sigma-Aldrich, St. Louis, MO) and sonicated in a water bath with ice until the pellet went into solution. The samples
were briefly spun and transferred to PCT MicroTube barocycler pulse tubes with 100 µl caps (Pressure Biosciences Inc., South Easton, MA). The MicroTubes were placed in a MicroTube cartridge and barocycled for 10 cycles (20 seconds at 35,000 psi back down to ambient pressure for 10 seconds). All of the material was removed from the MicroTubes and transferred to 1.5 mL micro-centrifuge tubes. A Coomassie Plus (Thermo Scientific, Rockford, IL) assay was used to determine protein concentration. Dithiothreitol (DTT) was added to each sample at a concentration of 5 mM (Sigma-Aldrich, St. Louis, MO) and incubated at 60 °C for 1 hour. The samples were then diluted 10-fold with 100 mM NH₄HCO₃, and tryptic digestion (Promega, Madison, WI) was performed at a 1:50 (w/w) ratio with the addition of 1 mM CaCl₂ to stabilize the trypsin and reduce autolysis. The sample was incubated for 3 hours and cleaned via C-18 solid phase extraction (Supelco, Bellefonte, PA). The samples were dried to 50 µl and assayed with Bicinchoninic acid (Thermo Scientific, Rockford, IL) to determine the final peptide concentration and vialled for MS analysis.

**Capillary LC-MS Analysis**

The HPLC system consisted of a custom configuration of 65-mL Isco Model 65D syringe pumps (Isco, Inc., Lincoln, NE), 2-position Valco valves (Valco Instruments Co., Houston, TX), and a PAL autosampler (Leap Technologies, Carrboro, NC), allowing for fully automated sample analysis across four separate HPLC columns [226]. Reversed-phase capillary HPLC columns were manufactured in-house by slurry packing 3-µm Jupiter C₁₈ stationary phase (Phenomenex, Torrence, CA) into a 60-cm length of 360 µm o.d. x 75 µm i.d. fused silica capillary tubing (Polymicro Technologies Inc., Phoenix, AZ) using a 1-cm sol-gel frit (unpublished PNNL variation of [227]) for retention of the packing material. Mobile phase consisted of 0.1% formic acid in water (A) and 0.1% formic acid acetonitrile (B). The mobile phase was degassed by using an in-line Degassex Model DG4400 vacuum degasser (Phenomenex, Torrence, CA). The HPLC system was equilibrated at 10 kpsi with 100% mobile phase A, and then a mobile phase selection valve was switched 50 min after injection, which created a near-exponential gradient as mobile phase B displaced
A in a 2.5 mL active mixer. A 40-cm length of 360 µm o.d. x 15 µm i.d. fused silica tubing was used to split ~17 µL/min of flow before it reached the injection valve (5 µL sample loop). The split flow controlled the gradient speed under conditions of constant pressure operation (10 kpsi). Flow through the capillary HPLC column when equilibrated to 100% mobile phase A was ~500 nL/min.

MS analysis was performed using a LTQ Orbitrap Velos ETD mass spectrometer (Thermo Scientific, San Jose, CA) outfitted with a custom electrospray ionization interface. Electrospray emitters were custom made using 150 µm o.d. x 20 µm i.d. chemically etched fused silica [178]. The heated capillary temperature and spray voltage were 250 ºC and 2.2 kV, respectively. Data was acquired for 100 min, beginning 65 min after sample injection (15 min into gradient). Orbitrap spectra (AGC 1×10⁶) were collected from 400-2000 m/z at a resolution of 100k followed by data dependent ion trap CID MS/MS (collision energy 35%, AGC 1×10⁴) and orbitrap ETD MS/MS (activation time 100 ms, AGC 2×10⁵) of the six most abundant ions. A dynamic exclusion time of 60 sec was used to discriminate against previously analyzed ions.

Calculation of Protein Abundance

Tandem MS spectra were collected from 99 datasets from a combination of Thermo Electron LTQ instruments and hybrid LTQ-Velos Orbitrap instruments (See instrument methods section). These MS/MS spectra were searched against a 2006 version of the Candidatus P. ubique strain HTCC1062 proteome [92]. The search program SEQUEST [280] was used to identify the peptides, and the MSGF score [281] was used to select peptides with a 1% False Discovery Rate (FDR) for identification and population into the AMT database. All peptides were assigned a normalized reverse-phase LC elution time [282], and the combination of the elution time and accurate mass was matched to LC-MS features from the experimental datasets.
High-resolution MS spectra were collected from duplicate runs for each biological sample on a Thermo Electron LTQ-Velos Orbitrap instrument. The deisotoped monoisotopic mass [229] and elution times were matched to the peptides in the AMT tag database [283] to obtain an identification and the abundance from the MS feature was used in the quantification.

Raw abundances for peptides were log\textsubscript{2} transformed, normalized using EigenMS in which 5 biases were detected and removed, and outliers removed using the open source software tool DanteR [284]. Finally, peptide abundances were mean centered then un-log\textsubscript{2} transformed.

Each of the 29 biological samples were measured with quantitative mass spectrometry in duplicate, producing 58 unique analyses that were binned into six conditions according the method described in "Time Point Classification" above. Calculating the difference in protein abundance between two conditions was a three step process. First, an average peptide abundance in each condition was calculated by averaging together the values for individual peptides across all samples assigned to a particular condition. Peptides with fewer than three observations in a condition were marked as "not observed" in that condition. Next, the peptide average from condition 1 was divided by the peptide average from condition 2, then log\textsubscript{10} transformed. Finally, all log\textsubscript{10} peptide ratios from the same protein were averaged together, and deconvoluted (10\textsuperscript{3}) to yield a single protein fold change ratio. Protein ratios are calculated in this manner when even a single peptide is measured in both conditions. However, at least three different peptides were required before marking a ratio as “∞” or “1/∞” to indicate the detection of a protein in one condition and not another.

To represent the likelihood that a protein was equally abundant in both samples, the multiple peptide measurements were combined into a single statistic as previously described [180,213]. Briefly, p-values for individual peptides were calculated using a one-tailed Student’s t-test. A two-tailed Student’s t-test was not used because p-values
reflecting a large increase would be indistinguishable from p-values reflecting a large decrease. Instead, peptides which changed in the opposite direction from the protein average were assigned a p-value of 1 for their one-tailed Student’s t-test. All peptide p-values for a single protein were then combined into a single chi-square statistic by using Fisher’s method with a Bonferroni correction:

\[ \chi^2 = -2 \sum_{i=1}^{n} \ln \left( \frac{\text{if } n \times P_i < 1 \text{ then } 1 \text{ if } n \times P_i \geq 1} \right) \]

Where \( n \) is the total number of peptides for a given protein and \( P_i \) is an individual peptide’s p-value as determined by a Student’s t-test. The \( \chi^2 \) value is then transformed into a p-value using a chi-square probability table and \( 2n \) degrees of freedom. A Perl script for performing this computation is available as Appendix 2.

Q-values were calculated via the Benjamini-Hochberg method [285] to estimate the false discovery rate arising from multiple comparisons in the microarray and proteomics analyses. In this procedure, each p-value was multiplied by the number of proteins, T, and then divided by that p-value’s rank, k, among the set of p-values (e.g. the most significant p-value was divided by 1, the second most significant p-value was divided by 2, etc).

\[ \text{q-value} = \text{p-value} \times \frac{T}{k} \]

**ACKNOWLEDGEMENTS**

The authors would like to thank PNNL’s mass spectrometer operator Robbie Heegel for his assistance with data collection.
Comparing sulfur-limited (n=5) and sulfur-replete (n=4) stationary phase cultures revealed 40 genes which differed significantly (q-value ≤ 0.05) and by at least threefold in either mRNA or protein expression. Values in the rightmost two columns are computed by dividing the expression in sulfur-limited cultures by the expression in sulfur-replete cultures – e.g. mRNA encoding *hflC* is five times more abundant in sulfur-replete cultures. Missing values indicate that differences in expression were not statistically significant (q-value ≤ 0.05); “ND” is used in instances where peptides were never detected. Bolded genes are downstream of a SAM-V riboswitch; ↑ denotes genes that were more abundant in sulfur-limited cultures.

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>Gene</th>
<th>Description</th>
<th>mRNA</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR11_0007</td>
<td><em>hflC</em></td>
<td>Integral membrane proteinase</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>SAR11_0051</td>
<td><em>copZ</em></td>
<td>Heavy metal transport/detoxification protein</td>
<td></td>
<td>1/∞</td>
</tr>
<tr>
<td>SAR11_0078</td>
<td><em>groEL</em></td>
<td>Chaperonin</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>SAR11_0162</td>
<td><em>ccmC</em></td>
<td>Heme exporter membrane protein</td>
<td>4.74</td>
<td>ND</td>
</tr>
<tr>
<td>SAR11_0368</td>
<td><em>dnaK</em></td>
<td>Chaperonin</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>SAR11_0399</td>
<td><em>rbr</em></td>
<td>Rubrerythrin; peroxidase</td>
<td>0.17</td>
<td>2.16</td>
</tr>
<tr>
<td>SAR11_0585</td>
<td><em>pgk</em></td>
<td>Phosphoglycerate kinase</td>
<td>3.41</td>
<td>1.90</td>
</tr>
<tr>
<td>SAR11_0641</td>
<td><em>recA</em></td>
<td>DNA repair/recombination protein</td>
<td>3.01</td>
<td></td>
</tr>
<tr>
<td>SAR11_0655</td>
<td><em>braC</em></td>
<td>Leu/Ile/Val ABC transporter; substrate binding</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>SAR11_0750</td>
<td><em>mmuM</em></td>
<td>Homocysteine S-methyltransferase</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>SAR11_0756</td>
<td><em>aldA</em></td>
<td>Acetaldehyde dehydrogenase</td>
<td>0.23</td>
<td>0.63</td>
</tr>
<tr>
<td>SAR11_0765</td>
<td></td>
<td>Hypothetical protein</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>SAR11_0794</td>
<td></td>
<td>Hypothetical protein</td>
<td>0.55</td>
<td>0.26</td>
</tr>
<tr>
<td>SAR11_0807</td>
<td><em>tauA</em></td>
<td>Sulfonate/nitrate/taurine ABC transporter; substrate binding</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>SAR11_0818</td>
<td><em>aprA</em></td>
<td>Adenylyl-sulfate reductase</td>
<td>0.32</td>
<td></td>
</tr>
<tr>
<td>SAR11_0837</td>
<td></td>
<td>Hypothetical protein</td>
<td>0.22</td>
<td>0.57</td>
</tr>
<tr>
<td>SAR11_0864</td>
<td></td>
<td>Mannitol TRAP transporter</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>SAR11_0866</td>
<td></td>
<td>Mannitol TRAP transporter</td>
<td>0.33</td>
<td>ND</td>
</tr>
<tr>
<td>SAR11_0882</td>
<td><em>ion</em></td>
<td>Protease, ATP-dependent</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>SAR11_0991</td>
<td></td>
<td>Peptidase family M48</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>SAR11_1093</td>
<td><em>rpoA</em></td>
<td>DNA-directed RNA polymerase</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>SAR11_1129</td>
<td></td>
<td>Hypothetical protein</td>
<td>0.27</td>
<td>ND</td>
</tr>
<tr>
<td>SAR11_1130</td>
<td><em>tufB</em></td>
<td>Translation elongation factor EF-Tu</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>SAR11_1163</td>
<td></td>
<td>Hypothetical protein</td>
<td>4.69</td>
<td>ND</td>
</tr>
<tr>
<td>SAR11_1164</td>
<td></td>
<td>Hypothetical protein</td>
<td>7.67</td>
<td>ND</td>
</tr>
<tr>
<td>SAR11_1172</td>
<td><em>osmC</em></td>
<td>Organic hydroperoxidase</td>
<td>0.23</td>
<td>0.21</td>
</tr>
<tr>
<td>SAR11_1173</td>
<td><em>bhmT</em></td>
<td>Betaine-homocysteine S-methyltransferase</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>SAR11_1179</td>
<td><em>pstS</em></td>
<td>Phosphate ABC transporter; substrate binding</td>
<td>3.35</td>
<td></td>
</tr>
<tr>
<td>SAR11_1274</td>
<td><em>cspL</em></td>
<td>Cold shock DNA-binding protein</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>SAR11_1290</td>
<td><em>opuAC</em></td>
<td>Glycine betaine ABC transporter; substrate binding</td>
<td>0.22</td>
<td>0.63</td>
</tr>
<tr>
<td>SAR11_1305</td>
<td><em>glnT</em></td>
<td>Glutamine synthetase</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>SAR11_1309</td>
<td></td>
<td>Hypothetical protein</td>
<td>0.28</td>
<td>0.52</td>
</tr>
<tr>
<td>SAR11_1310</td>
<td><em>amt</em></td>
<td>Ammonium transporter</td>
<td>0.29</td>
<td>0.37</td>
</tr>
<tr>
<td>SAR11_1311</td>
<td></td>
<td>Hypothetical protein</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>SAR11_1334</td>
<td><em>potC</em></td>
<td>Spermidine/putrescine ABC transporter; permease</td>
<td>0.29</td>
<td>0.50</td>
</tr>
<tr>
<td>SAR11_1361</td>
<td></td>
<td>Leu/Ile/Val ABC transporter; substrate binding</td>
<td>0.23</td>
<td></td>
</tr>
</tbody>
</table>
Table 7: Post-log phase upregulated mRNAs and proteins.

All 13 genes that increase five-fold or more in mRNA or protein abundance after ceasing log phase growth. Values indicate the fold change in expression relative to sulfur-replete log phase growth. Values are only displayed if a difference in expression is supported by a q-value of ≤ 0.05. ND = never detected by mass spectrometry. Bolded genes are immediately downstream of a SAM-V riboswitch. %S = Percentage of sulfur-containing amino acids. † Q-values for these changes were between 0.05 and 0.08.

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>%S</th>
<th>Gene</th>
<th>Description</th>
<th>mRNA fold change</th>
<th>Protein fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Late</td>
<td>One week in</td>
<td>One week in</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Exponential (S-limited)</td>
<td>stationary</td>
<td>stationary</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(S-limited)</td>
<td>(S-replete)</td>
</tr>
<tr>
<td>C134_0181</td>
<td>3.7</td>
<td>ibpA</td>
<td>Heat shock protein</td>
<td>3.69</td>
<td>6.07</td>
</tr>
<tr>
<td>C134_0259</td>
<td>2.7</td>
<td>Hypothetical protein</td>
<td>7.38</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>C134_0287</td>
<td>3.0</td>
<td>ccmC</td>
<td>Heme exporter membrane protein</td>
<td>10.35</td>
<td>2.18</td>
</tr>
<tr>
<td>C134_0641</td>
<td>3.7</td>
<td>recA</td>
<td>DNA repair/recombination protein</td>
<td>7.31</td>
<td>3.11</td>
</tr>
<tr>
<td>C134_0750</td>
<td>2.6</td>
<td>mmuM</td>
<td>Homocysteine S-methyltransferase</td>
<td>5.22</td>
<td>2.58</td>
</tr>
<tr>
<td>C134_1019</td>
<td>2.8</td>
<td>xerD</td>
<td>Integrase/recombinase</td>
<td>8.50</td>
<td>7.63</td>
</tr>
<tr>
<td>C134_1030</td>
<td>2.0</td>
<td>metY</td>
<td>O-acetyl homoserine (thiol)-lyase</td>
<td>7.14†</td>
<td>6.37</td>
</tr>
<tr>
<td>C134_1163</td>
<td>7.2</td>
<td>Hypothetical protein</td>
<td>6.11</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>C134_1164</td>
<td>2.1</td>
<td>Hypothetical protein</td>
<td>11.13</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>C134_1171</td>
<td>2.0</td>
<td>orlL</td>
<td>Oxidoreductase</td>
<td>17.28†</td>
<td>4.88</td>
</tr>
<tr>
<td>C134_1172</td>
<td>1.6</td>
<td>osmc</td>
<td>Organic hydroperoxidase</td>
<td>50.04</td>
<td>9.56</td>
</tr>
<tr>
<td>C134_1173</td>
<td>7.1</td>
<td>bhmT</td>
<td>Betaine-homocysteine S-methyltransferase</td>
<td>33.77</td>
<td>8.91</td>
</tr>
<tr>
<td>C134_1354</td>
<td>2.6</td>
<td>Hypothetical protein</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8: Composition of organosulfur-limited artificial seawater media.

<table>
<thead>
<tr>
<th>Chemical Name</th>
<th>Formula</th>
<th>Molecular Weight</th>
<th>Sulfur Limited</th>
<th>Sulfur Replete</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMSP</td>
<td>C₅H₁₀O₂S•HBr</td>
<td>214.1</td>
<td>100 nM</td>
<td>1 µM</td>
</tr>
<tr>
<td>Glycine</td>
<td>C₂H₇NO₂</td>
<td>75.1</td>
<td>10 µM</td>
<td>10 µM</td>
</tr>
<tr>
<td>Glycine Betaine</td>
<td>C₃H₅NO₂•H₂O</td>
<td>135.2</td>
<td>500 nM</td>
<td>500 nM</td>
</tr>
<tr>
<td>Pyruvate</td>
<td>C₃H₅NaO₃</td>
<td>110.0</td>
<td>500 µM</td>
<td>500 µM</td>
</tr>
<tr>
<td>Ammonium</td>
<td>NH₄Cl</td>
<td>53.5</td>
<td>100 µM</td>
<td>100 µM</td>
</tr>
<tr>
<td>Phosphate</td>
<td>NaH₂PO₄</td>
<td>120.0</td>
<td>50 µM</td>
<td>50 µM</td>
</tr>
<tr>
<td>Sigma #S7653 NaCl</td>
<td></td>
<td>58.4</td>
<td>460 mM</td>
<td>460 mM</td>
</tr>
<tr>
<td>Sigma #M5921 MgSO₄•7H₂O</td>
<td>246.5</td>
<td>28 mM</td>
<td>28 mM</td>
<td></td>
</tr>
<tr>
<td>Sigma #M2670 MgCl₂•6H₂O</td>
<td>203.3</td>
<td>27 mM</td>
<td>27 mM</td>
<td></td>
</tr>
<tr>
<td>Sigma #C5080 CaCl₂•2H₂O</td>
<td>147.0</td>
<td>10 mM</td>
<td>10 mM</td>
<td></td>
</tr>
<tr>
<td>Sigma #P9333 KCl</td>
<td></td>
<td>74.6</td>
<td>9 µM</td>
<td>9 µM</td>
</tr>
<tr>
<td>Sigma #S6297 NaHCO₃</td>
<td></td>
<td>84.0</td>
<td>2 mM</td>
<td>2 mM</td>
</tr>
<tr>
<td>Sigma #P8881 KBr</td>
<td></td>
<td>119.0</td>
<td>800 µM</td>
<td>800 µM</td>
</tr>
<tr>
<td>Baker 0084-01 H₂BO₃</td>
<td></td>
<td>61.8</td>
<td>400 µM</td>
<td>400 µM</td>
</tr>
<tr>
<td>Sigma #S0390 SrCl₂•6H₂O</td>
<td>266.6</td>
<td>91 µM</td>
<td>91 µM</td>
<td></td>
</tr>
<tr>
<td>Sigma #S1504 NaF</td>
<td></td>
<td>42.0</td>
<td>68 µM</td>
<td>68 µM</td>
</tr>
<tr>
<td>Iron</td>
<td>FeCl₂•6H₂O</td>
<td>270.3</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Zinc</td>
<td>ZnSO₄•7H₂O</td>
<td>287.5</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Molybdenum</td>
<td>Na₂MoO₄•2H₂O</td>
<td>241.9</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>Copper</td>
<td>CuSO₄•5H₂O</td>
<td>249.7</td>
<td>10 nM</td>
<td>10 nM</td>
</tr>
<tr>
<td>Manganese</td>
<td>MnCl₂•4H₂O</td>
<td>197.9</td>
<td>10 nM</td>
<td>10 nM</td>
</tr>
<tr>
<td>Selenite</td>
<td>Na₂SeO₃</td>
<td>172.9</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Nickel</td>
<td>NiCl₂•6H₂O</td>
<td>237.7</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Aluminum</td>
<td>AlK(SO₄)₂•12H₂O</td>
<td>474.4</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Cobalt</td>
<td>CoCl₂•6H₂O</td>
<td>237.9</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
<tr>
<td>Cadmium</td>
<td>CdCl₂•2½H₂O</td>
<td>228.3</td>
<td>500 pM</td>
<td>500 pM</td>
</tr>
<tr>
<td>myo-Inositol</td>
<td>C₅H₁₀O₅</td>
<td>180.2</td>
<td>5 µM</td>
<td>5 µM</td>
</tr>
<tr>
<td>Thiamine HCl</td>
<td>C₃H₁₇N₃O₄S•HCl</td>
<td>337.3</td>
<td>5 µM</td>
<td>5 µM</td>
</tr>
<tr>
<td>Niacin</td>
<td>C₃H₆N₂O₂</td>
<td>123.1</td>
<td>1 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>Pantothenate</td>
<td>(C₃H₆N₂O₂)₂Ca</td>
<td>476.5</td>
<td>1 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>Pyridoxine HCl</td>
<td>C₃H₆N₂O₂•HCl</td>
<td>169.2</td>
<td>1 µM</td>
<td>1 µM</td>
</tr>
<tr>
<td>PABA</td>
<td>C₅H₁₀N₂O₂</td>
<td>137.1</td>
<td>100 nM</td>
<td>100 nM</td>
</tr>
<tr>
<td>d-Biotin</td>
<td>C₃H₁₀N₂O₅S</td>
<td>244.3</td>
<td>5 nM</td>
<td>5 nM</td>
</tr>
<tr>
<td>Folic Acid</td>
<td>C₃H₁₃N₃O₆</td>
<td>441.4</td>
<td>5 nM</td>
<td>5 nM</td>
</tr>
<tr>
<td>B12</td>
<td>C₆₃H₉₅CoN₁₄O₁₄P</td>
<td>1355.4</td>
<td>1 nM</td>
<td>1 nM</td>
</tr>
</tbody>
</table>
Table 9: Phylogenetic distribution of genes related to sulfur metabolism.

The numbers in the table indicate the number of orthologs for each gene present in each genome.

| Species                                      | bhmT | metY | mmuM | orL | osmC | ccmC | SAR11_1163 | SAR11_1164 | sarA | cysA | cysC | cysD | cysG | cysH | cysI | cysJ | cysN | cysQ | cysU | cysW | cytW | sbp | serA | serB | serC | metE | metH | aprA |
|----------------------------------------------|------|------|------|-----|------|------|------------|------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Candidatus Pelagibacter ubique HTCC1002      | 1    | 1    | 1    | 1   | 1    | 1    | 1          | 1          | 1    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Candidatus Pelagibacter ubique HTCC1062      | 1    | 1    | 1    | 1   | 1    | 1    | 1          | 1          | 1    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Candidatus Pelagibacter ubique HTCC7211      | 1    | 1    | 1    | 1   | 1    | 1    | 1          | 1          | 1    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Candidatus Pelagibacter ubique HTCC9565      | 1    | 1    | 1    | 1   | 1    | 1    | 2          | 1          |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Alphaproteobacterium sp. SAR11 HIMB114       | 1    | 1    |      | 1   |      |      | 1          | 2          | 1    |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Alphaproteobacterium sp. SAR11 HIMB59        | 1    | 1    | 1    | 1   | 1    | 1    | 1          | 1          |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Alphaproteobacterium sp. SAR11 HIMB59        | 3    | 2    | 1    | 2   | 1    | 1    | 1          | 1          |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| Alphaproteobacterium sp. SAR11 HIMB59        |      |      |      | 2   | 1    | 1    | 1          | 2          |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
**FIGURES**

**Figure 10: Effect of stationary phase on methionine metabolism.**

MetY, BhmT, and MmuM work independently to produce methionine, a key organic sulfur compound. Red lines indicate statistically significant (p-value \( \leq 0.05 \)) increased protein expression relative to sulfur-replete exponential phase, with line thickness representing the magnitude of the change. Dashed green lines denote inhibition of translation by SAM-sensing riboswitches. Glycine betaine and DMSP are underlined to indicate their availability in the growth medium. Abbreviations: DMSP, dimethylsulphoniopropionate; SAHc, S-adenosyl homocysteine; SAM, S-adenosyl methionine. *Comparison is between treatment groups. Other arrows indicate comparisons of the same replicate cultures at different time points.
Yields of Ca. P. ubique cells were dependent on initial organosulfur concentrations (A). Sulfur-limited cultures ceased exponential growth near the theoretical maximum density for the DMSP-sulfur provided; whereas control cultures with 10 times as much organosulfur yielded only three times more cells. Replicate cultures of each treatment were harvested independently at multiple time-points for proteomic and transcriptomic analysis (B). Patterns in transcript abundance enabled the classification of time-point samples into exponential growth (e), transitioning (t), and stationary phase (s) categories.
Figure 12: Protein abundance does not depend on mRNA abundance.

(A) Protein and mRNA abundance was analyzed at five time points from a single sulfur-replete culture of Ca. P. ubique. (B) Protein and mRNA of selected genes varied in correlation from -0.93 to +0.99. (C) Plotting all genes according to their level of protein-to-mRNA correlation shows that larger fold changes in expression are not better correlated. *0529 and 1265 are abbreviations for genes SAR11_0529 and SAR11_1265.
Figure 13: Genes initially induced as sulfur became limiting.

Genes downstream of S-adenosyl methionine riboswitches (A) were up-regulated in both mRNA and protein (B). Proteins expressed at least five-fold higher under sulfur limiting conditions are bolded. All 133 genes in (B) were significantly different (p-value ≤ 0.05) between the sulfur-limited and control treatments. OsmC and SAR11_1354 peptides were not detected in control treatments. The inner box in (B) encloses genes with less than a five-fold change.
Figure 14: The bhmT gene in *Candidatus Pelagibacter ubique* shows evidence of horizontal gene transfer.

The *bhmT* gene from three strains of *Ca. P. ubique* – HTCC1062, HTCC1002, and HTCC7211 – do not form a single clade, and are instead bisected by *bhmT* genes from actinobacteria and firmicutes. Beginning with the *bhmT* protein sequence from *Ca. P. ubique* HTCC7211, 481 similar sequences were selected from NCBI’s nr databases using ten iterations of psi-blast. These were aligned with MUSCLE, and a bootstrapped maximum likelihood tree constructed by invoking RaxML with the following parameters: `-m PROTCATDAYHOFF -f a -# 500 -x 9078563412`. 
CHAPTER 5. GENERAL CONCLUSIONS

This work is the first to describe how the transcriptome and proteome of Candidatus Pelagibacter ubique changes in response to iron-, nitrogen-, and organosulfur-limited conditions. Data and results derived from these experiments expands the interpretive power of metagenomic, metatranscriptomic, and metaproteomic surveys of natural marine communities that are increasingly being used to further our understanding of interactions between microbes and Earth’s biogeochemical processes. In agreement with previous studies, this work concludes that nutrient transporters such as those for ammonium (AmtB) and iron (SfuC) dominate the transcriptomes and proteomes of cells limited by these nutrients, thereby supporting the use of these genes as biomarkers for nutrient limitation in environmental gene expression surveys. However, despite the fact that dramatic increases in AmtB and SfuC were equally prominent in both the transcriptome and proteome of nutrient-limited Ca. P. ubique cells, overall correlation between transcript and protein abundances, even for highly variable transcripts, was generally observed to be non-existent. This finding is common among integrated transcriptomics/proteomics studies (see chapter 1) and adds more weight to the argument that transcriptomic analyses are severely limited in their ability to describe cellular physiology defined by protein interactions. This decoupling of translation from transcription may stem from expression of RNA-binding proteins such as CspL and CspE, or occur through riboswitch structures on mRNA that regulate translation based on metabolite concentrations; this work presents evidence that both systems play an active role in Ca. P. ubique’s adaption to iron, nitrogen, and organosulfur availability.

In vivo and in situ proteomic studies are highly complementary; the former can be employed to thoroughly monitor a response to a controlled stimulus and the latter to detect the most abundant components of a natural population’s steady-state physiology. Thus, one of the primary contributions of this in vivo study was measuring the changes in protein expression across a large percentage of the proteome. Based on relative increases and decreases of enzymes involved in metabolic pathways, this work
was able to propose the routes by which nutrients are routed into biomass under specific nutrient-limiting conditions (Figure 7; Figure 10). These physiological aspects of the cell are important for understanding the fates of dissolved nutrients – fates that may change as the availability of nutrients change. In these experiments, limiting amounts of nitrogen, sulfur, and iron appear to be directed into glutamate, methionine, and iron-sulfur centers, respectively.

Several proteins of unknown function were significantly up-regulated in each nutrient limitation experiment. This finding is consistent with analyses of gene expression in natural marine assemblages (see chapter 1), where unknown/hypothetical genes were sometimes the most abundant transcripts or proteins in the environment. As a result of this study, several previously uncharacterized genes can be associated with the regulatory response for a particular nutrient stress: SAR11_0144, SAR11_0401, SAR11_0599, SAR11_0785, SAR11_0793, SAR11_1233, SAR11_1239, and SAR11_1279 increase in response to iron stress (Table 1); SAR11_1166 increases in response to nitrogen stress (Table 3); SAR11_0259, SAR11_1163, and SAR11_1164 increase in response to organosulfur stress (Table 6). Identification of the conditions under which these genes are induced not only increases their informativeness when observed in future experiments, but also defines a starting point for future studies that seek to characterize the functional role of these proteins in Ca. P. ubique and other bacteria.

In summary, this dissertation has investigated how a widely successful marine bacterium adapts to nutrient-limited environments. The experimental results presented here implicate novel gene regulatory mechanisms and identify specific transporters and metabolic pathways that are differentially expressed upon depletion of iron, nitrogen, and organosulfur. These results are key to interpreting the growing collection of marine meta-omic data, as well as for understanding how anthropogenic additions of nutrients to the environment influence larger biogeochemical cycles.
BIBLIOGRAPHY


APPENDICES
APPENDIX 1. PERL SCRIPT FOR GENERATING RADIAL VISUALIZATION GENE PLOTS

#!/usr/bin/perl

# Description: Makes a Radial Visualization Plot
#
# Author: Daniel Patrick Smith
# Version: 2.1 / June 14th, 2010
# Affiliation: Giovannoni Laboratory, Oregon State University
# License: GNU General Public License v3
#
# Reference: Smith, D.P. et al. Transcriptional and Translational
# Regulatory Responses to Iron Limitation in the Globally
# Distributed Marine Bacterium Candidatus Pelagibacter
#
use strict;
use warnings;

use PostScript::Simple;
use constant PI => 3.1415926535897932384626433832795;
use constant IMG_SIZE => 450;
use constant RAYLEN => 200;
use constant CENTER => 225;

my $file_in = shift;
die qq{USAGE: $0 input.txt
INPUT FORMAT:
Header line is of the format [sample_name]\@[angle]
Angle 0 is at x=1, y=0; Angles proceed counter-clockwise.
Following lines begin with the gene identifier, then
values for that gene in each of the samples.
Any whitespace(s) is considered a delimiter.
Example:
-----------------------------------------------------------
+T1@0  +T2@860  -T2@120  -T1@150  -T0@200  +T0@280
-----------------------------------------------------------
OUTPUT:
An image will be saved to input.txt.eps
A list of gene coordinates will be saved to input.txt.list

} unless (length($file_in) && -r $file_in);

# Create a New PostScript Object (For Handling Drawing Commands)
#------------------------------------------
my $p = new PostScript::Simple(xsize => IMG_SIZE, ysize => IMG_SIZE,
                           colour => 1, eps => 1, units => "pt");
# Inscribe the Bounding Circle

$p->setlinewidth(1);
$p->setcolour('grey60');
$p->circle(CENTER, CENTER, RAYLEN);

# Read In & Draw Dimensional Anchors - Open Circles w/ Labels

$p->setlinewidth(1.5);
$p->setcolour('black');
$p->setfont("ArialBoldItalic", 10);
open (FILE, $file_in) or die "$!
";
my $DA_line = <FILE>;
my @DA = ();
while ($DA_line =~ m/\s*(\S.*\S)\[\s\n\]*$/g) {
    my ($name, $angle) = ($1, $2);
    # Transform angles into (x,y) coordinates
    my $radians = $angle * PI / 180;
    my ($x, $y) = (cos($radians), sin($radians));
    push @DA, { 'name' => $name, 'x' => $x, 'y' => $y };
    # Temporary scaling and recentering for plotting
    $x = $x * RAYLEN + CENTER;
    $y = $y * RAYLEN + CENTER;
    $p->circle($x, $y, 5);
    # Add the text label next to the DA's circle
    my $aln;
    if ($x < CENTER) { $aln = "right"; $x -= 8; } else { $aln = "left"; $x += 6; }
    $p->text( { 'align' => $aln }, $x, $y - 3, $name );
}

# Calculate the (x,y) Coordinates For Each Gene, Using (0,0) as the Origin

my @final_pts = ();
while (my $line = <FILE>) {
    # Remove leading/trailing whitespace before splitting
    $line = $1 if ($line =~ m/^\s*(\S.*\S)[\s\n]*$/);  
    my @values = split (/\s+/, $line);
    my $pt = { 'x' => 0, 'y' => 0 };  
    $pt->{name} = shift @values;  
    $pt->{size} = (max(@values) - min(@values)) || next;
    foreach my $i (0..($#values - 1)) {
        foreach my $j ($i+1..$#values) {
            $pt->{'x'} += ($DA[$i]->{'x'} - $DA[$j]->{'x'}) * ($values[$i] - $values[$j]) / $pt->{'size'};
        }
    }
}
$pt->('y') += (DA[$i]->('y') - DA[$j]->('y')) * (values[$i] - values[$j]) / $pt->('size');

push @final_pts, $pt;
}
close FILE;

# Output the Raw Coordinates, then Re-Center, Scale, and Plot Them
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
my $max = max(map(max(abs($_->('x')), abs($_->('y'))), @final_pts));

open (FILE, ">$file_in.list") or die "$!
";
print FILE "Gene\tx\ty\tsize\n";
$p->setcolour('black');
foreach my $pt (@final_pts) {
  my $x = $pt->('x') / $max;
  my $y = $pt->('y') / $max;

  print FILE $pt->('name') . "\t\t$y\t$y\t" . $pt->('size') . "\n";

  $x = $x * RAYLEN + CENTER;
  $y = $y * RAYLEN + CENTER;
  $p->circle({ 'filled' => 1 }, $x, $y, $pt->('size'));
}
close FILE;

# Write the Graphical Output to a File
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
$p->output("$file_in.eps");

# Helper Functions
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
sub max {
  my $max = shift;
  foreach (@_) { $max = $_ if ($_ > $max); }
  return $max;
}
sub min {
  my $min = shift;
  foreach (@_) { $min = $_ if ($_ < $min); }
  return $min;
}
APPENDIX 2. PERL SCRIPT FOR COMBINING PEPTIDE MEASUREMENTS

#!/usr/bin/perl

# Description: Combines multiple peptide readings into protein-centric data
# Author: Daniel Patrick Smith
# Version: 2.0 / November 29th, 2011
# Affiliation: Giovannoni Laboratory, Oregon State University
# License: GNU General Public License v3

use strict;
use warnings;
use Statistics::Distributions;

my $file = "Peptides.txt";

# Data File - tab-delimited peptide readings.
# 1st column is the peptide sequence (ignored)
# 2nd column is the protein ID
# Remaining columns are abundance measurements for each sample
# 1st row contains sample IDs for each sample
# Technical replicates should be given the same sample ID

my %sets = (Le => ["SL-2", "SL-3", "SL-5"],
             Lt => ["SL-6", "SL-8", "SL-9", "SL-10", "SL-11", "SL-12"],
             Ce => ["SL-1", "SL-4", "SL-7", "SL-14"],
             Ct => ["SL-15", "SL-16", "SL-17", "SL-25"],
             Cs => ["SL-26", "SL-27", "SL-28", "SL-29"]);

my $sets = {
    "Le" => ["SL-2", "SL-3", "SL-5"],
    "Lt" => ["SL-6", "SL-8", "SL-9", "SL-10", "SL-11", "SL-12"],
    "Ce" => ["SL-1", "SL-4", "SL-7", "SL-14"],
    "Ct" => ["SL-15", "SL-16", "SL-17", "SL-25"],
    "Cs" => ["SL-26", "SL-27", "SL-28", "SL-29"]
};
Sets are used to group biological replicates together into treatments. Above, "Le" and "Cs" represent "Limited exponential" and "control stationary", respectively. However, any treatment identifier may be used.

my @compare = ("Ce", "Le"), ("Ce", "Ls"), ("Ce", "Cs"), ("Cs", "Ls");

These are the treatments that should be compared. E.g., the first pair, "Ce", "Le" will compute the fold change in expression between those two treatments by dividing the abundance in one by the other: Le / Ce

Read in peptide abundances from user-specified file

my @raw_data = ();
open(FILE, $file) or die ("Can't open file '$file': $!");
my $line = <FILE>;
chomp $line;
my @headers = split("\t", $line);
while ($line = <FILE>) {
  chomp $line;
  my @f = split("\t", $line);
  my $pep_readings = {'protein' => $f[1]};
  push @$pep_readings->{@headers[$_]}{$_} foreach 2..$#f;
  push @raw_data, $pep_readings;
}
close FILE;

Average peptides together by treatments, set undef if >=50% are NA

my @avg_data = ();
my $prot2idx = {};
foreach my $raw (@raw_data) {
  my $new = {};
  $new->{'protein'} = $raw->{'protein'};
  while (my ($cond, $SL_ID_arr) = each $sets) {
    my @vals = ();
    push @vals, @$raw->{$_} foreach @{$SL_ID_arr};
    @vals = grep(/\^[0-9.]+$/{$_}, @vals);
    if (scalar(@vals) < 3) {
      $new->{$_} = undef;
      $raw->{$_} = undef;
    } else {
      $new->{$_} = 2 ** avg(@vals);
      $raw->{$_} = @vals;
    }  
  }
APPENDIX 2 (Continued)

}]
push @{$prot2idx->{@raw->{'protein'}}}, scalar(@avg_data);
push @avg_data, $new;
}

# Output headers
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
print "Protein ";
foreach my $cond_pair (@compare) {
  my ($ref_condition, $exp_condition) = @{$cond_pair};
  print "\t" . $ref_condition . "->" . $exp_condition . "\tP-Value\t";
}
print "\n";

# Calculate fold changes in protein abundances
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
my @protlist = sort { $a cmp $b } (keys %{$prot2idx});
foreach my $protein (@protlist) {
  print $protein;

  # Iterate over each pair of treatments specified by the user
  #-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
  foreach my $cond_pair (@compare) {
    my ($ref_condition, $exp_condition) = @{$cond_pair};
    my @ratios = ();
    my @ttests = ();
    my $expONLY = 0;
    my $refONLY = 0;
    my $fold_change = undef;
    my $p_value = 0;

    # Loop over each peptide for the current protein
    #-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
    foreach my $idx @{$prot2idx->{($protein)}} {
      my $ref_val = $avg_data[$idx]->{$ref_condition};
      my $ref_arr = $raw_data[$idx]->{$ref_condition};

      my $exp_val = $avg_data[$idx]->{$exp_condition};
      my $exp_arr = $raw_data[$idx]->{$exp_condition};

      # If the peptide was detected in both treatments,
      # calculate the ratio (fold change).
      #-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
      if ($ref_val && $exp_val) {
        if ($exp_val < $ref_val) {
          ($exp_arr, $ref_arr) = ($ref_arr, $exp_arr);
        }
        push @ratios, log10($exp_val / $ref_val);
        push @ttests, ttest($exp_arr, $ref_arr);
      }
    }

    # Keep track of peptides only occurring in one treatment

APPENDIX 2 (Continued)

#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
elsif ($ref_val) { $refONLY++; }
elsif ($exp_val) { $expONLY++; }
}

# Loop over each peptide for the current protein
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
if (@ratios) {
    # Set p-value to 1 where the ratio is opposite the average
    my $avg_ratio = avg(@ratios);
    foreach (0..$#ratios) {
        $ttests[$_].= 1 if ($ratios[$_] > 0 xor $avg_ratio > 0);
    }

    $p_value = $ttests[0];
    $fold_change = sprintf("%.2f", (10 ** $avg_ratio));
    if (scalar(@ttests) > 1) {
        # Bonferroni correction
        foreach (0..$#ttests) {
            $ttests[$_].= $ttests[$_]. * scalar(@ttests);
            $ttests[$_].= 1 if ($ttests[$_]. > 1);
        }

        # Combine multiple p-values into one
        my $df = 2 * scalar(@ttests);
        my $chi_sq = 0;
        $chi_sq += log($_) foreach (@ttests);
        $chi_sq *= -2;
        $p_value = Statistics::Distributions::chisqprob($df,$chi_sq);
    }
}

# Report instances where the peptides for a given protein were
# only detected in one treatment
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
elsif ($expONLY > $refONLY && $expONLY >= 3) {
    $fold_change = "INF";
    $p_value = "O*";
}
elsif ($expONLY < $refONLY && $refONLY >= 3) {
    $fold_change = "0";
    $p_value = "0*";
}

# Also report instances where the peptides for a given protein were
# never detected in either treatment
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
else {
    $fold_change = "ND";
    $p_value = "1*";
}
APPENDIX 2 (Continued)

print "\t$fold_change\t$p_value";
}

print "\n";

# Two-tailed Student's t-test calculation
# Statistical Sleuth, 2nd ed. pg 44
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
sub ttest {
    my @i_arr = @{$_[0]};
    my @j_arr = @{$_[1]};
    my $i_num = scalar(@i_arr) || return 1;
    my $j_num = scalar(@j_arr) || return 0;
    my $i_df = ($i_num - 1) || return 1;
    my $j_df = ($j_num - 1) || return 0;
    my $i_ave = 0; $i_ave += $_ foreach @i_arr; $i_ave /= $i_num;
    my $j_ave = 0; $j_ave += $_ foreach @j_arr; $j_ave /= $j_num;
    return 1 unless ($i_ave > $j_ave);
    my $i_sd = 0; $i_sd += ($_ - $i_ave) ** 2 foreach @i_arr;
    $i_sd = sqrt($i_sd / $i_df);
    my $j_sd = 0; $j_sd += ($_ - $j_ave) ** 2 foreach @j_arr;
    $j_sd = sqrt($j_sd / $j_df);

    my $pooled_df = $i_df + $j_df;
    my $pooled_sd = ($i_df * ($i_sd ** 2)) + ($j_df * ($j_sd ** 2));
    $pooled_sd = sqrt($pooled_sd / $pooled_df);
    my $pooled_se = $pooled_sd * sqrt((1 / $i_num) + (1 / $j_num));
    return 1 unless ($pooled_se != 0);
    my $t_statistic = ($i_ave - $j_ave) / $pooled_se;
    my $p_value = Statistics::Distributions::tprob($pooled_df, $t_statistic);
    return $p_value;
}

# Helper Functions
#-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
sub avg {
    my $sum = 0;
    $sum += $_ foreach @_; 
    my $avg = scalar(@_) ? $sum / scalar(@_) : 0;
    return $avg;
}
sub log10 {
    my $n = shift;
    return log($n)/log(10);
}