Western Washington University
Western CEDAR

WWU Graduate School Collection

WWU Graduate and Undergraduate Scholarship

Winter 2016

Microbial Diversity Across an Oxygen Gradient Using Large-scale Phylogenetic-based Analysis of Marine Metagenomes

Ryan J. (Ryan James) McLaughlin
Western Washington University, mclaugr4@students.wwu.edu

Microbial diversity across an oxygen gradient using large-scale phylogenetic-based analysis of marine metagenomes

By

Ryan James McLaughlin

Accepted in Partial Completion

Of the Requirements for the Degree

Master of Science

Kathleen L. Kitto, Dean of the Graduate School

ADVISORY COMMITTEE

Chair, Dr. Robin Kodner, Department of Biology

Dr. Craig Moyer, Department of Biology

Dr. Perry Fizzano, Department of Computer Science

MASTER'S THESIS

Signature: Ryan James McLaughlin

Date: February 17th, 2016

Microbial diversity across an oxygen gradient using large-scale phylogenetic-based analysis of marine metagenomes

A Thesis

Presented to

The Faculty of

Western Washington University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

Ryan James McLaughlin

February 2016

TABLE OF CONENTS

ABSTRACT

Insufficient reference sequence data for annotation of unknown environmental sequences and metagenomes has driven efforts to find alternative annotation methods that mitigate biases from missing information. The use of phylogenetic-placement algorithms shows promise as a robust sequence annotation technique that deals with missing reference information by allowing for annotation of sequences at internal nodes of a phylogenetic tree. However, using these methods for community level surveys of the thousands of genes found in metagenomes requires powerful computational systems and sophisticated software workflows. The main goal of this thesis is to outline a phylogenetic analysis pipeline built to process environmental metagenomic samples using the pplacer software suite, and a pilot study performed with this software pipeline to investigate community-level patterns in gene diversity for a marine oxygen minimum zone (OMZ) off the coast of Chile, South America. Reference sequence data was used to create a custom database and custom reference packages for 9,204 functional housekeeping genes, along with small sub-unit ribosomal genes (SSU) by Domain. A comparative analysis of metagenomic samples from the OMZ using our pipeline shows that while functional and SSU genes show similar spatial patterns of diversity across the oxygen gradient, higher overall diversity was identified via the functional genes. Ecologically relevant functional genes showed higher levels of diversity than either the total from all functional genes or SSU ribosomal genes, underlining the importance of diversity in ecosystem functions.

ACKNOWLEDGMENTS

My thesis project would not have been possible without support of some key individuals throughout my time at Western Washington University (WWU). First and foremost I would like to thank my advisor Dr. Robin Kodner for taking me on as her first graduate student. She challenged my abilities at each step, while fostering a positive, inclusive, and nurturing work environment. I appreciated that she did not simply answer my questions, but rather taught me how to answer my own and to collaborate with fellow academics. To Eric Hervol, thank you for continuing to pass on your decades of programming experience on to make me a better data scientist and helping me both optimize and clean up my code for this project. To Dr. Perry Fizzano for letting me bounce ideas off him to develop my final project idea. To Dr. Craig Moyer for his help getting up-to-date with microbial literature. To all those in biology stockroom for always being available to answer my questions and help me get the equipment for my teaching and research. Thank you to Mary Ann Merrill and supporting biology office staff for being another vital resource throughout my project. Thank you to the WWU Computer Science Department cluster staff for their help with my special computer requirements and responsive feedback. To my fellow graduate students for their unbelievable support. A special thanks to Andrew Rothstein, Kelsey Jesser, and Erin D’Agnese for their advice, friendship, and fantastic times. I especially want to thank my partner Jenna Brooks, whom I met here at WWU, for her unyielding encouragement and support. Finally, I thank my parents and my brother and sister, James, Patricia, Nathan, and Aimee McLaughlin. As a family they have been infinitely supportive of my scientific exploits. Without their guidance, support, and love throughout my life I would not have had the opportunities to pursue my passions, particularly my love of science.

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

INTRODUCTION

The contributions and overall importance of microbial organisms to marine ecosystem function is well established (Sunagawa et al (2015), Fuhrman (2009)). However, the intricacies of their evolutionary relationships and full extent of functional diversity remain largely under-characterized (Vargas et al (2015), Rusch et al (2007), Venter et al (2004)). This gap in our understanding has narrowed in the last several decades with advances in sequencing and computer technologies (Armbrust and Palumbi (2015), Sunagawa et al (2015), Vargas et al (2015), Villar et al (2015), Lima-Mendez et al (2015), Iverson et al (2012)). However, work in building computational methods for community-wide remote homology detection of functional genes and quantification of their overall contribution to ecosystem biodiversity is an ongoing field of research.

High-throughput sequencing using next generation sequencing (NGS) platforms has become common practice when characterizing the microbial community in an environment. NGS systems are capable of producing extremely large sequence libraries, 10^6-10^9 reads of 100-700 base-pairs in length per run (Logares et al (2012)). Application of NGS to environmental DNA samples has led to the emergence of a new type of genomic sequence data, a metagenome, and field of study, environmental metagenomics. Creating a metagenome forgoes isolation and cultivation techniques used by targeted sequencing methods, resulting in an unbiased data-set containing sequences from the entire community. These methods are advantageous for surveying under-characterized microbial assemblages; however they require sophisticated computational analysis pipelines in order to analyze the large and complex data-sets.

The field of bioinformatics has responded to the ever-growing biological sequence data by producing a multitude of software pipelines capable of robust and efficient data handling, processing, and annotation. Typically, processing a metagenome requires multiple steps in order to address a research question. It is necessary to build these steps into an analysis pipeline, executing each step consecutively and automatically. This allows for the larger-scale application of a method on diverse data-sets. In the last several decades, numerous annotation methods have been developed and implemented in pipelines to analyze metagenomic data. Many of these pipelines are capable of performing large-scale taxonomic and functional annotations, some examples of pipelines include: MG-RAST, CARMA, MEGAN, and QIIME (Meyer et al (2008), Krause et al (2008), Huson et al (2007), Caporaso et al (2010)).

Currently the most popular methods for sequence annotation are based on pair-wise comparison of query sequences with reference sequence databases of model organisms; the most common example being BLAST (Basic Local Alignment Search Tool), (Altschul S.F. (1990)). The goal of a pair-wise comparison is to locate a reference sequence that is similar to the query sequence. The name of the organism and functional annotation of the best match, “hit”, is used to append annotate the query sequence. Although these types of analyses are convenient, there are known issues when dealing with the shorter reads of metagenomic libraries. A 2008 study found that when BLAST annotation was applied to two versions of a data-set, a long read (750 bp) and a short read (100-200 bp), up to 72% of annotations for long reads were not identified in the short reads (Wommack et al (2008)). This limitation is compounded when BLAST is used to annotate metagenomes containing highly divergent organisms with no established model system, as is common with most microbial communities. Insufficient reference information and annotation techniques have driven efforts to find alternative comparison methods specifically focused on metagenomic data-sets.

Hidden Markov Model (HMM) based methods designed for detection of remote homologies using sequence alignment profiles have helped to address the issue of inadequate reference information with respect to annotating metagenomes. HMMER is a software suite designed to evaluate sequence comparisons for the purpose of identifying homology using profile HMMs (Eddy (1998)). HMMs work by calculating discrete probabilities of each nucleotide base or amino acid in a query sequence. Unlike the arbitrary score-based algorithms, such as BLAST, these probabilities have a stronger statistical framework and therefore can implemented in biological statistical models. Alignments of orthologs, homologus genes from multiple organisms sharing a common ancestor and a shared function, can be used to create a profile HMM for that gene. This profile is used by HMMER, to search against a sequence database to identify new potential orthologs from an unknown set of sequences (in this case environmental sequences from a metagenome). HMMER outputs matches between queries and HMMs, as well as the probabilities associated with those matches, and if a sequence match meets the confidence threshold set by the user, then the query sequence is considered an ortholog to the sequences in the profile. Therefore, HMMER is a mathematically robust annotation method for functional assignment of environmental reads. However, these analyses do not give information on the taxonomic identity of the sequence. Coupling HMM searches with phylogenetic placement methods that can identify the taxonomic or phylogenetic affinities of a sequence, further resolving the identity of the environmental reads.

Phylogenetic-based analysis used for taxonomic assignment improves on annotations based on sequence similarity by including assessment of the evolutionary relationships of the sequences. Furthermore sequences with no appropriate reference sequence matches can be placed on internal nodes of phylogenetic trees, giving some insight into what group they might be most closely related. This is currently the best way to deal with the known biases that exist from incomplete reference databases. Unlike pair-wise scoring algorithms, which only suggest if a query is similar to a single reference sequence or group of sequences; phylogenetic placement uses existing reference trees as a map of how multiple sequences from the environment relate to each other and to known references. Examples of analysis pipelines that allow for phylogenetic or diversity analysis of communities include: MOTHUR and MLST (Schloss et al (2009), Jolley et al (2004)).

In this thesis I discuss a metagenome or environmental amplicon sequence analysis pipeline that uses a combination of HMM searches with phylogenetic placement to annotate metagenomes. Although they are a powerful combination, HMMER and phylogenetic analyses require significant computational power and high quality pre-built reference information. Performing large-scale metagenomic surveys using these methods require thousands of genes to be assembled into profile HMMs and a sophisticated analysis pipeline to direct processing of samples. The main topic of my thesis is to outline the analysis pipeline built to process environmental metagenomic samples using the pplacer software suite and a pilot study performed to demonstrate the utility of our pipeline to investigate gene diversity in a marine oxygen minimum zone (OMZ) off the coast of Chile, South America.

CHAPTER 1: Completion of the Phylogenetic Analysis Workflow

INTRODUCTION

A phylogenetic analysis workflow

Our approach to utilizing the power of phylogenetic analysis for metagenomic annotation is to use the well-established program HMMER in combination with the phylogenetic placement software pplacer (Matsen et al (2010)). Our phylogenetic analysis workflow (PAW) is a powerful and robust series of analyses designed for large-scale, high-throughput and comprehensive surveys of these important, yet largely unexplored, microbial communities. The PAW is a previously created semi-automated high-throughput analysis pipeline specifically designed to help investigate uncharacterized, diverse microbial communities (Land et al (2015)). It is designed to search short-read shotgun metagenomes for potential orthologs of a user specified reference gene or group of genes. The PAW has two main components: 1) creating automated workflow for generating *reference packages* and 2) running a large set of reference packages across a metagnome to annotate environmental sequence reads.

Building reference packages

The PAW first creates *reference packages* for each gene of interest from available multiple sequence alignments (MSAs), profile HMMs, and a custom built reference DB containing a tailored collection of sequence information for taxa found in a given MSA. This package contains several important components built from reference sequences for that gene. The components include: a multiple sequence alignment, hidden Markov model (HMM), un-rooted phylogenetic tree, taxonomy list, and controls files. In order to scale this project to include many thousands of genes the production of packages was built into a semi-autonomous pipeline inside the PAW, referred to as the reference package pipeline (RPP).

We have chosen to generate reference packages from a set of known orthologs from the COGs, TIGRfams, and NCBI clusters (Tatusov et al (2012), Haft (2003), Klimke et al (2009)). This reference package pipeline has generated a total of 9,207 reference packages that can be used to annotate metagenomic sequences.

Functional and marker seed data

The initial reference information for each gene, identified as a "seed", must be in the form of a profile HMM. This seed is used as the core molecular and taxonomic representatives of the gene, so seeds must be carefully selected and built. There are several long-term functional gene projects with available HMM seeds via download from FTP sites. The projects selected for this study are: Clusters of Orthologous Genes (COGs), TIGRfams, NCBI Protein Clusters (CHLs, PTZs, MTHs) (Tatusov et al (2012), Haft (2003), Klimke et al (2009)). These genes are well established, with many years of investigative effort contributed to support the sequences they contain. Standard marker genes (SSUs) were also included in this study, requiring their seeds to be custom built before package building. These genes included small sub-unit ribosomal genes for Bacteria, Archaea, and Eukaryotes.

Building SSUs seeds

The non-redundant 99% identity SSU reference DB release 119 was downloaded from the ARB-SILVA web server to be used to create SSU seeds (Quast et al (2013)). The DB was de-duplicated for both identical sequences and taxa to reduce its complexity using the seqmagick utility. PhyloSift v1.0.1, a suite of tools for phylogenetic analysis, was used to recruit sequences from the DB to one of the three seeds based on included SSU markers packages (Darling et al (2014)). The tool was used with default out of the box settings for the version and output sequence alignments for the SSU genes containing reads from the ARB-SILVA reference DB.

Reference DB for RPP

The gene seeds are the sequence core for making packages, but they only contain the most well established sequences for each gene. This can affect their usefulness when investigating a specific environment or community. This is mitigated by incorporating sequences specifically associated with the study setting. A custom reference DB was created for this purpose by combining Archaea, Bacteria, fungi, invertebrates, plants, plasmid, protozoa, and viral data from RefSeq release 66 (Pruitt et al (2007)). Sequences from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP), available in July 2014, were added to this reference DB to increase resolution of Eukaryotic taxa bringing total sequence reads to 35,205,636 (Keeling et al (2014)). RPP requires an NCBI Taxonomy DB to be downloaded and installed locally. It was crucial that any taxonomy identification numbers (tax ids) be synchronized with this DB version, 4.0, as many downstream functions and analyses relied on this assumption. As such, custom file checking scripts were built in the Python programming language to rename, delete, and merge tax ids for all reference information (Sanner (1999)).

Running through RPP

The functional gene seeds, including all sets but the SSUs, were run through the RPP. Briefly, the seeds are compared to the reference DB using HMMER 3.0 and the sequences, using a threshold of similarity, e^{-5}, are recruited (Finn et al (2011)). RPP then proceeds to build all the necessary components of a *reference package* listed in the first paragraph of the "Building Reference Packages" section.

The SSUs, however, needed to be run in a different fashion as they are not translated into protein-space and have a significantly larger data pool from which to draw. Since the seeds were built from custom SSU data, there was no need to recruit from the reference DB using HMMER. Several data preparation steps needed to be modified to handle cDNA instead of peptides. Lastly, during the step where each reference tree is pruned to remove polytomies at the end of branches, a SSU-specific configuration was required to sufficiently trim the trees while preserving their quality.

All packages were reviewed using package_checker.py, a custom quality checking script. The files required for a complete package were counted, if there were missing files the package was deemed incomplete and was not used for further analyses. The removed packages may lack sufficient reference information or have other computational reasons for not completing successfully. A full review of this topic is beyond the topic of this study, but this is an on-going area of investigation.

Once a set of reference packages are established, they are used to annotate environmental reads using HMM searches and phylogenetic placement. At the core of the PAW is HMMER and pplacer, software that employs phylogenetic placement algorithms on short shotgun sequences. pplacer places metagenomic reads on the fixed branches of each reference tree using probability calculations to append a confidence score to each placement (Matsen et al (2010)) (Figure 1).

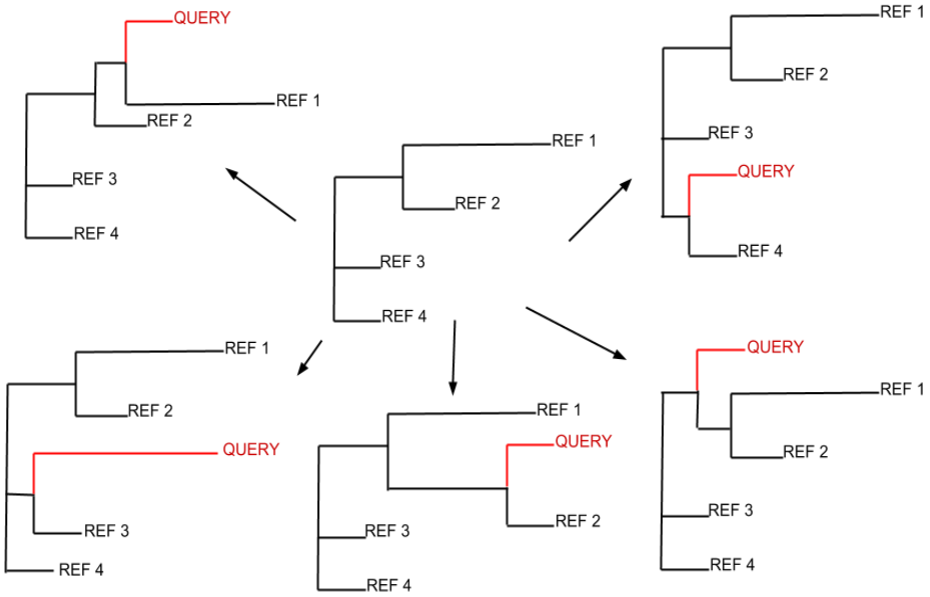

Figure 1: Visual representation of phylogenetic placement. The reads (red) are placed on different branches of the reference tree (center) until the placement of highest probability is determined. (Figure after E. Matsen)

Metagenome annotation using PAW

Using the reference packages created by the RPP, the PAW can then annotate metagenomes extracted from environmental samples. The hmmsearch function from the HMMER suite is used to recruit reads from the metagenomes with a e-value threshold of e^{-5}. A read is recruited to the reference package with the lowest e-value from the hmmsearch comparison. The recruited reads are then aligned to the MSA for that reference package using hmmalign from the HMMER suite and this output is piped into a pplacer analysis. The recruitment process is run in parallel to improve run-time and each pplacer analysis per gene is performed in parallel when multiple gene reference packages are being used.

The resulting output from the PAW is an un-rooted phylogenetic tree for each gene with query reads, likely orthologs, placed on its branches (Matsen et al (2010)). Reads may have several possible placements on the tree, each of which can be assessed by an associated probability score. The file format of a post-PAW tree containing placements is a subtype of the JavaScript Object Notation (JSON) format, referred to as a *jplace* file in this study. The PAW outputs one *jplace* file per gene for each sample. Due to the architecture of the PAW, large-scale gene surveys quickly produce a quantity of *jplace* files, unmanageable by manual manipulation methods. As part of my thesis, I created a downstream analysis pipeline (DAP) for the purpose of managing and analyzing PAW output of large-scale projects (Figure 2).

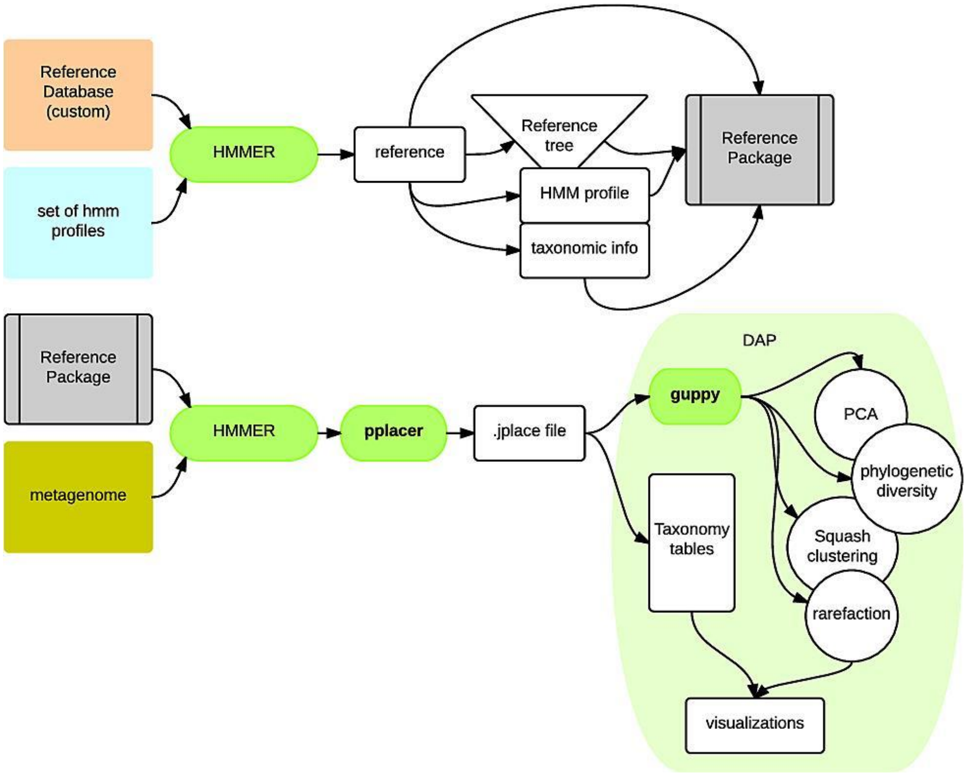

Figure 2: Visual breakdown of the PAW and DAP. The PAW is everything outside of the shaded region labeled as DAP (Figure from R. Kodner).

METHODS

The downstream analysis pipeline

I built a series of scripts into a downstream analysis pipeline (DAP) in the Python computer language to help with the handling of the PAW output, as the *jplace* files are complex and tend to be numerous. For each *jplace* file, the DAP performs: 1) initial quality filtering, 2) parse *jplace* files by sample and taxonomic criteria, 3) run general statistics and calculations, 4) visualize summary data for further investigation (Figure 2). These functions are designed to be run autonomously, to allow for large amounts of data to be processed in a consistent and efficient way.

Pre-filtering, quality control

The first pre-stats script is built to extract only the placements within a specific threshold of confidence based on the maximum-likelihood weight ratio score (MLWR) appended to each score by the PAW. The threshold is defined as: "If the difference between the MLWR of the first and the second placement on a branch of a tree is > 0.05, then the first placement is marked as *confident* and the others are discarded as *junk* or bad placements. If, however, the first and second placements MWLR are within 0.05 of each other, then all placements on that branch are marked as *fuzzy* or uncertain." This function is combined with others in the lineage.py script, described in the next paragraph.

Applying lineage annotations

The National Center for Biotechnology Information (NCBI) has an online resource for taxonomic annotations, including taxonomic and lineage information for all established lineages of described organisms. When the lineage of an organism is established but not officially described at a level in the classical hierarchical taxonomy such as genus or phylum, it is designated as a "no rank" by NCBI Taxonomy (Sayers et al (2011), Benson et al (2015)).

This is common for microeukaryote taxonomic categories that have been more recently established due to molecular phylogenetics but have not yet been officially described in the literature. Due to this naming convention, most of the taxa-based annotations during the PAW analysis are unable to be used by pplacer's built-in classification functionality. To remedy this we built lineage.py, a script that appends the correct annotations to the PAW outputs so that taxonomic information can be used for comparisons. The lineage.py script: accepts the standard output of PAW in *jplace* file format, creates a full lineage of all known taxa from the NCBI Taxonomy database, accesses taxonomy identification codes (taxids) from the *jplace* files, adds specified levels of the lineage, and utilizes the previously stated filtering functions to output to confident, fuzzy or junk files. The taxonomic levels automatically appended are the top three under cellular organism, referred to as Domain, Division, and Clade. The outputs of lineage.py are comma-separated variable files for confident, fuzzy, and junk placements, all with associated taxonomic annotations appended to them.

Mapping Domain and splitting jplace files

It is very useful for a variety of analyses to split jplace files by a taxonomic level or group, such as Domain or Division. We built taxmapper.py, a taxon mapping tool, for the purpose of separating each *jplace* by any specified taxonomic level. In each *jplace* file there are reads that have been placed on the reference tree. Those reads, known as placements, have names that pplacer can uses to run other functions. The pplacer suite includes a program called guppy, which can split *jplace* files by sub-strings in each placements name. The taxmapper.py script utilizes this function by first appending the taxon annotations from lineage.py output to each of their respect placement names in each *jplace* file. After the taxonomic information is added to the name of each placement, guppy is used to split the *jplace* files by Domain name via a wrapper script called guppy_quick_split.py. This then allows for all following calculations to be easily performed separately for each Domain of life.

Basic calculations and stats

A traditional method for initially describing a microbial community structure is to quantify read counts for each organism by gene. This is achieved in the DAP by countbot.py, a simple quantifying script for calculating gene abundances for specified groups of data. For this study each gene was quantified by sample, Domain, Division, Clade, and functional category. The countbot.py script utilizes the standard output of lineage.py, counting the occurrences of each previously mentioned category in the confident output file.

There are many possible statistical measures and calculations that could be useful when investigating microbial communities. Several calculations have been incorporated into the DAP to give a starting point for more in depth analyses. The DAP utilizes pplacer functions like edge-principle components analysis (edge-PCA), quadratic entropy, phylogenetic entropy, faith phylogenetic diversity (PD), abundance-weighted phylogenetic diversity (AWPD), expected distance between placements (EDPL) (Matsen and Evans (2013), McCoy and Matsen (2013), Matsen et al (2010)). Each of the previously mentioned pplacer functions has a wrapper script built around it in order to manage the input and output data.

The DAP can also calculate the Shannon Diversity Index (SDI) (Hamilton (2005)) (1), paired/unpaired student t-tests from the scipy pythonic library, and determine differential abundance between communities using the DESeq2 R package. The SDI calculation is run by a custom script called SDI_calc.py that uses countbot.py standard output. The count data for the lowest possible taxids are used for the SDI calculation and the diversity measures are collected by gene, sample, and Domain.

$$\boldsymbol{SDI} = -\sum_{i=l}^{R} \frac{c}{C} \ln\left(\frac{c}{C}\right) \tag{1}$$

c = count of lowest taxa

C = total count for gene/sample/Domain

Differential abundances are calculated using the DESeq2 R package and custom data prep script called deseq_prep.py, which accept the standard output from countBot.py. DESeq2 was originally built to compare transcriptome data to identify whether differences in expression levels between data sets from different conditions could be explained by simply biological variance (Love et al (2014)). Using these same principles and functions this analysis can be applied to metagenomes, given that there are two testable condition types present in the data (Jonsson et al (2016), Xu et al (2015)). An added advantage to this analysis method is that it does not require sequence libraries to be normalized before-hand, a commonly required pre-analysis step (McMurdie and Holmes (2014)). For this study differential abundance was calculated between oxic and suboxic zones using an R control script, DESeq2_cmds.R. The genes found to be differentially abundance were then visualized to explore the functional diversity of each sample.

Visualizations

After the filtering, collecting, and calculating scripts have been run, the DAP can then output a series visualizations. There are many base functionalities for visualization in the DAP. The functionalities include: scatter plots, bar charts, pie charts, histograms, heatmaps, and phylogenetic trees. Several scripts were built to use these base functions to automatically build report graphs for this project. They scripts include: bar_bell.py, scat_man.py, heating_up.py, histo_listo.py, and guppy (last script from pplacer suite)

All of the scripts were built using the Python programming language in a Linux environment and are built to be run from the command-line.

Computational resources

The PAW and DAP are housed on the computer cluster located in the computer science department at Western Washington University. The cluster has 8 nodes capable of running 24 single thread jobs per node for a total of 192 parallel processes. We also used the Computer Science department data storage facilities for all input and output data for this project.

Code repository

All code associated with the PAW and the DAP are freely available on the Kodner lab repository located on GitHub (https://github.com/McGlock/cluster_pipeline, https://github.com/McGlock/DAP).

RESULTS

The Downstream analysis pipeline

The DAP performs multiple functions required to mass process thousands of *jplace* files for a community analysis. The *jplace* files are collected and read into a single data file allowing for quality filtering, parsing to be performed on the entire data-set. Once the data checks are completed, there are many other post PAW functions to help with further investigation including: sorting by a specified taxonomic levels, basic statistics such as edge

PCA, EDPL, AWPD, and other phylogenetic analyses, and data report visualizations. All of these scripts are freely available at the Kodner lab Github repository, along with documentation for running the PAW and DAP (https://github.com/McGlock/cluster_pipeline, https://github.com/McGlock/DAP).

Reference package production

A total of 9,207 genes were successfully run through the RPP to produce reference packages for use in the PAW placement analysis. This included 9,204 functional from COGs, TIGRfams, and NCBI clusters and 3 custom built SSU genes. There were 122 functional genes that did not pass the inspection stage of the RPP due to lack of reference sequences or insufficient quality.

On average, bacteria comprise over 50% of the taxa recruited for each gene, with the exception being the MTHs (Figure 3). It is not surprising that the MTHs have less than 50% contribution from bacteria because these are mitochondrial gene packages. However, it is also puzzling that the CHL (chloroplast) genes do not show the same trend. Evolutionary studies for mitochondrial and chloroplast origins have suggested that the endosymbiosis of the former was much earlier and that the latter is a more modern addition. Over time more gene transfer and hybridization may have occurred in the mitochondrial genome, effectively masking its bacterial signal. This effect would be weaker for the younger relationship of chloroplasts, preserving the bacterial signal in the gene packages. It must also be noted that the MTHs had the smallest number of genes overall, a possible source of bias for the taxonomic representation in the packages.

Eukaryotes had a range between 14%-50% of taxa and Archaea made up less than 7% of taxa for all projects. The project with the highest average taxa per gene was the SSUs, with a total of 5118 taxa. Then CHLs, PTZs, COGs, TIGRs, and MTH in descending order (Figure 3). The average length for reference sequences in gene trees was highest for the SSUs genes at 1941 base pairs (bp) (Appendix: Table 1). The other averages in descending order were: MTHs, TIGRs, PTZs, COGs, CHLs at 216, 212, 197, 190, 165 bp.

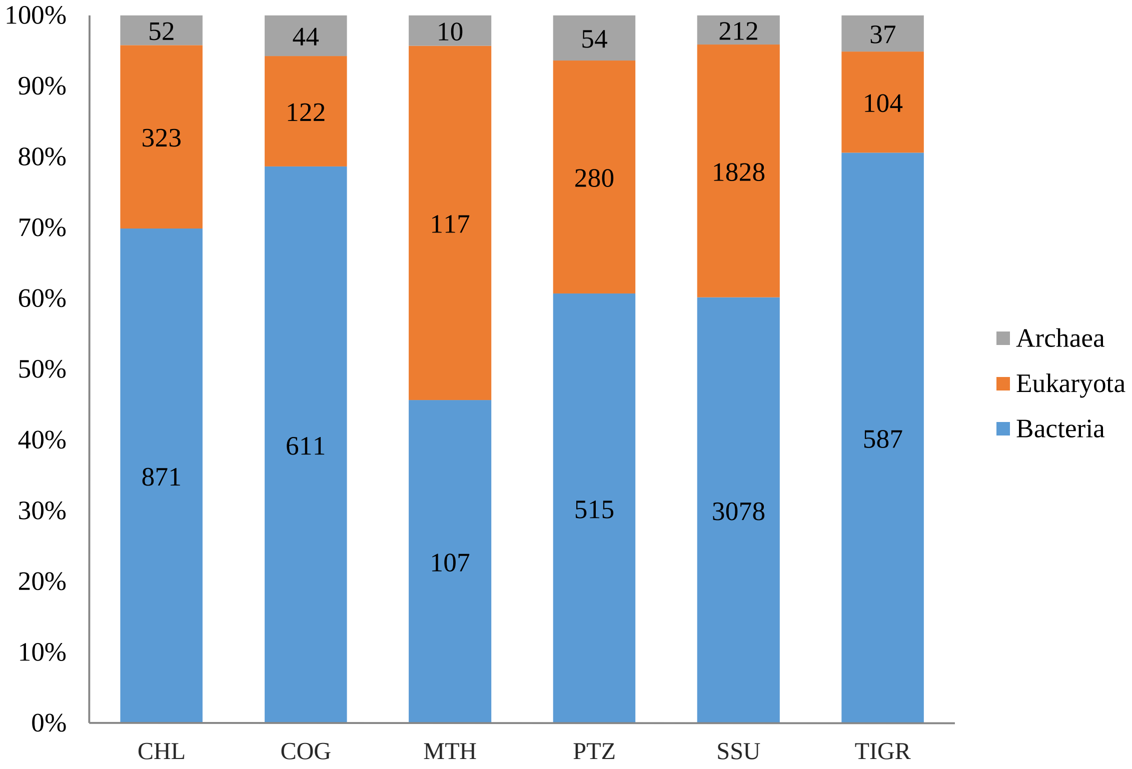

Figure 3: Average number of taxa by Domain contributing to each gene project. Archaea (gray), Eukaryotes (orange), and Bacteria (blue).

Originally, the standard RPP was to be used to create the SSU packages using a custom built reference sequence library including SSU sequences Bacteria, Eukaryotes and Archaea. However, software and hardware limitations did not allow RPP to complete successfully. It was discovered that the cluster computer did not have a sufficient amount of RAM to complete the more intensive steps of the package creation, namely multiple sequence alignment (MSA) with MUSCLE (Edgar (2004)). In order to remedy this issue the PhyloSift step was incorporated in the SSU package production, and this seemed to allow for the creation of the reference packages. In future studies, if higher resolution is needed for SSUs, packages for specific groups should be created, allowing for the inclusion of more SSU information for that group. Improving on the limitations of current MSA software is not a simple undertaking, so for biologists refining the reference sequence selection process through the use of software like PhyloSift is a very important pre-analysis step.

Efficient computer usage is currently one of the biggest issues in bioinformatics. The majority of analysis software is built to handle small numbers of files at a time, i.e., one profile-hmm or one MSA. In the building of the DAP and the running of the PAW these programs needed to be executed many thousands of times in order to complete the processing of the entire data-set. This requires many wrapper scripts to be built and a protocol for the format and content of input data to be created. While the scripts built in this project perform their function properly, due to limits of time and software development resources, optimization would be a necessary next step. There are many processes during the PAW and DAP that could benefit from a more mathematics-based or parallel-computing-based approach. The majority of wrapper scripts are built in Python, but many functions could be migrated to a lower-level language to improve efficiency and therefore overall run-time.

Currently, the DAP have a package checking function to quickly identify packages that have not be correctly created. A further investigation of the genes that did not pass the quality checking should be performed. It is unclear as to why these packages are not successfully created, although a cursory check showed that many of them had a limited number of reference sequences, which could have effects on the quality of the package. In the future, it would be helpful for the DAP user to be able to read out a report on each package, providing statistics for the quality of the build. This would require a significant effort to review the building process and possible weak points in the production of packages.

CONCLUSIONS

Reference packages are a valuable resource for studying metagenomes, but require computational infrastructure and specialized software to create on a large-scale. This is not ideal for all research projects; however availability of pre-built packages from this project can provide a solution for researchers who lack the expertise or budget to create their own. Taxonomic, functional, and phylogenetic information is contained in these packages is a more accessible format and in combination with pplacer, can provide high quality sequence annotations for any study with a metagenomic component.

The DAP successfully completed the PAW, making it more user friendly for biologists in future sequence-based analyses. The semi-autonomous workflow of the PAW/DAP allow for large-scale high-throughput surveys of metagenomic libraries against thousands of genes. The DAP collects large output volumes and presents the user with manageable analysis files, more easily accessible for further manual investigation into possible biological signals. The combination of methods in this pipeline allow for a query sequence to be annotated with both taxonomic and functional information, further improving on current annotation standards. Direct connection of organisms to ecosystem functions will lead to better understanding of the structure and interactions of microbial communities as a whole.

CHAPTER 2: PAW/DAP Capabilities & A New Study on Diversity

INTRODUCTION

Studying biodiversity

Biodiversity has been shown to influence an ecosystems ability to resist and recover from environmental variation (Norberg (2013), Hillebrand et al (2007), Loreau et al (2001)). However, a consensus of the most suitable methods for measuring diversity in microbial systems has not yet been reached (Caron et al (2009), Rosselló-Mora and Amann (2001)). Traditional diversity components of a microbial study include a gene survey using the small sub-unit ribosomal RNA genes (SSUs) and a functional richness (FR) measure, commonly identification and quantification of unique functional genes. Although, these methods can give insight into both evolutionary relationships of organisms and the total functional capabilities of a community, there are inherent problems with both when investigating microbial groups.

SSU surveys have been used extensively to investigate the evolutionary relationships between many groups including Bacteria and macro-Eukaryotes. These highly conserved genes can be helpful when looking at ancient lineages and distantly related organisms, but definitions of evolution are largely based on macro-Eukaryotic biology, much of which cannot be directly applied to microbes. Genetic recombination from lateral gene transfer is suggested to be a major influence of the genetic diversity in bacterial groups (Ochman et al (2000)). Genomic plasticity can lead to organisms with identical or similar SSU sequence identity having significantly different genomic content and distinct ecological influences (Thompson et al (2005)). The propensity of some groups to have more than a single copy of the SSU gene can also lead to artifacts in diversity measurements (Acinas et al (2004)). The implications of these findings are that phylogenetic diversity (PD) analyses based on SSUs do not directly represent the functional diversity (FD) of the community, and in some cases could drastically underestimate the overall evolutionary diversity.

A review study containing data from 29 grassland plant experiments found that PD and FD were both predictors of the influence of biodiversity on ecosystem function (Flynn et al (2011)). FR had the lowest predictive power of all measures, indicating that it shows less utility in understanding the relationship between biodiversity and ecosystem function. Similar studies support these findings and also suggest that both FR and species richness (SR) are the least informative predictors (Cadotte et al (2009), Petchey et al (2004)). Utilizing the PAW/DAP effectively combines PD and FD into one analysis allowing for both taxonomic and functional traits to be examined and directly linked with each other. Functional phylogenetic diversity (FPD) incorporates sequence similarity information and functional annotations to get a high resolution of a community's functional stability and architecture.

There are large repositories of functional housekeeping genes currently available from online resources. Along with their high conservation among divergent lineages, the functions of these genes have been studied and are curated. This makes them a valuable annotation resource for a phylogenetic study of an under-characterized community. The Clusters of Orthologous Genes (COGs) represent a well-studied set of conserved functional genes. These genes can give insight into the present community's functional capabilities as well as the evolutionary relationships for the organisms contributing to these functions.

Understanding the relationship between microbial biodiversity and ecosystem function is a critical component when attempting to characterize a community. Understanding the evolutionary history of organisms and the functions they perform can give insight into current global distributions and how that might change in the coming years. Diversity can also be used as a metric to find members or functions, which may be under selective pressure in an ecosystem.

Applying the analysis to an oxygen minimum zone data-set

Oxygen minimum zones (OMZs) influence global biogeochemical processes and have a significant influence on community structure in the oceans. Naturally occurring OMZs are found in areas of nutrient upwelling allowing for high levels of photosynthetic primary production. The resulting biomass is decomposed by microbial heterotrophs via aerobic respiration. This, in conjunction with insufficient ventilation and low circulation, can lead to large areas of the mesopelagic having reduced levels of dissolved oxygen (Ulloa et al (2012), Stewart (2011), Stramma et al (2008), Diaz and Rosenberg (2008), Wyrtki (1962)). OMZs are defined as having dissolved oxygen concentrations of <20uM, necessitating the use of alternative terminal electron acceptors during cellular respiration, such as nitrate, nitrite, manganese, iron, sulfate, and carbon dioxide. Current research estimates that OMZs make up approximately 7% of the total volume of the oceans and contribute to over 33% of fixed nitrogen loss in this global ecosystem (Hawley et al (2014), Wright et al (2012), Paulmier and Ruiz-Pino (2009), Galloway et al (2004), Codispoti et al (2001)).

Recently, studies have concluded that agricultural nutrient runoff and climate change are contributing to the expansion of OMZs on a global scale (Stewart (2011), Stramma et al (2008), Diaz and Rosenberg (2008)). OMZ expansions driven by anthropogenic sources can potentially have large ecologic and economic implications as they have distinct biochemical properties, distinct from oxygen-rich zones. Correctly identifying natural variation in an OMZ community will allow for future studies to investigate and understand the consequences of human input into these systems.

Eastern tropical south Pacific oxygen minimum zone (ETSP OMZ)

The ETSP OMZ is a permanent low oxygen zone located off the western coast of Chile. The OMZ is located at 100-500m, with seasonal variation of the boundaries. The data-set was collected from the high dissolved oxygen (>200µmol/L) surface through the low dissolved oxygen (<5µmol/L) core (Bryant et al (2012)).

The ETSP OMZ dataset has shown that redox pathways in sulfur-cycling bacteria may contribute to up to 30% of the organic carbon mineralization (Canfield et al (2010)). High abundance of crenarchaeal-like Archaea were identified in the nitrification transitional zone between oxic and suboxic regions of the water column (Stewart et al (2012)). Finally, a 2012 study found that taxonomic richness, faith phylogenetic diversity, and functional richness all decreased as oceanic depth increased (Bryant et al (2012)).

This bacteria-centric data-set is interesting because it was collected across the oxygen gradient in the OMZ over a period of three years with increasing sequencing effort each year. This allows for an investigation of a highly dynamic physiochemical environment with a diverse uncharacterized community, but also an investigation of the influence of sequencing effort on diversity measurements. The goal of this work is to investigate the utility of functional genes for exploring community function and diversity as well as the influence of sequencing effort on patterns of diversity.

This study will use the previously reviewed bioinformatics pipeline (chapter 1) to investigate the utility of functional gene for calculating diversity in comparison to the current standard, which utilizes SSU marker genes. We calculate phylogenetic diversity (PD) using the PAW/DAP pipline, and functional gene PD measurements will then be compared to SSUs and information about sequencing effort. The sub-set of genes found through DA analysis will be compared to the patterns for the full set of genes. These analyses and comparisons will help to test the pipeline and functional genes utility in community-level functional and diversity studies using metagenomes.

METHODS

OMZ metagenome preparation

The raw data was collected and processed by the Microbial Oceanography of Oxygen Minimum Zones (MOOMZ) project and stored in the NCBI Sequence Read Archive (SRA) (Leinonen et al (2011)). This study included 17 previously published metagenomic samples collected from Station #3 (20°07'S, 70°23'W) off the coast of Iquique (Appendix: Table 2), Chile during the austral fall (June 2008), winter (August 2009), and summer (January 2010) as part of the Microbial Oceanography of Oxygen Minimum Zone (MOOMZ) cruises aboard the R/V *Vidal Gormaz* (Bryant et al (2012), Stewart (2011), Canfield et al (2010)). Specific collection methods can be found in previous publications on the data-set (Stewart (2011), Canfield et al (2010)). The samples were pre-filtered through 1.6µm filters and collected on 0.22µm filters, making the size fraction 0.22-1.6µm. Genomic DNA extraction and sequencing methods can be found in Stewart (2011) and Canfield et al (2010). The HTS technology used for the pyrosequencing was a Roche Genome Sequencer FLX instrument using either FLX or Titanium series reagents, see previous methods for specifics (Appendix: Table 2).

The raw nucleotide sequence reads for each OMZ sample were downloaded from the NCBI SRA database. The data was de-duplicated by sequence and by read name using the seqmagick command line utility available via GitHub (https://github.com/fhcrc/seqmagick). The European Molecular Biology Open Software Suite (EMBOSS) program getorf was used to translate the metagenomes into protein-space (Rice et al (2000)). After deduplication the raw nucleotide dataset equated to 15,832,111 reads and after translation 438,239,102 open reading frames (ORFs). The SRA identification codes for each sample library were added to their respective reads for later use in the DAP (Appendix: Table 2).

Running PAW/DAP on OMZ

The 3 SSU and 4,425 COG reference packages were used for this study, as they represent well studied groups for both marker and functional genes. The PAW was used with the OMZ data-set and reference packages as input, producing 4,428 *jplace* files. The output *jplace* files were then run through the DAP, using scripts outlined in chapter 2. Briefly, taxonomic annotations were mapped to each read, allowing for abundance and diversity measures to be calculated for functional, taxonomic, spatial, and temporal groups. The diversity measure used for this study was AWPD defined and employed by the pplacer suite (2,3,4) (McCoy and Matsen (2013)).

(2) $$PD_u(s) = \sum_i l_i g\big(D_s(i)\big)$$

(3) $$g_\theta(x) = min\big(x^\theta, (1-x)^\theta\big)$$

(4) $$AWPD_\theta(s) = \sum_i l_i g_\theta\big(D_s(i)\big)$$

$where\ \theta = 1$

Abundance Statistics

Differential count analysis is a commonly used method in transcript-level investigations to find genes that have statistically significant differential expression in samples or treatments. However, it may also be useful in metagenomic analyses in the form of differential abundance (DA). The output from our analysis of the OMZ metagenomes presented thousands of genes for further comparisons. Because of the size and complexity of this data, a sub-set of candidate genes showing different patterns of abundance between oxic and suboxic zones were identified using DA analysis. This sub-set of functional genes for oxic and suboxic zones are supported by statistical measures of the DA analysis and can be directly linked to ecologically important functions for their respective zones.

Differential abundance (DA) between oxic and suboxic zones was determined using DESeq2 (described in chapter 1). Metagenomes were grouped by the zone, oxic (>5ug/L) and suboxic (<5ug/L) and by year. Gene abundances were compared between the zones for each year and stats collected on those comparisons. If a gene showed higher abundance in one zone it was passed on to undergo quality filtering. DESeq2 also gave a magnitude of the difference in abundance and two probability scores, a standard p-value and an adjusted p-value (padj). The padj is a p-value adjusted using the Benjamini-Hochberg procedure to control for false discovery rates, R function p.adjust. A threshold of the significance of DA genes was set at less than or equal to 0.05 padj. The functional and taxonomic annotations for each of these genes were investigated to identify important community features for each zone.

A pythonic implementation of the students t-test was used to identify diversity differences between the all COGs and DA COGs (Oliphant (2007)).

Visualizations were created with a combination of DAP functions and standard graphing tools, i.e., R-stat and Microsoft Excel 2013 (Hunter (2007)). All scripts used are available via the GitHub open repository, along with a workflow document (https://github.com/McGlock/cluster_pipeline, https://github.com/McGlock/DAP).

RESULTS & DISSCUSSION

To show the capabilities of the PAW/DAP, an overall observations section is included below. These results outline the taxonomic and functional information extracted from the raw OMZ metagenomes using the PAW/DAP scripts and features. While none of this section's results are new or novel, they show the successful testing and provided output that is made available through the use of the semi-autonomous execution of the PAW/DAP on raw metagenomic data.

Package placement distributions

The majority of reads from the OMZ data-set were confidently recruited and placed in the COGs. A combined total of 5,505,404 reads for SSUs and COGs met the "confident" quality threshold, constituting 34.77% of 15,832,111 open-reading frames (ORFs) from the data-set. There were 5,319 reads placed in SSUs and 5,500,085 placed in COGs, which are

0.1% and 99.9% of total confident placements for SSUs and COGs respectively (see chapter 1: *pre-filtering and quality control* for confidence threshold).

Taxonomic packages comparison

The confident placement distribution by Domain for all genes in total is seen in Figure 4 . Bacteria made up 87.95% of the placements for SSUs and COGs combined, with Eukaryotes and Archaea making up 6.03% and 5.29% respectively (Figure 4, Bar 1). Bacteria were most abundant for the SSU total confident placements at 84.68%, with Archaea at 8.14% and Eukaryotes at 6.52% (Figure 4, Bar 3). A previous study on this data reported an average of 3.8% Eukaryotes for SSUs, suggesting our methods have an increased sensitivity for that Domain (Bryant, 2012). This increased coverage may be influenced in-part by the inclusion of the MMESP transcriptomes as reference information. The COGs had the same distribution as the combined genes for Bacteria, Eukaryotes, and Archaea at 87.95%, 6.03%, and 5.29% (Figure 4, Bar 2).

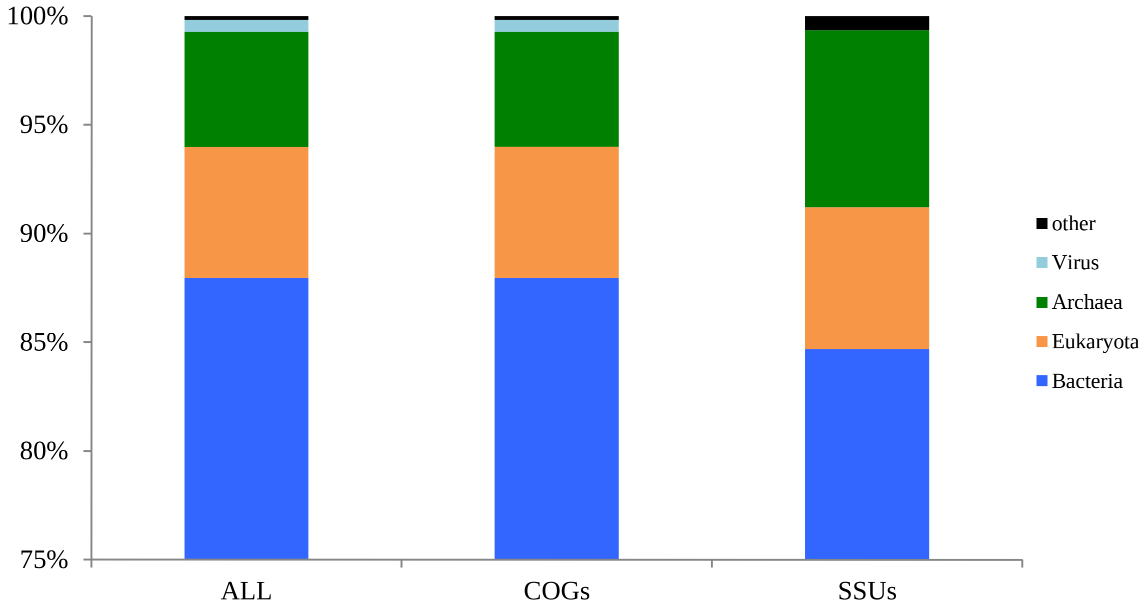

Figure 4: Distribution of confident placements from ETSP OMZ data-set across biological Domain and virus.

The observed distribution of the confident placements is not surprising because bacteria: 1) have higher abundance than both Archaea and Eukaryotes in marine systems, 2) have more reference information and sequenced genomes, 3) were the focus for the original OMZ project and therefore dictated the sampling methods (Heike, 2008). There is also the possibility of bias due to the reference sequences used in the creation of the reference packages. The COG reference packages contain an average of >70% bacterial sequences per gene. However, it is currently unknown to what extent the results are influenced by the taxonomic distribution of the reference packages.

The taxonomic annotations for the COG placements showed a similar distribution to SSUs for biological Domain. This is evidence that using functional housekeeping genes when taken together gives similar taxonomic information as traditional marker genes. However, functional genes are rarely used for diversity based studies. The following analyses investigate the application of functional housekeeping genes for a study in community diversity, in comparison to diversity of traditional marker genes.

Bacteria SSU

There were 16 Division level groups contributing to the observed trends in diversity, 3 of which contributed to 79% of the placements, in descending order: Proteobacteria (48%), environmental samples (21%), and Fibrobacteres/Acidobacteria group (10%). Other groups contributing less than 10% but more than 1% were: Bacteroidetes/Chlorobi group, unclassified bacteria, Actinobacteria, and Planctomycetes. Groups with 1% or less of total placements were: Cyanobacteria, Chlamydiae/Verrucomicrobia group, Spirochaetes, NO MATCH group, Chloroflexi, Firmicutes, Gemmatimonadetes, Tenericutes, and Deferribacteres (Figure 5, Bar 1).

B*a*cteria COG

There were 26 Division level groups annotated as COGs, but only Proteobacteria, at 67%, contributed more than 10% on its own (Figure 5, Bar 2). Groups which contributed less than 10% but more than 1% were: Bacteroidetes/Chlorobi group, Firmicutes, Actinobacteria, Chlamydiae/Verrucomicrobia group, and Cyanobacteria (Figure 5, Bar 2). The remaining 20 groups contributed 1% or less and included: Spirochaetes, Planctomycetes, NO MATCH group, unclassified bacteria, Chloroflexi, Fibrobacteres/Acidobacteria group, Deinococcus-Thermus, Nitrospirae, Aquificae, Tenericutes, Thermotogae, Deferribacteres, Fusobacteria, Synergistes, Thermodesulfobacteria, Elusimicrobia, Dictyoglomi, Armatimonadetes, Chrysiogenetes, Caldiserica (Figure 5, Bar 2).

Bacteria DA COGs

A total of 24 Division level groups contributed to the AWPD for the DA COGs (Figure 5, Bar 3). The Proteobacteria made up 67% of the placements for DA COGs, also the only group contributing over 10% (Figure 5, Bar 3). Groups with more than 1% but less than 10% were: Bacteroidetes/Chlorobi group, Firmicutes, Actinobacteria, and Cyanobacteria (Figure 5, Bar 3). Groups with 1% or less of total placements for DA were: Planctomycetes, Chlamydiae/Verrucomicrobia group, Spirochaetes, NO MATCH group, unclassified bacteria, Chloroflexi, Fibrobacteres/Acidobacteria group, Deinococcus-Thermus, Aquificae, Nitrospirae, Synergistes, Tenericutes, Thermotogae, Deferribacteres, Fusobacteria, Thermodesulfobacteria, Elusimicrobia, Dictyoglomi, and Chrysiogenetes (Figure 5, Bar 3).

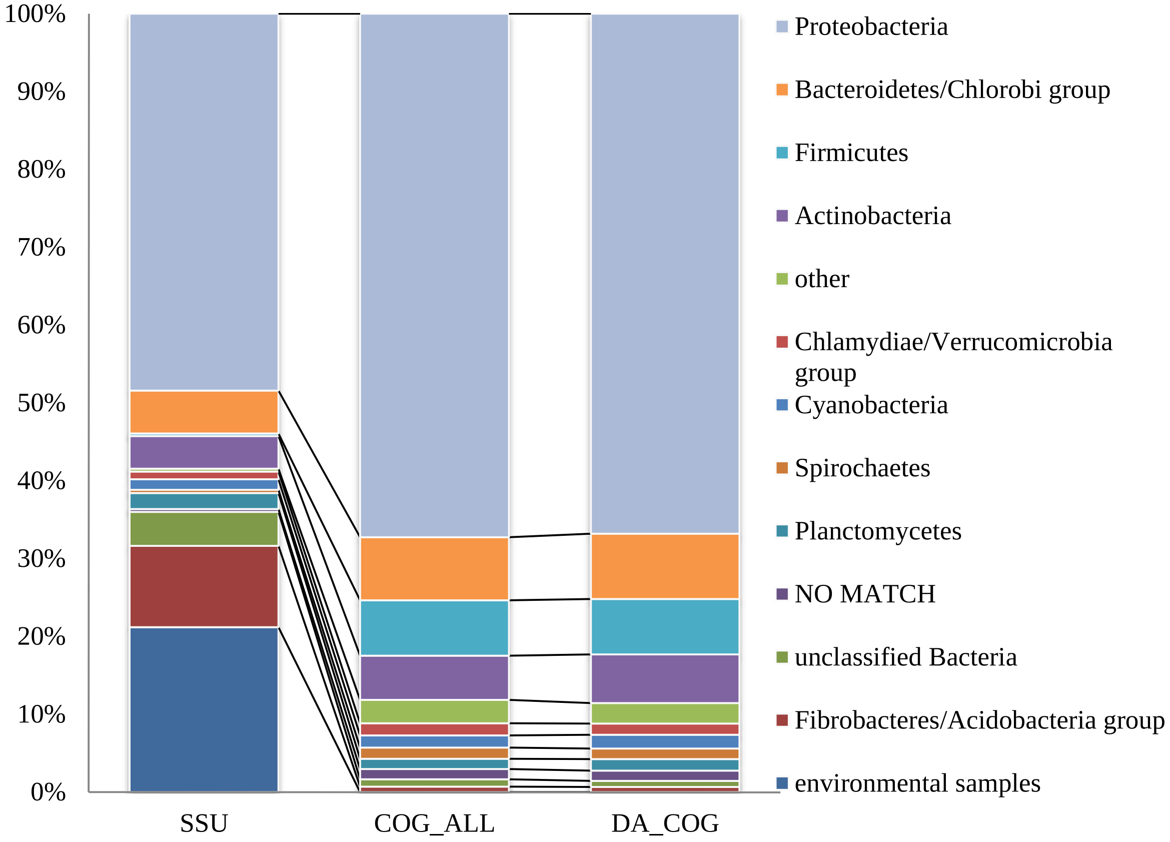

Figure 5: Taxonomic breakdown of confidently placed reads at Division level for bacterial SSU. Taxa contributing <1% in all 3 columns were grouped into the “other” category.

Conclusion for analysis of taxonomic data

The results for SSUs, COGs, and DA COGs all show the Proteobacteria as the dominant Division in the overall data-set. This supports previous work in this region (Stevens and Ulloa (2008)) and importance of this group in dynamic and disturbed systems (Yeo et al (2013)). The presence in the DA COGs also illuminates the metabolic breadth and importance of this group. There were three Divisions shared between the three gene packages that made up more than 1% of the placements: Proteobacteria, Bacteroidetes/Acidobacteria group, and Actinobacteria. The top 4 groups for COGs and DA COGs were shared and similarly ranked. This included the previously stated 3, along with the Firmicutes. The less understood environmental sample group found to contribute a large percentage (21%) to SSUs suggests that there are still many unknown groups present in the bacterial community, but also the lack of resolution when using only a small portion of the sequence reads. The relative proportions of taxa for DA COGs are very similar to the total COGs, suggesting that the DA COGs provide similar taxonomic representation of the community. The COGs and DA COGs did not have the environmental sample or the unknown bacteria as major contributions to confident placements. Also, COGs and DA COGs included all Division level taxonomic annotations found for SSUs, as well as several additional groups. This outcome suggests that COGs may provide more taxonomic information when SSUs give little insight into the source of the more abundant reads.

New analysis of OMZ data: An exploration of diversity measures

It is currently unknown how overall phylogenetic functional diversity compares to measures of diversity for SSU marker genes in metagenomes. The OMZ data was explored using the AWPD metric to compare SSU diversity to that of COGs. This was preformed to investigate the utility of functional genes for studies in microbial diversity.

SSU PD

The traditional diversity measure used in current studies is the Faith PD applied to SSU OTUs (Faith (1992)). As a reference, PD was calculated for the overall SSUs by year and depth (Figure 6). The average PD for SSUs was highest, 9.5, in 2010 at 150m and lowest, 4.25, in 2008 at 200m, both of which are located in the suboxic zone. In 2008, average PD decreased from surface samples through the oxic-suboxic transition, although an increase at 500m was observed. Conversely, 2009 and 2010 showed apparent increases in diversity from oxic to suboxic, with the highest being in suboxic (Figure 6).

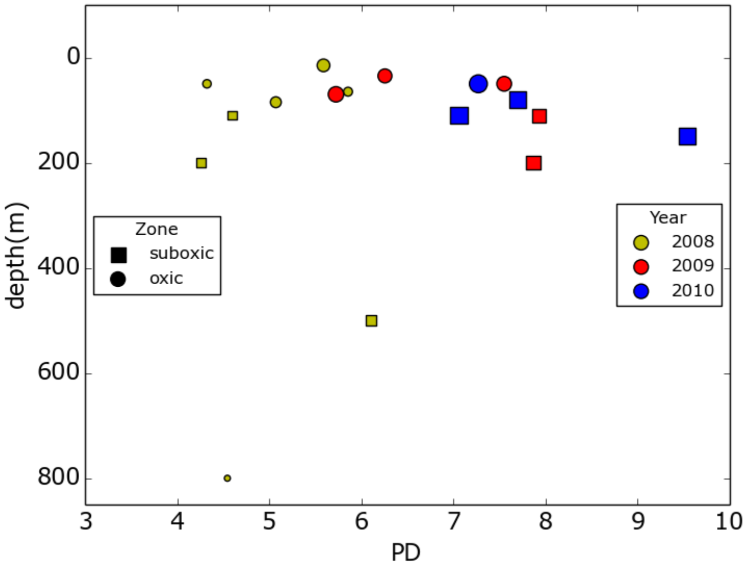

Figure 6: Average PD for 3 SSU genes by depth; color = year, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

Traditional PD does not normalize for abundance in its calculation and therefore does not give an accurate representation of the diversity of a community. This is particularly important when characterizing highly dynamic microbial systems, which tend to be dominated by a small subset of taxa the majority of time, have episodic blooms, and a diverse rare biosphere contributing to overall community processes (Sogin et al (2006)). Community unevenness must be incorporated in diversity measures via abundance information if a true understanding of these biomes is to be achieved. Abundance-weighted phylogenetic diversity (AWPD, *see methods equation 4*) incorporates abundance information into traditional Faith phylogenetic diversity (PD) calculations to account for shifts in community evenness.

SSU & COG AWPD

The overall AWPD by depth, as well as by zone, for SSUs and COGs did not share similar trends (Figure 7, 8). Average AWPD for SSUs was highest and lowest in 2008, 1.0 at 65m (oxic) and 0.61 at 200m (suboxic), respectively. Both 2008 and 2009 have an increase in diversity at the transition between oxic and suboxic, where 2010 have no increase present. The 2009 and 2010 highest average AWPD were in suboxic samples, 110m and 150m respectively. The same increase in diversity observed in PD for the 2008 500m sample was also present in AWPD (Figure 7). COGs average AWPD did not share overall trends with SSUs (Figure 8). The maximum and minimum AWPD was observed in 2009 at 50m, 1.84, and 2008 at 800m, 1.56. AWPD decreased steadily with depth, with the decrease being more rapid through the transition from oxic to suboxic. An outlier in 2008 at 500m showed an increase in AWPD from the 200m sample in that same year. Finally, all samples had higher average AWPD for COGs than SSUs, in some cases over 2x the AWPD for COGs (Figure 7, 8).

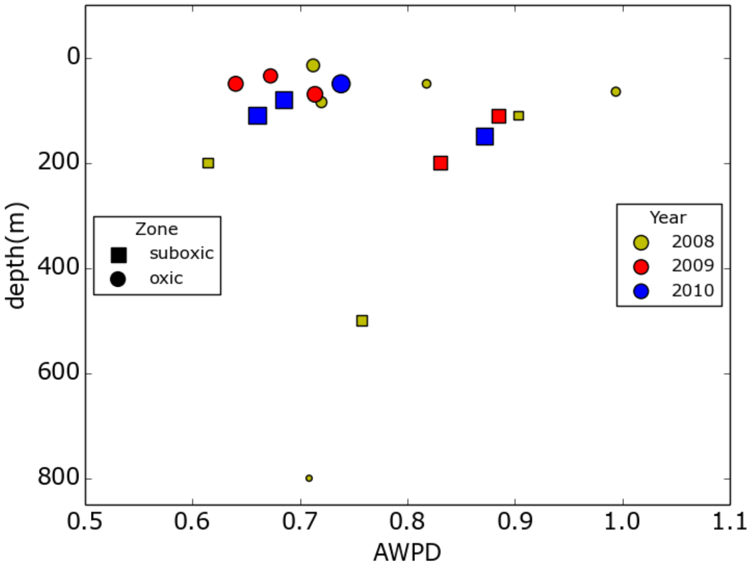

Figure 7: Average AWPD for 3 SSU genes by depth; color = year, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

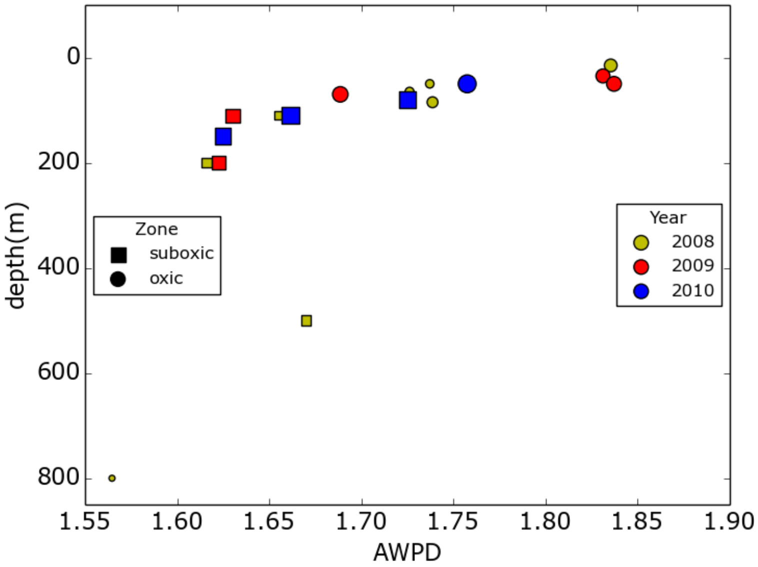

Figure 8: Average AWPD for 4,425 COG genes by depth; color = year, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

The AWPD for COGs shows an increase in AWPD for all samples when compared to SSUs and does not show similar trends with respect to depth. The clear trend in diversity with respect to depth is observed for COGs agrees with previously published trends in diversity for this data-set (Bryant et al (2012)). However, the AWPD for each SSU package should not be averaged to get an overall AWPD for all Domains due to the properties of AWPD itself. When averaged, the diversity scores were unevenly weighted towards the less abundant Archaea and Eukaryotes. SSU reference packages are Domain specific, not allowing for a direct comparison to the combined Domain SSU diversities of Bryant et al (2012). In fact, the direct comparison of SSU results to COGs was not possible either, as the COGs were not built to be Domain specific. To compare diversity measures across Domain for a specific sample, Domains were separated in jplace files using DAP functions to allow for Domain specific calculations of community diversity.

Bacteria: SSU & COG AWPD

The average AWPD for bacterial SSU showed a range from 0.81 to 1.07 which was observed in the 15m to 150m samples. A spike was seen in 2008 and 2009 from 110m to 200m and then decreases again for 2008 at 500m. The deep oxic AWPD is higher than all 2008 suboxic samples (Figure 9).

Average AWPD for COGs was higher than SSUs for all samples, with the lowest for COGs (800m) being higher than the highest (15m) SSU. The maximum and minimum AWPD were 1.83 and 1.54, for the 2009 35m and 2008 800m samples. COG AWPD decreased from surface to 200m samples, with an increase at 500m and then dropping back down at 800m (Figure 10).

The SSU and COG AWPD for Bacteria (Figure 9, 10) show a similar trend of decreasing diversity from surface to the transition between oxic and suboxic. The main driver of this trend is the Division Proteobacteria, making up the majority of placements for both SSUs and COGs. A notable difference between SSU and COG is that COGs have higher diversity for all libraries, in some cases over 2x the average AWPD score. So, while the spatial and physiochemical trends in diversity are similar, the higher average AWPD for COGs indicates a higher overall genetic diversity found in this set of functional genes.

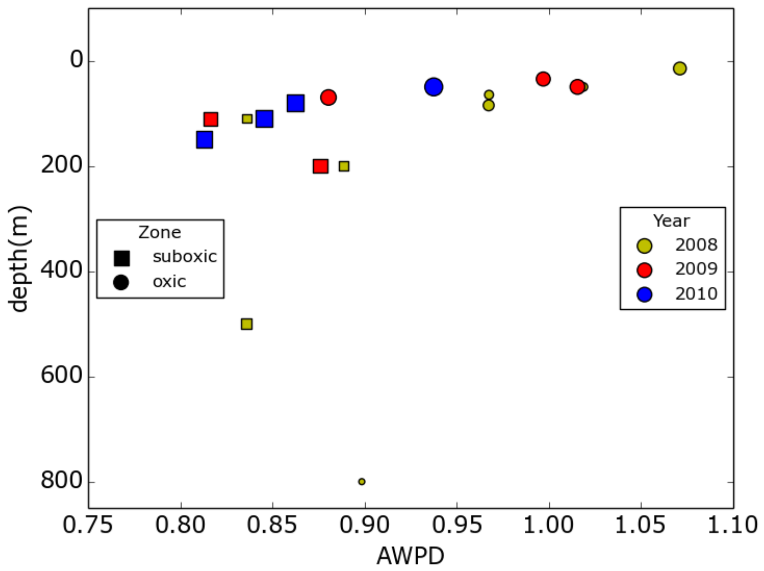

Figure 9: Bacterial average AWPD for SSU genes for the 2008-2010 data by depth; color = year, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

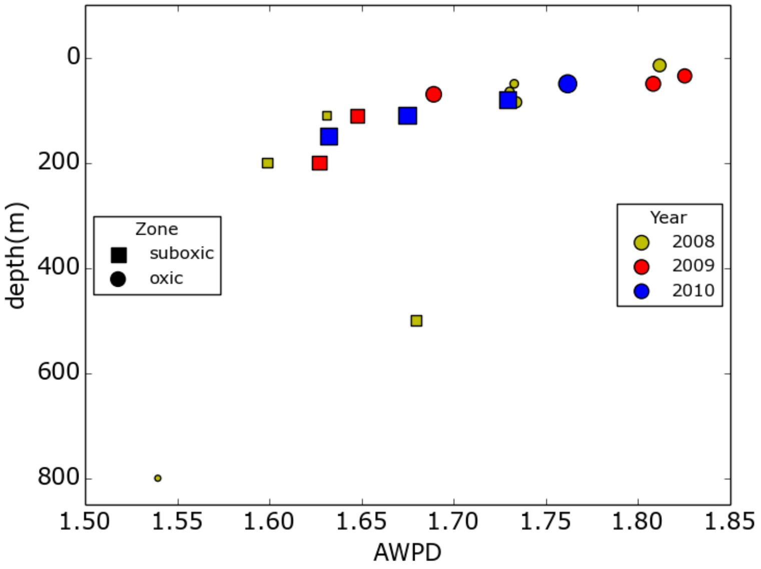

Figure 10: Bacterial average AWPD for COG genes for the 2008-2010 data by depth; color = year, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

DA COGs

Although the previous reviewed results give insight into the advantages of functional genes in diversity studies, another goal of this study was to test a method for the identification of important functions for specific environments. The DESeq2 analysis was employed in order to identify genes that possibly play an important role in oxic or suboxic processes. The differentially abundant (DA) analysis with DESeq2 returned a subset of COGs for each year that showed differential abundance, defined as having an adjusted p-value of less than 0.05, between oxic and suboxic zones. In the 2008 samples 60 DA genes were identified, 31 in oxic and 29 in suboxic (Appendix: Table 3, Figure 11). For 2009, 174 genes were found to be DA, 126 in oxic and 48 suboxic (Appendix: Table 4, Figure 12). The 2010 samples had the most DA genes at 214, 64 oxic and 150 suboxic (Appendix: Table 5, Figure 13). For the scope of this project a specific DA COG for Bacteria was compared to functional analyses from previous work on the ETSP OMZ, followed by a diversity analysis for the complete set of DA genes..

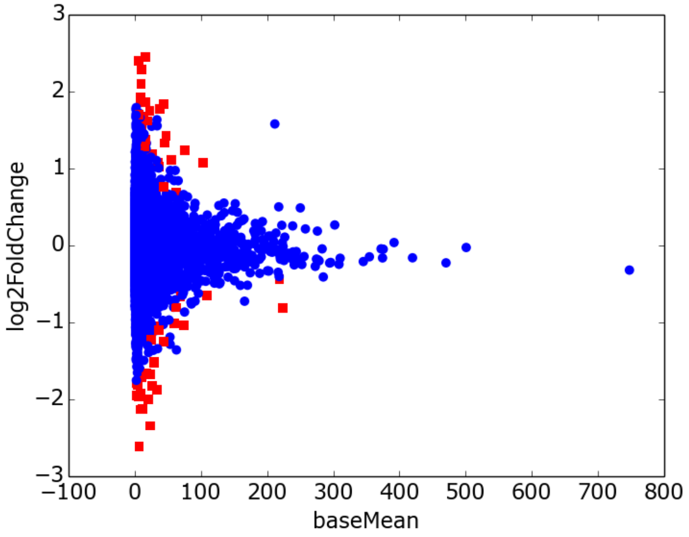

Figure 11: DESeq2 analysis for 2008 bacterial data. x-axis is geometric mean of abundance for genes across libraries. y-axis is the log base 2 of the fold change between oxic and suboxic zones. Greater than 0 on y-axis indicates higher expression in suboxic zones, less than 0 indicates higher abundance in oxic zones. Each point represents a COG or SSU gene; blue circles = padj > 0.05 (not significant), red squares padj < 0.05 (significant).

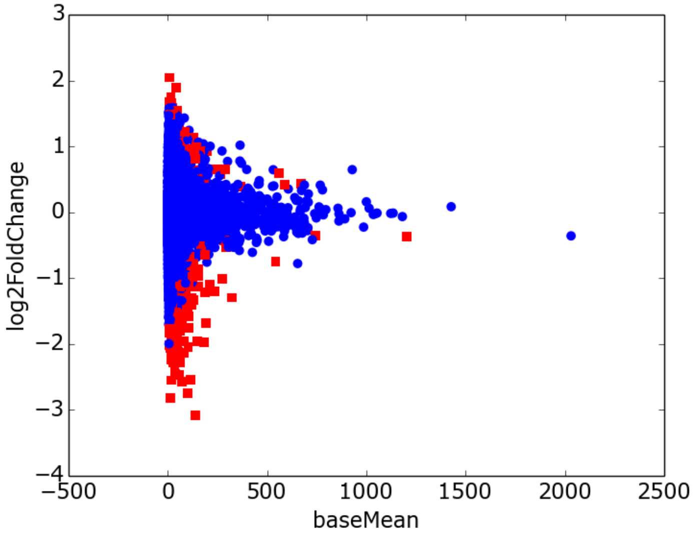

Figure 12: DESeq2 analysis for 2009 bacterial data. x-axis is geometric mean of abundance for genes across libraries. y-axis is the log base 2 of the fold change between oxic and suboxic zones. Greater than 0 on y-axis indicates higher expression in suboxic zones, less than 0 indicates higher expression in oxic zones. Each point represents a COG or SSU gene; blue circles = padj > 0.05 (not significant), red squares padj < 0.05 (significant)

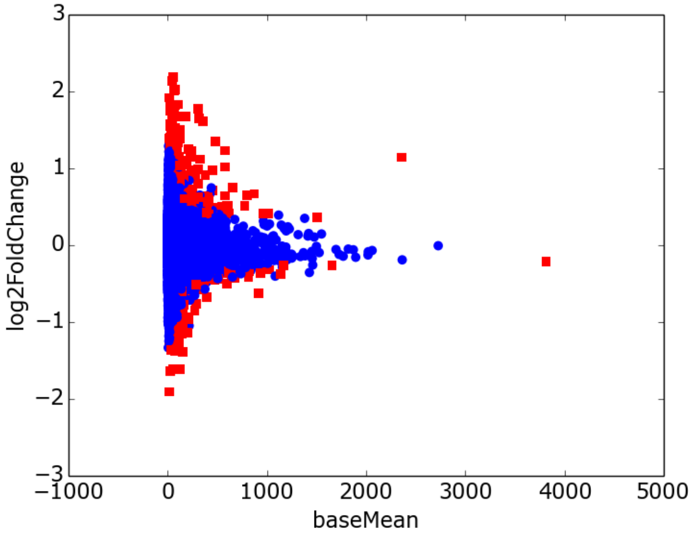

Figure 13: DESeq2 analysis for 2010 bacterial data. x-axis is geometric mean of abundance for genes across libraries. y-axis is the log base 2 of the fold change between oxic and suboxic zones. Greater than 0 on y-axis indicates higher expression in suboxic zones, less than 0 indicates higher expression in oxic zones. Each point represents a COG or SSU gene; blue circles = padj > 0.05 (not significant), red squares padj < 0.05 (significant).

Comparison of previous results for narG gene

Previous work on the ETSP OMZ has highlighted specific functional pathways when transitioning from the oxic to suboxic zone including: ammonia oxidation, ammonium transport, anaerobic nitrogen metabolism, and sulfur energy metabolism (Stewart et al (2012), Stewart (2011), Canfield et al (2010)). To advocate for the reliability of our pipeline for functional annotations, we included a brief comparison of one of the DA COGs from the suboxic zone. Transcripts of *narG* (COG5013), a gene that codes for the alpha sub-unit of dissimilatory nitrate reductase, increased with depth and transition to the OMZ-core (Stewart et al (2012)). Our DA analysis found that *narG* had the highest base mean of any DA gene for the suboxic zone as compared to oxic samples (Table 5). As expected from the overall taxonomic distribution of DA genes, *narG* annotations were primarily placed under the Proteobacteria Division, approximately 84% of reads. The Class breakdown of Proteobacteria for oxic and suboxic revealed that a major contributor to the differences in gene abundance between zones were the Gammaproteobacteria, supporting the previous findings for this data-set (Figure 14)(Stewart et al (2012), Stewart (2011)). Yet, the abundance distributions alone do not paint a clear picture of the significance of Gammaproteobacteria, due to the similar increases in abundance for all other Classes from oxic to suboxic zones. Visualizing the phylogenetic information in a KR heat tree, a function from the pplacer suite, for *narG* gave a better perspective on key taxonomic groups for oxic versus suboxic (Figure 15)(Evans and Matsen (2012)). A KR heat tree visualizes only the areas of a tree which differ in placement distribution between zones. In both oxic and suboxic, Gammaproteobacteria contributed to the overall differences in placement distributions on the tree. The highest abundance classifications in the suboxic zone from the Gammaproteobacteria were the Family Ectothiorhodospiraceae (purple sulfur bacteria) and unclassified Gammaproteobacteria (Table 6). This investigation of the *narG* gene in this data-set has supported the previous studies, highlighting the importance of sulfur oxidizing bacteria in anaerobic nitrogen metabolisms. However, reads for the suboxic zone were placed in high-level internal nodes, observable on the KR heat tree, underlining the need for further investigation of the functional contributions of this Class in OMZ anaerobic nitrogen metabolism and how this functional pathway might be coupled with sulfur oxidation (Figure 15).

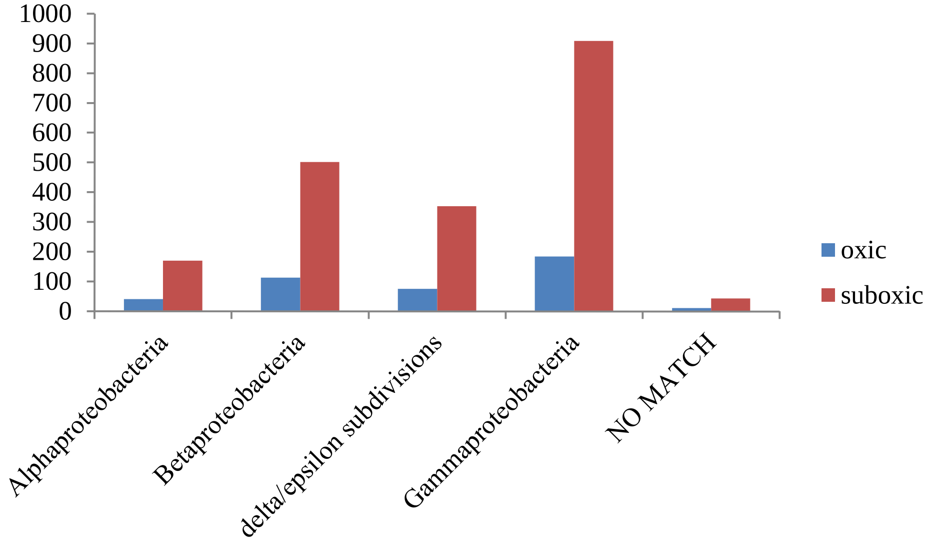

Figure 14: Confident read counts of oxic and suboxic zones for the DA gene *narG* (COG5013), broken down into Proteobacterial Classes. Counts normalized to largest sample library.

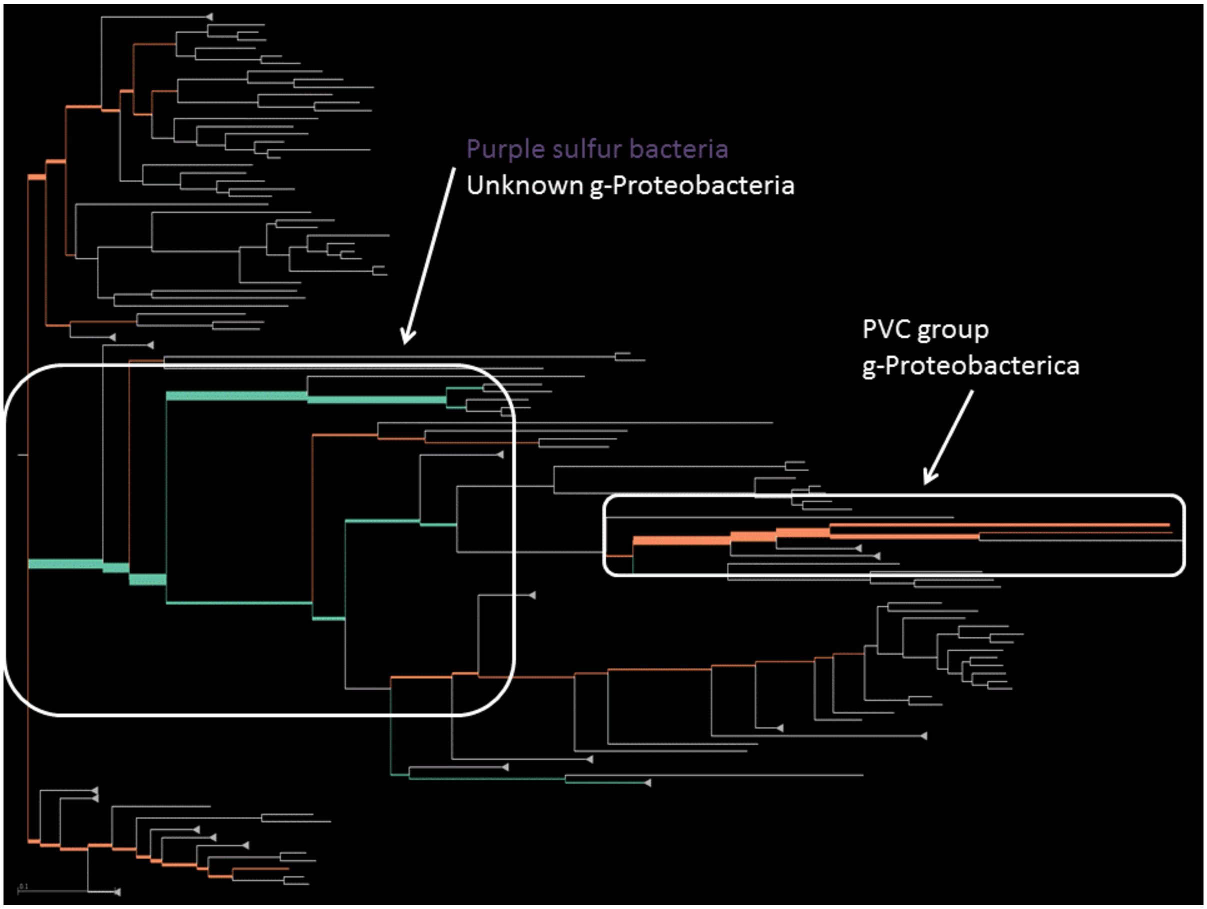

Figure 15: KR heat tree of *narG* gene for oxic (orange) vs suboxic (blue). Thickness of edges indicates number of placements from ETSP OMZ.

Diversity analysis of DA COGs

DA analysis revealed patterns of diversity for DA COGs differing from that of all COGs combined. A paired student t-test showed that the DA COGs had significantly lower average AWPD when compared to total COGs for 2008 and 2009, with p-values of 0.015 and 0.036 respectively (Figure 15). The 2010 samples showed higher average AWPD in the DA COGs when compared to total COGs, with a p-value of 0.017 (Figure 15).

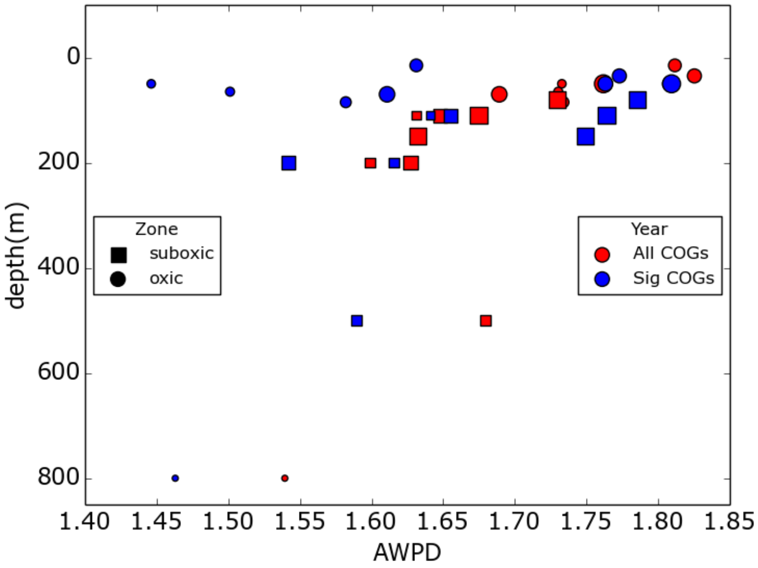

Figure 16: Bacterial average AWPD for all COG genes and DA genes from DESeq2 analysis for the 2008-2010 data by depth; color = All or DA COGs, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

The average AWPD for DA genes was different than all COGs, but the differences were not the same for each year. The 2010 DA COGs had the least number of sample depths and the highest sequencing effort, which contributed to a diversity trend similar to the combined COGs. The range of AWPD for the DA COGs is 1.45 to 1.82 with both minimum and maximum located in the oxic zone, 2008 and 2010 respectively. Overall AWPD decreases with depth, with 2010 showing the most uniform trend. In both 2008 and 2009 AWPD increases at the oxic-suboxic transition, 110m, then decrease until their lowest sample (Figure 15). The trend in diversity for DA COGs is also similar to total COGs, although all depths have a higher diversity for DA COGs. Differentially abundant genes may represent functions with a higher diversity than the average diversity of all functional genes.

Conversely, this may be evidence that functional genes with high diversity are likely to pertain to important functions in a specific environment.

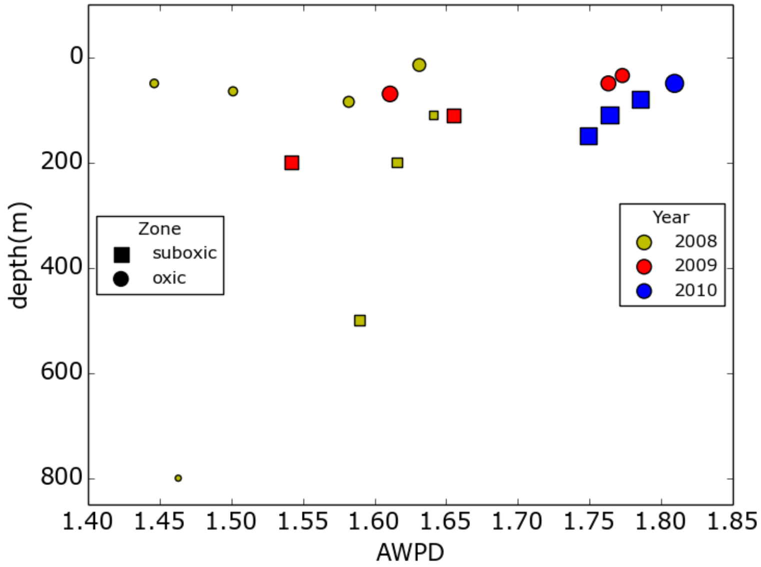

Figure 17: Bacterial average AWPD for significant COG genes found to be significantly differentially abundant for the 2008-2010 data by depth; color = year, shape = zone, size = read count for library. Suboxic threshold = <5umol/L dissolved O2.

Effects of sequencing effort

The PD for SSUs showed evidence of influence by sequencing effort, with the least effort (2008) having the lowest diversity, followed by medium effort (2009), and finally most effort (2010) with the highest overall AWPD (Figure 6). Neither the SSU nor the all COG average AWPD showed signs of being influenced by sequencing effort (Figure 7, 8, 9, and 10). The DA COGs for Bacteria, however, did have the 2010 samples grouping in the higher AWPD region of the graph. The lower sampling effort years had lower AWPD for all samples, with the exception being the 35m and 50m samples for 2009 (Figure 15). Variation of AWPD between years seemed to be reduced for the overall COGs, while the DA COGs were influenced by sequencing effort. The fact that the 2010 bacterial DA COGs are very similar to the diversity scores for overall COGs in Bacteria, might hint at the possibility of an identifiable sequencing threshold for this data-set.

CONCLUSIONS

The analysis of AWPD for three gene-sets has produced promising results supporting the use of functional housekeeping genes for studies in diversity. Measures in bacterial diversity for the SSU genes supported previously published trends of the OMZ community (Bryant et al (2012)). The trends in overall and bacterial diversity for all COGs are similar to SSUs suggesting that the functional genes used in this analysis can serve to answer the same question of diversity as the traditional marker genes. In the 2010 samples, diversity was highest in DA COGs, followed by all COGs, and finally SSUs, hinting at untapped novel diversity in the functional genes. This is also supported by the findings in the sequencing effort section, where increased effort leads to identification of increase diversity. These results are evidence that suggests the DA COGs not only have more ecological significance to community function, but may also be more sensitive to novel diversity.

The functional and taxonomic annotations, as well as the DA analysis results, for the *narG* gene agree with previous work supporting the efficacy of the PAW/DAP. High abundance of sulfur oxidizing bacteria, such as Ectothiorhodospiraceae stresses the importance of these organisms in anaerobic regions of the OMZ.

The DESeq2 comparison method identified functional genes to be differentially abundant between the oxic and suboxic regions of the ETSP OMZ. This is an important result, as these genes represent ecologically important functions. The AWPD of DA genes was lower for the lower sampling efforts, but higher for the highest effort when compared to the all COGs AWPDs. This may support a minimum sequencing threshold for the functional genes in this community somewhere between the 2009 and 2010 sampling effort.

Phylogenetic diversity of functional genes shows promised as a alternative method to measure the total diversity of an ecosystem. In all cases the functional AWPD was higher than the SSU AWPD, although the trends for Bacteria remained similar for both gene categories. This suggests that by using the COGs for measuring AWPD, more novel diversity of the community is detected. As biodiversity can be directly related to ecosystem stability and recovery, characterizing novel diversity is an important step to understanding the overall ecology of a community.

FUTURE DIRECTIONS

This study included an in-depth analysis of bacteria in an OMZ because the available data-set and reference packages were bacteria-centric. A future study that would add significantly to further testing of the PAW/DAP would be to use metagenomes sampled equally for all three Domains of life and viruses. This would allow for a more inclusive and encompassing test of the methods and capabilities of the pipeline.

The reference packages remain mostly generalized to the available reference information, but could be customized for very specific questions. An interesting experiment could include an organism specific package or a package built on a single protein domain instead of an entire gene. The ability to customize the packages via the included reference information allows for a large degree of flexibility in experimental design.

A deeper investigation of the functions of the Proteobacteria would be the next logical step for the functional aspect of this study. It was shown that Proteobacteria dominated the oxic and suboxic and that different Classes contributed to that overall primary position in the community. Further resolving the community composition by including more DA functional genes for alpha, beta, and gamma-proteobacterial classes could shed light on the community dynamics in an oxic versus suboxic zone.

Overall this study has helped to test a method built for rapid hypothesis testing on a large scale. Creating an analysis that combines taxonomic, functional, and phylogenetic annotation methods, such as the PAW/DAP, is vital to gaining a better understanding of the incredible diversity of microbial ecosystems. Resolving the role of biodiversity in the underlying mechanisms driving community functions will assist in future efforts to predict the effects of environmental variation on global ecosystems.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, et al. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* 430: 551-54

Altschul S.F. GWMWMEWLDJ. 1990. Basic local alignment search tool. In *J Mol Biol*, pp. 403-10

Armbrust BEV, Palumbi SR. 2015. Uncovering hidden worlds of ocean biodiversity. *Science (New York, N.Y.)* 348: 865-67

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015. GenBank. *Nucleic Acids Research* 43: D30-D35

Bryant JA, Stewart FJ, Eppley JM, DeLong EF. 2012. Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. *Ecology* 93: 1659-73

Cadotte MW, Hamilton Ma, Murray BR. 2009. Phylogenetic relatedness and plant invader success across two spatial scales. *Diversity and Distributions* 15: 481-88

Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, et al. 2010. A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. *Science (New York, N.Y.)* 330: 1375-78

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7: 335-36

Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, et al. 2009. Defining DNA-Based Operational Taxonomic Units for Microbial-Eukaryote Ecology. *Applied and Environmental Microbiology* 75: 5797-808

Codispoti La, Brandes Ja, Christensen JP, Devol aH, Naqvi SWa, et al. 2001. The oceanic fixed nitrogen and nitrous oxide budgets : Moving targets as we enter the anthropocene? *Scientia Marina* 65: 85-105

Darling AE, Jospin G, Lowe E, Matsen Fa, Bik HM, Eisen Ja. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2: e243

Diaz RJ, Rosenberg R. 2008. Spreading Consequences Dead for Marine. *Science (New York, N.Y.)* 321: 926-29

Eddy S. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755-63

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Research* 32: 1792-97

Evans SN, Matsen FA. 2012. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 74: 569-92

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1-10

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 39: W29-37

Flynn DFB, Mirotchnick N, Jain M, Palmer MI, Naeem S. 2011. Functional and phylogenetic diversity as predictors of biodiversity–ecosystem-function relationships. *Ecology* 92: 1573-81

Fuhrman JA. 2009. Microbial community structure and its functional implications. *Nature* 459: 193-99

Galloway JN, Dentener FJ, Capone DG, Boyer EW, Howarth RW, et al. 2004. Nitrogen cycles: Past, present, and future. *Biogeochemistry* 70: 153-226

Haft DH. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Research* 31: 371-73

Hamilton AJ. 2005. Species diversity or biodiversity? *Journal of Environmental Management* 75: 89-92

Hawley aK, Brewer HM, Norbeck aD, Pa a-Toli L, Hallam SJ. 2014. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proceedings of the National Academy of Sciences* 111: 11395-400

Hillebrand H, Bennett DM, Cadotte MW. 2007. Consequences of Dominance: A Review of Evenness Effects on Local and Regional Ecosystem Processes. *Ecology* 88: 1622-33

Hunter JD. 2007. Matplotlib: A 2D graphic environment. *Computing in Science & Engineering* 9: 90-95

Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Research* 17: 377-86

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012. Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science (New York, N.Y.)* 335: 587-90

Jolley Ka, Chan M-S, Maiden MCJ. 2004. mlstdbNet - distributed multi-locus sequence typing (MLST) databases. *BMC bioinformatics* 5: 86

Jonsson V, Nerman O, Kristiansson E. 2016. Statistical evaluation of methods for comparative metagenomics. *BMC Genomics*: 1-14

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology* 12

Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, et al. 2009. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Research* 37: 216-23

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, et al. 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research* 36: 2230-39

Land T, Fizzano P, Kodner R. 2015. Measuring cluster stability in a large scale phylogenetic analysis of functional genes in metagenomes using pplacer. In *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, pp. 1

Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. *Nucleic Acids Research* 39: D19-D21

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, et al. 2015. Determinants of community structure in the global plankton interactome. *Science (New York, N.Y.)* 348: 1262073_1-73_9

Logares R, Haverkamp THA, Kumar S, Lanzén A, Nederbragt AJ, et al. 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. *Journal of Microbiological Methods* 91: 106-13

Loreau M, Loreau M, Naeem S, Naeem S, Inchausti P, et al. 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science (New York, N.Y.)* 294: 804-8

Love MI, Anders S, Huber W. 2014. Differential analysis of count data - the DESeq2 package. 1-41

Matsen Fa, Evans SN. 2013. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PloS one* 8: e56859

Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11: 538

McCoy CO, Matsen Fa. 2013. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ* 1: e157

McMurdie PJ, Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology* 10

Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, et al. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386

Norberg J. 2013. Biodiversity and ecosystem functioning : A complex adaptive systems approach. *Limnology and Oceanography* 49: 1269-77

Ochman H, Lawrence JG, Groisman Ea. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405: 299-304

Oliphant TE. 2007. Python for Scientific Computing. *Comp Sci Eng* 9: 10-20

Paulmier A, Ruiz-Pino D. 2009. Oxygen minimum zones (OMZs) in the modern ocean. *Progress in Oceanography* 80: 113-28

Petchey OL, Hector A, Gaston KJ, Petchey OL, Hector A, Gaston KJ. 2004. HOW DO DIFFERENT MEASURES OF FUNCTIONAL DIVERSITY PERFORM ? *Ecology* 85: 847-57

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* 35: D61-D65

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41: D590-D96

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16: 276-77

Rosselló-Mora R, Amann R. 2001. The species concept for prokaryotes. *FEMS Microbiology Reviews* 25: 39-67

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology* 5: e77

Sanner MF. 1999. Python: a programming language for software integration and development. *Journal of molecular graphics & modelling* 17: 57-61

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. 2011. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 39: D38-D51

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* 75: 7537-41

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* 103: 12115-20

Stevens H, Ulloa O. 2008. Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. *Environmental Microbiology* 10: 1244-59

Stewart FJ. 2011. Dissimilatory sulfur cycling in oxygen minimum zones: an emerging metagenomics perspective. *Biochemical Society Transactions* 39: 1859-63

Stewart FJ, Ulloa O, DeLong EF. 2012. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environmental microbiology* 14: 23-40

Stramma L, Johnson GC, Sprintall J, Mohrholz V. 2008. Expanding Oxygen-Minimum. *Science (New York, N.Y.)* 2006: 2006-09

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, et al. 2015. Ocean plankton. Structure and function of the global ocean microbiome. *Science (New York, N.Y.)* 348: 1261359

Tatusov RL, Koonin EV, Lipman DJ. 2012. A Genomic Perspective on Protein Families. *Science (New York, N.Y.)* 631: 631-37

Thompson JR, Pacocha S, Pharino C, Klepac-ceraj V, Dana E, et al. 2005. Genotypic Diversity within a Natural Coastal Bacterioplankton Population. *Science (New York, N.Y.)* 307: 1311-13

Ulloa O, Canfield DE, DeLong EF, Letelier RM, Stewart FJ. 2012. Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences* 109: 15996-6003

Vargas Cd, Audic S, Henry N, Decelle J, Mahé F, et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science (New York, N.Y.)* 348: 1-11

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. 2004. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science (New York, N.Y.)* 304: 66-75

Villar E, Farrant GK, Follows M, Garczarek L, Speich S, et al. 2015. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science (New York, N.Y.)* 348: 1261447

Wommack KE, Bhavsar J, Ravel J. 2008. Metagenomics: Read length matters. *Applied and Environmental Microbiology* 74: 1453-63

Wright JJ, Konwar KM, Hallam SJ. 2012. Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology* 10: 381-94

Wyrtki K. 1962. The oxygen minima in relation to ocean circulation. *Deep Sea Research and Oceanographic Abstracts* 9: 11-23

Xu X, Passey T, Wei F, Saville R, Harrison RJ. 2015. Amplicon-based metagenomics identified candidate organisms in soils that caused yield decline in strawberry. *Horticulture Research* 2: 15022

Yeo SK, Huggett MJ, Eiler A, Rappé MS. 2013. Coastal bacterioplankton community dynamics in response to a natural disturbance. *PloS one* 8: e56207

APPENDIX:

Table 1: Reference package sequence count and stats by gene project and domain-level, domain columns are number of taxa.

	stat	Bacteria	Eukaryota	Archaea	seq_len	num_seqs
CHL	average	870.955056	322.803371	52.258427	165.342697	1247.8427
	max	2860	1604	204	1857	3540
	min	0	10	0	6	39
COG	average	611.184994	121.526044	44.4913187	190.312112	779.920628
	max	3170	1638	288	2219	3909
	min	0	0	0	5	4
MTH	average	106.869823	117.159763	10.0769231	216.852071	234.106509
	max	979	737	130	534	1413
	min	0	0	0	32	1
PTZ	average	515.402655	279.84292	54.0199115	197.225664	852.325221
	max	2613	1058	261	1584	3245
	min	0	1	0	8	1
SSU	average	3078.33333	1827.66667	211.666667	1941	5127.33333
	max	9234	5206	568	2733	9595
	min	0	2	0	1508	580
TIGR	average	587.006168	104.320998	37.1376507	212.186992	730.319596
	max	2728	1658	251	3162	3281
	min	0	0	0	5	1

Table 2: Metadata for OMZ metagenomes

sra_id	year	depth(m)	seq_type	lib_size	ave_read_len	zone
SRR304684	2008	15	DNA	771623	238	oxic
SRR064444	2008	50	DNA	341163	256	oxic
SRR304656	2008	65	DNA	382821	251	oxic
SRR064446	2008	85	DNA	569046	253	oxic
SRR064448	2008	110	DNA	380764	243	suboxic
SRR064450	2008	200	DNA	485911	249	suboxic
SRR304668	2008	500	DNA	515676	248	suboxic
SRR304683	2008	800	DNA	173051	242	oxic
SRR304671	2009	35	DNA	937420	333	oxic
SRR304672	2009	50	DNA	1042057	339	oxic
SRR070081	2009	70	DNA	1147856	385	oxic
SRR304673	2009	110	DNA	905059	403	suboxic
SRR070082	2009	200	DNA	930359	246	suboxic
SRR304674	2010	50	DNA	1530891	386	oxic
SRR070083	2010	80	DNA	1359823	428	suboxic
SRR304680	2010	110	DNA	1456854	409	suboxic
SRR070084	2010	150	DNA	1301664	431	suboxic

Table 3: Bacteria 2008 differentially abundant genes w/ padj < 0.05, green are oxic, blue are suboxic

gene	zone	baseMean	log2FoldChange	functional_description	Padj
COG4338	oxic	6.715856171	-2.616524044	Uncharacterized_protein_conserved_in_bacteria	7.67E-04
COG3067	oxic	23.67836869	-2.342707063	Na+/H+_antiporter	4.95E-03
COG3476	oxic	8.845768204	-2.131064639	Tryptophan-rich_sensory_protein__mitochondrial_benzodiazepine_receptor_homolog	1.02E-02
COG3223	oxic	12.01465024	-2.120301553	Predicted_membrane_protein	8.70E-03
COG3496	oxic	20.64512487	-2.002298323	Uncharacterized_conserved_protein	1.25E-02
COG5454	oxic	5.627329439	-1.960376118	Predicted_secreted_protein	1.61E-02
COG3380	oxic	3.303176453	-1.950106751	Predicted_NAD/FAD-dependent_oxidoreductase	2.28E-02
COG3564	oxic	9.466037371	-1.928383393	Uncharacterized_protein_conserved_in_bacteria	2.28E-02
COG2907	oxic	33.28331514	-1.873376171	Predicted_NAD/FAD-binding_protein	2.28E-02
COG1485	oxic	26.65717818	-1.821086455	Predicted_ATPase	5.28E-03
COG2509	oxic	5.035571655	-1.810211967	Uncharacterized_FAD-dependent_dehydrogenases	3.34E-02
COG4635	oxic	3.225175535	-1.790133692	Flavodoxin	4.36E-02
COG0586	oxic	2.648603235	-1.789178593	Uncharacterized_membrane-associated_protein	4.36E-02
COG4787	oxic	3.112455405	-1.785289398	Flagellar_basal_body_rod_protein	4.36E-02
COG1733	oxic	10.7604238	-1.711144289	Predicted_transcriptional_regulators	2.28E-02
COG2941	oxic	23.88544273	-1.676737876	Ubiquinone_biosynthesis_protein_COQ7	8.30E-03
COG2249	oxic	8.335793376	-1.667900675	Putative_NADPH-quinone_reductase__modulator_of_drug_activity_B	3.14E-02
COG3752	oxic	18.08386226	-1.664861956	Predicted_membrane_protein	1.02E-02
COG2855	oxic	29.10266248	-1.525323104	Predicted_membrane_protein	3.63E-02
COG1054	oxic	29.69075752	-1.516026679	Predicted_sulfurtransferase	6.06E-03
COG1805	oxic	43.49485151	-1.244912271	Na+-transporting_NADH_ubiquinone_oxidoreductase__subunit_NqrB	3.53E-02
COG3565	oxic	24.56484925	-1.220783937	Predicted_dioxygenase_of_extradiol_dioxygenase_family	1.37E-02
COG4531	oxic	36.27232983	-1.099535756	ABC-type_Zn2+_transport_system__periplasmic_component/surface_adhesin	3.49E-02
COG2076	oxic	29.89720607	-1.046457166	Membrane_transporters_of_cations_and_cationic_drugs	3.23E-02
COG0397	oxic	74.74743312	-1.041359996	Uncharacterized_conserved_protein	3.38E-02
COG1953	oxic	59.50706095	-1.007925726	Cytosine/uracil/thiamine/allantoin_permeases	7.74E-03
COG2609	oxic	224.1616686	-0.814590538	Pyruvate_dehydrogenase_complex__dehydrogenase__E1__component	9.31E-03

Table 3: continued

COG1233	oxic	62.83422785	-0.805517302	Phytoene_dehydrogenase_and_related_proteins	3.30E-02
COG1194	oxic	69.01690221	-0.66708947	A/G-specific_DNA_glycosylase	3.36E-02
COG0765	oxic	109.1960823	-0.649182573	ABC-type_amino_acid_transport_system__permease_component	7.74E-03
COG0508	oxic	218.6666299	-0.438193382	Pyruvate/2-oxoglutarate_dehydrogenase_complex__dihydrolipoamide_acyltransferase__E2__component__and_related_enzymes	3.93E-02
COG0635	suboxic	62.52438079	0.690322406	Coproporphyrinogen_III_oxidase_and_related_Fe-S_oxidoreductases	3.19E-02
COG2870	suboxic	43.64629637	0.763395678	ADP-heptose_synthase__bifunctional_sugar_kinase/adenylyltransferase	3.57E-02
COG0007	suboxic	36.0892051	1.028693842	Uroporphyrinogen-III_methylase	9.31E-03
COG1883	suboxic	34.52441487	1.071121238	Na+-transporting_methylmalonyl-CoA/oxaloacetate_decarboxylase__beta_subunit	1.73E-02
COG0674	suboxic	103.3065075	1.076267254	Pyruvate_ferredoxin_oxidoreductase_and_related_2-oxoacid_ferredoxin_oxidoreductases__alpha_subunit	1.25E-02
COG3347	suboxic	55.06728128	1.112597673	Uncharacterized_conserved_protein	3.25E-03
COG0053	suboxic	26.40438016	1.193031752	Predicted_Co/Zn/Cd_cation_transporters	2.08E-02
COG1013	suboxic	75.94231288	1.239968961	Pyruvate_ferredoxin_oxidoreductase_and_related_2-oxoacid_ferredoxin_oxidoreductases__beta_subunit	6.74E-03
COG0758	suboxic	16.16644126	1.257415417	Predicted_Rossmann_fold_nucleotide-binding_protein_involved_in_DNA_uptake	4.36E-02
COG4864	suboxic	15.48721428	1.285924419	Uncharacterized_protein_conserved_in_bacteria	1.61E-02
COG2170	suboxic	44.95652611	1.330778505	Uncharacterized_conserved_protein	1.61E-02
COG1994	suboxic	16.02906553	1.371172331	Zn-dependent_proteases	3.30E-02
COG0685	suboxic	47.38101277	1.426752736	5_10-methylenetetrahydrofolate_reductase	6.31E-05
COG0658	suboxic	10.47666326	1.539667328	Predicted_membrane_metal-binding_protein	3.57E-02
COG1254	suboxic	6.041619138	1.591377104	Acylphosphatases	3.30E-02
COG2826	suboxic	19.4784619	1.618605454	Transposase_and_inactivated_derivatives__IS30_family	2.06E-02
COG3039	suboxic	12.71244391	1.662380016	Transposase_and_inactivated_derivatives__IS5_family	2.32E-02
COG1271	suboxic	14.07241765	1.677148862	Cytochrome_bd-type_quinol_oxidase__subunit_1	1.37E-02

Table 3: continued

COG4660	suboxic	6.658052413	1.703851759	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfE	3.35E-02
COG3328	suboxic	22.66493098	1.74680062	Transposase_and_inactivated_derivatives	1.02E-02
COG0826	suboxic	37.41398446	1.774496093	Collagenase_and_related_proteases	8.70E-03
COG3243	suboxic	43.75412262	1.833301409	Poly_3-hydroxyalkanoate__synthetase	5.62E-03
COG2180	suboxic	16.26969621	1.857372068	Nitrate_reductase_delta_subunit	1.86E-02
COG1964	suboxic	8.693461413	1.85776998	Predicted_Fe-S_oxidoreductases	2.28E-02
COG1355	suboxic	9.008339013	1.923933491	Predicted_dioxygenase	8.30E-03
COG3676	suboxic	9.71094718	2.103708295	Transposase_and_inactivated_derivatives	7.95E-03
COG5394	suboxic	10.36715833	2.290896028	Uncharacterized_protein_conserved_in_bacteria	7.67E-04
COG2963	suboxic	5.885282022	2.397067331	Transposase_and_inactivated_derivatives	1.08E-03
COG4656	suboxic	15.71582672	2.450677216	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfC	6.31E-05

Table 4: Bacteria 2009 differentially abundant genes w/ padj < 0.05, green are oxic, blue are suboxic.

gene	zone	baseMean	log2FoldChange	functional_description	Padj
COG0376	oxic	137.1072874	-3.076361504	Catalase__peroxidase_I	3.75E-05
COG3241	oxic	14.48784985	-2.815262514	Azurin	2.07E-04
COG1201	oxic	101.4558476	-2.74149716	Lhr-like_helicases	5.76E-05
COG3651	oxic	69.95687864	-2.571904296	Uncharacterized_protein_conserved_in_bacteria	1.33E-04
COG3489	oxic	16.4616228	-2.548161126	Predicted_periplasmic_lipoprotein	6.92E-04
COG2907	oxic	112.852164	-2.541280771	Predicted_NAD/FAD-binding_protein	4.95E-04
COG3496	oxic	59.16164737	-2.467059451	Uncharacterized_conserved_protein	1.09E-03
COG3670	oxic	38.15844913	-2.453281582	Lignostilbene-alpha_beta-dioxygenase_and_related_enzymes	6.78E-04
COG3476	oxic	34.38619977	-2.316578828	Tryptophan-rich_sensory_protein__mitochondrial_benzodiazepine_receptor_homolog	3.60E-03
COG1054	oxic	62.15334171	-2.283870939	Predicted_sulfurtransferase	1.43E-03
COG1398	oxic	46.50417567	-2.283803949	Fatty-acid_desaturase	3.75E-05
COG5135	oxic	25.29738858	-2.278474569	Uncharacterized_conserved_protein	1.48E-03
COG4121	oxic	18.94850999	-2.2325597	Uncharacterized_conserved_protein	4.84E-03
COG3502	oxic	20.19448329	-2.222745258	Uncharacterized_protein_conserved_in_bacteria	5.26E-03
COG2409	oxic	51.90738446	-2.191752744	Predicted_drug_exporters_of_the_RND_superfamily	7.47E-04
COG1562	oxic	79.67186414	-2.129900568	Phytoene/squalene_synthetase	2.73E-03
COG0369	oxic	44.00514783	-2.12725443	Sulfite_reductase__alpha_subunit__flavoprotein	8.98E-03
COG3239	oxic	50.40008309	-2.12034945	Fatty_acid_desaturase	3.82E-04
COG0346	oxic	17.09400805	-2.108412358	Lactoylglutathione_lyase_and_related_lyases	3.70E-03
COG4989	oxic	20.63737294	-2.10629536	Predicted_oxidoreductase	6.00E-03
COG2326	oxic	47.60345272	-2.061399828	Uncharacterized_conserved_protein	1.82E-04
COG3380	oxic	14.69728839	-2.055817419	Predicted_NAD/FAD-dependent_oxidoreductase	1.27E-02
COG2124	oxic	102.284878	-2.044517658	Cytochrome_P450	6.78E-04
COG3733	oxic	6.692635994	-2.022027557	Cu2+-containing_amine_oxidase	1.43E-02
COG1705	oxic	14.91634806	-2.009407639	Muramidase__flagellum-specific	1.45E-02
COG0397	oxic	180.9737162	-1.968356181	Uncharacterized_conserved_protein	4.48E-04
COG0415	oxic	149.9830568	-1.963552146	Deoxyribodipyrimidine_photolyase	1.81E-02
COG3752	oxic	53.13815051	-1.958491486	Predicted_membrane_protein	4.17E-03
COG2855	oxic	83.36148022	-1.95596418	Predicted_membrane_protein	8.70E-03
COG2107	oxic	35.08101877	-1.937165388	Predicted_periplasmic_solute-binding_protein	1.23E-02
COG4338	oxic	27.6528933	-1.906923046	Uncharacterized_protein_conserved_in_bacteria	2.27E-02
COG4270	Oxic	23.50059187	-1.875476875	Predicted_membrane_protein	2.08E-02

Table 4: continued

COG0387	oxic	16.85277934	-1.872291745	Ca2+/H+_antiporter	1.92E-02
COG1448	oxic	29.87376795	-1.851822648	Aspartate/tyrosine/aromatic_aminotransferase	2.33E-02
COG4454	oxic	6.439637977	-1.823010975	Uncharacterized_copper-binding_protein	3.29E-02
COG2717	oxic	45.69982685	-1.807816205	Predicted_membrane_protein	1.25E-02
COG3128	oxic	21.16418631	-1.793604383	Uncharacterized_iron-regulated_protein	2.54E-02
COG2941	oxic	62.58435197	-1.79227569	Ubiquinone_biosynthesis_protein_COQ7	6.44E-03
COG1914	oxic	17.72057849	-1.786979919	Mn2+_and_Fe2+_transporters_of_the_NRAMP_family	8.06E-03
COG2268	oxic	34.88004032	-1.779338172	Uncharacterized_protein_conserved_in_bacteria	7.93E-03
COG5515	oxic	9.936259101	-1.761326965	Uncharacterized_conserved_small_protein	3.87E-02
COG5184	oxic	106.0794387	-1.759579424	Alpha-tubulin_suppressor_and_related_RCC1_domain-containing_proteins	2.73E-02
COG2035	oxic	59.44920501	-1.756877418	Predicted_membrane_protein	3.82E-04
COG2309	oxic	43.25508618	-1.74933687	Leucyl_aminopeptidase__aminopeptidase_T	1.33E-02
COG1679	oxic	22.93664303	-1.730260972	Uncharacterized_conserved_protein	3.60E-03
COG0027	oxic	25.12619856	-1.700180508	Formate-dependent_phosphoribosylglycinamide_formyltransferase__GAR_transformylase	3.58E-02
COG4772	oxic	72.95230794	-1.692668211	Outer_membrane_receptor_for_Fe3+-dicitrate	2.57E-02
COG3046	oxic	192.7187622	-1.686133579	Uncharacterized_protein_related_to_deoxyribodipyrimidine_photolyase	4.79E-02
COG3104	oxic	66.13878427	-1.680733262	Dipeptide/tripeptide_permease	8.98E-03
COG1222	oxic	14.99639651	-1.679412846	ATP-dependent_26S_proteasome_regulatory_subunit	3.58E-02
COG4445	oxic	12.08945386	-1.668251623	Hydroxylase_for_synthesis_of_2-methylthio-cis-ribozeatin_in_tRNA	4.79E-02
COG1786	oxic	11.09642335	-1.64964576	Uncharacterized_conserved_protein	4.07E-02
COG0855	oxic	54.63169784	-1.640392656	Polyphosphate_kinase	3.41E-03
COG3509	oxic	32.52527483	-1.632595617	Poly_3-hydroxybutyrate__depolymerase	3.58E-02
COG4623	oxic	21.18953963	-1.630571317	Predicted_soluble_lytic_transglycosylase_fused_to_an_ABC-type_amino_acid-binding_protein	4.61E-02
COG4067	oxic	17.03410523	-1.626317902	Uncharacterized_protein_conserved_in_archaea	3.81E-02
COG3540	oxic	25.35299684	-1.597451318	Phosphodiesterase/alkaline_phosphatase_D	4.59E-02
COG2802	oxic	55.92329593	-1.589304817	Uncharacterized_protein__similar_to_the_N-terminal_domain_of_Lon_protease	1.43E-02

Table 4: continued

COG1796	oxic	21.16584281	-1.585171249	DNA_polymerase_IV__family_X	4.07E-02
COG3491	oxic	103.6562566	-1.583968339	Isopenicillin_N_synthase_and_related_dioxygenases	4.95E-04
COG3000	oxic	101.7443279	-1.578374193	Sterol_desaturase	3.40E-03
COG3565	oxic	73.18501609	-1.576601628	Predicted_dioxygenase_of_extradiol_dioxygenase_family	1.22E-02
COG2041	oxic	62.66954538	-1.549710893	Sulfite_oxidase_and_related_enzymes	3.33E-03
COG3825	oxic	70.28410493	-1.549234883	Uncharacterized_protein_conserved_in_bacteria	4.09E-03
COG0464	oxic	45.86456253	-1.545248657	ATPases_of_the_AAA+_class	1.49E-02
COG0561	oxic	26.12619784	-1.543978756	Predicted_hydrolases_of_the_HAD_superfamily	1.51E-02
COG4276	oxic	28.25098652	-1.531917757	Uncharacterized_conserved_protein	1.41E-02
COG4558	oxic	24.30019811	-1.531045714	ABC-type_hemin_transport_system__periplasmic_component	4.79E-02
COG3555	oxic	29.00662316	-1.529253346	Aspartyl/asparaginyl_beta-hydroxylase_and_related_dioxygenases	3.71E-02
COG3425	oxic	43.97079448	-1.525751206	3-hydroxy-3-methylglutaryl_CoA_synthase	4.85E-02
COG1443	oxic	18.10983095	-1.517967816	Isopentenyldiphosphate_isomerase	2.27E-02
COG2317	oxic	98.62611875	-1.516025417	Zn-dependent_carboxypeptidase	1.02E-02
COG2820	oxic	25.13682214	-1.484381129	Uridine_phosphorylase	1.65E-02
COG1946	oxic	32.77331547	-1.473191309	Acyl-CoA_thioesterase	1.42E-02
COG2013	oxic	14.76661248	-1.468266033	Uncharacterized_conserved_protein	4.40E-02
COG4233	oxic	24.40063333	-1.452086263	Uncharacterized_protein_predicted_to_be_involved_in_C-type_cytochrome_biogenesis	2.54E-02
COG2947	oxic	86.29830123	-1.437941819	Uncharacterized_conserved_protein	1.66E-03
COG2115	oxic	30.42395996	-1.435257631	Xylose_isomerase	4.94E-02
COG5524	oxic	117.7191333	-1.40198752	Bacteriorhodopsin	3.58E-02
COG1279	oxic	62.33103903	-1.380948654	Lysine_efflux_permease	4.79E-02
COG3340	oxic	18.46720694	-1.378415039	Peptidase_E	4.90E-02
COG1363	oxic	35.14085134	-1.348772664	Cellulase_M_and_related_proteins	4.97E-02
COG2301	oxic	130.1932271	-1.336675111	Citrate_lyase_beta_subunit	1.66E-03
COG3818	oxic	32.72921656	-1.319542818	Predicted_acetyltransferase__GNAT_superfamily	3.81E-02
COG2175	oxic	117.6353099	-1.304611103	Probable_taurine_catabolism_dioxygenase	1.23E-02
COG2308	oxic	25.52670055	-1.295569113	Uncharacterized_conserved_protein	3.98E-02
COG1629	oxic	323.1643759	-1.293835276	Outer_membrane_receptor_proteins__mostly_Fe_transport	4.07E-02
COG0523	oxic	31.27058677	-1.269460248	Putative_GTPases__G3E_family	1.90E-02
COG1292	oxic	89.35340448	-1.266083793	Choline-glycine_betaine_transporter	2.08E-02
COG1164	oxic	69.44206938	-1.257666348	Oligoendopeptidase_F	4.20E-03

Table 4: continued

COG1726	oxic	84.53905328	-1.232487308	Na+-transporting_NADH_ubiquinone_oxidoreductase__subunit_NqrA	3.58E-02
COG1404	oxic	188.8421269	-1.215956666	Subtilisin-like_serine_proteases	4.85E-02
COG1233	oxic	234.2388274	-1.199042528	Phytoene_dehydrogenase_and_related_proteins	8.32E-03
COG0507	oxic	23.58587363	-1.19835905	ATP-dependent_exoDNAse__exonuclease_V___alpha_subunit_-_helicase_superfamily_I_member	4.07E-02
COG0657	oxic	36.89778486	-1.171914732	Esterase/lipase	1.05E-02
COG1070	oxic	39.10833661	-1.138215521	Sugar__pentulose_and_hexulose__kinases	2.57E-02
COG1172	oxic	154.2736484	-1.128584123	Ribose/xylose/arabinose/galactoside_ABC-type_transport_systems__permease_components	2.90E-04
COG2070	oxic	128.057176	-1.124508908	Dioxygenases_related_to_2-nitropropane_dioxygenase	1.32E-02
COG4760	oxic	60.99938002	-1.121651147	Predicted_membrane_protein	1.27E-02
COG2165	oxic	36.15407341	-1.106763001	Type_II_secretory_pathway__pseudopilin_PulG	3.06E-02
COG4638	oxic	209.6333487	-1.10344989	Phenylpropionate_dioxygenase_and_related_ring-hydroxylating_dioxygenases__large_terminal_subunit	6.96E-03
COG0733	oxic	120.1585029	-1.09761573	Na+-dependent_transporters_of_the_SNF_family	1.24E-02
COG4341	oxic	80.18589875	-1.096736193	Predicted_HD_phosphohydrolase	1.12E-02
COG2076	oxic	104.7638217	-1.083190552	Membrane_transporters_of_cations_and_cationic_drugs	3.96E-02
COG2154	oxic	59.62985004	-1.082617535	Pterin-4a-carbinolamine_dehydratase	3.79E-02
COG4147	oxic	276.0213096	-1.007464626	Predicted_symporter	2.08E-02
COG1879	oxic	55.238737	-1.002966152	ABC-type_sugar_transport_system__periplasmic_component	3.58E-02
COG4152	oxic	125.8760091	-0.975567556	ABC-type_uncharacterized_transport_system__ATPase_component	9.09E-03
COG0423	oxic	152.9377641	-0.964835898	Glycyl-tRNA_synthetase__class_II	3.70E-03
COG1953	oxic	132.0447677	-0.952363214	Cytosine/uracil/thiamine/allantoin_permeases	1.82E-02
COG3396	oxic	61.52562261	-0.931121423	Uncharacterized_conserved_protein	4.79E-02
COG3962	oxic	154.8974778	-0.85899205	Acetolactate_synthase	4.87E-02
COG1301	oxic	115.0774498	-0.808503907	Na+/H+-dicarboxylate_symporters	1.61E-02
COG1129	oxic	127.5751008	-0.764920441	ABC-type_sugar_transport_system__ATPase_component	3.71E-02
COG3119	oxic	542.6716379	-0.750995946	Arylsulfatase_A_and_related_enzymes	1.35E-04

Table 4: continued

COG1344	oxic	105.8107444	-0.745546064	Flagellin_and_related_hook-associated_proteins	1.61E-02
COG1834	oxic	130.8456495	-0.719878374	N-Dimethylarginine_dimethylaminohydrolase	1.22E-02
COG1429	oxic	93.78589314	-0.685696	Cobalamin_biosynthesis_protein_CobN_and_related_Mg-chelatases	4.57E-02
COG4102	oxic	112.0837573	-0.673199135	Uncharacterized_protein_conserved_in_bacteria	4.79E-02
COG5285	oxic	188.9103311	-0.646116524	Protein_involved_in_biosynthesis_of_mitomycin_antibiotics/polyketide_fumonisin	1.41E-02
COG0765	oxic	294.6073657	-0.528933362	ABC-type_amino_acid_transport_system__permease_component	3.06E-02
COG1508	oxic	169.4686571	-0.519586895	DNA-directed_RNA_polymerase_specialized_sigma_subunit__sigma54_homolog	2.00E-02
COG0811	oxic	222.5661822	-0.447863761	Biopolymer_transport_proteins	4.94E-02
COG0667	oxic	282.0770751	-0.425709515	Predicted_oxidoreductases__related_to_aryl-alcohol_dehydrogenases	3.44E-02
COG1960	oxic	1204.516428	-0.366749414	Acyl-CoA_dehydrogenases	4.79E-02
COG1024	oxic	744.7151203	-0.350588055	Enoyl-CoA_hydratase/carnithine_racemase	1.43E-02
COG0001	suboxic	367.5134247	0.387110313	Glutamate-1-semialdehyde_aminotransferase	2.45E-02
COG0574	suboxic	588.7338647	0.415598406	Phosphoenolpyruvate_synthase/pyruvate_phosphate_dikinase	4.79E-02
COG0542	suboxic	670.7159898	0.427353197	ATPases_with_chaperone_activity__ATP-binding_subunit	3.81E-02
COG1560	suboxic	192.0948594	0.463243712	Lauroyl/myristoyl_acyltransferase	4.87E-02
COG0696	suboxic	173.0740538	0.50760073	Phosphoglyceromutase	4.79E-02
COG2951	suboxic	182.5706231	0.514552204	Membrane-bound_lytic_murein_transglycosylase_B	4.94E-02
COG1932	suboxic	262.996577	0.577540373	Phosphoserine_aminotransferase	1.12E-02
COG0559	suboxic	560.8681826	0.591473657	Branched-chain_amino_acid_ABC-type_transport_system__permease_components	1.10E-02
COG0347	suboxic	145.7350338	0.636262142	Nitrogen_regulatory_protein_PII	4.07E-02
COG0135	suboxic	105.1454596	0.643501408	Phosphoribosylanthranilate_isomerase	2.54E-02
COG3914	suboxic	287.851967	0.646391341	Predicted_O-linked_N-acetylglucosamine_transferase__SPINDLY_family	5.99E-03
COG0156	suboxic	240.5946558	0.654882147	7-keto-8-aminopelargonate_synthetase_and_related_enzymes	4.79E-02
COG1066	suboxic	179.9788785	0.764767933	Predicted_ATP-dependent_serine_protease	1.99E-04

Table 4: continued

COG2878	suboxic	66.03230251	0.769065737	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfB	4.79E-02
COG5016	suboxic	138.5053282	0.807010885	Pyruvate/oxaloacetate_carboxyltransferase	5.19E-03
COG0352	suboxic	92.46047248	0.82142053	Thiamine_monophosphate_synthase	3.33E-03
COG2918	suboxic	117.7390477	0.851158	Gamma-glutamylcysteine_synthetase	3.34E-02
COG3347	suboxic	163.6583151	0.92726746	Uncharacterized_conserved_protein	4.79E-02
COG0422	suboxic	103.5591409	0.92762105	Thiamine_biosynthesis_protein_ThiC	3.33E-03
COG0635	suboxic	199.2551618	0.929891691	Coproporphyrinogen_III_oxidase_and_related_Fe-S_oxidoreductases	1.23E-02
COG0213	suboxic	118.4905159	0.930672145	Thymidine_phosphorylase	2.76E-02
COG4108	suboxic	134.829429	0.93200336	Peptide_chain_release_factor_RF-3	1.43E-03
COG1883	suboxic	109.3316013	0.960667175	Na+-transporting_methylmalonyl-CoA/oxaloacetate_decarboxylase__beta_subunit	3.65E-02
COG2518	suboxic	143.8435222	0.982033567	Protein-L-isoaspartate_carboxylmethyltransferase	2.54E-02
COG1636	suboxic	39.54697878	0.996919163	Uncharacterized_protein_conserved_in_bacteria	2.65E-02
COG1015	suboxic	63.42150289	1.073107241	Phosphopentomutase	1.23E-02
COG4137	suboxic	46.19293641	1.127437197	ABC-type_uncharacterized_transport_system__permease_component	1.82E-02
COG2046	suboxic	127.6963248	1.130558542	ATP_sulfurylase__sulfate_adenylyltransferase	3.32E-02
COG3954	suboxic	51.78636757	1.146036705	Phosphoribulokinase	8.00E-03
COG4579	suboxic	85.67333296	1.21747386	Isocitrate_dehydrogenase_kinase/phosphatase	3.34E-02
COG2923	suboxic	21.81877353	1.252034539	Uncharacterized_protein_involved_in_the_oxidation_of_intracellular_sulfur	4.29E-02
COG1469	suboxic	54.51446857	1.260516861	Uncharacterized_conserved_protein	6.92E-04
COG3114	suboxic	20.48984122	1.274777367	Heme_exporter_protein_D	4.59E-02
COG2920	suboxic	41.74110579	1.344661902	Dissimilatory_sulfite_reductase__desulfoviridin___gamma_subunit	8.98E-03
COG3205	suboxic	49.32580561	1.347736661	Predicted_membrane_protein	4.17E-03
COG5014	suboxic	13.67067112	1.486426231	Predicted_Fe-S_oxidoreductase	2.84E-02
COG2914	suboxic	21.27433656	1.486947366	Uncharacterized_protein_conserved_in_bacteria	1.51E-02
COG2922	suboxic	31.12777782	1.488516511	Uncharacterized_protein_conserved_in_bacteria	3.58E-02
COG3931	suboxic	29.83612838	1.491455525	Predicted_N-formylglutamate_amidohydrolase	1.09E-03
COG2833	suboxic	29.34471327	1.496031665	Uncharacterized_protein_conserved_in_bacteria	1.48E-02
COG3749	suboxic	48.45335335	1.54103816	Uncharacterized_protein_conserved_in_bacteria	3.02E-02
COG1415	suboxic	21.90563438	1.589279069	Uncharacterized_conserved_protein	1.73E-02

Table 4: continued

COG2168	suboxic	16.95437432	1.652663361	Uncharacterized_conserved_protein_involved_in_oxidation_of_intracellular_sulfur	3.65E-02
COG1687	suboxic	8.379195063	1.658458447	Predicted_branched-chain_amino_acid_permeases__azaleucine_resistance	4.79E-02
COG3260	suboxic	7.767104255	1.676314423	Ni_Fe-hydrogenase_III_small_subunit	4.32E-02
COG5456	suboxic	15.26804473	1.745471804	Predicted_integral_membrane_protein_linked_to_a_cation_pump	1.43E-02
COG4660	suboxic	41.01660852	1.897123397	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfE	1.12E-02
COG2069	suboxic	6.480769712	2.045273734	CO_dehydrogenase/acetyl-CoA_synthase_delta_subunit__corrinoid_Fe-S_protein	1.31E-02

Table 5: Bacteria 2010 differentially abundant genes w/ padj < 0.05, green are oxic, blue are suboxic.

gene	zone	baseMean	log2FoldChange	functional_description	Padj
COG3502	oxic	16.45184745	-1.907140039	Uncharacterized_protein_conserved_in_bacteria	1.02E-03
COG3476	oxic	24.41989512	-1.63304352	Tryptophan-rich_sensory_protein__mitochondrial_benzodiazepine_receptor_homolog	9.27E-03
COG2907	oxic	124.8300223	-1.611874721	Predicted_NAD/FAD-binding_protein	2.19E-03
COG3651	oxic	50.29077522	-1.608118027	Uncharacterized_protein_conserved_in_bacteria	1.74E-04
COG1953	oxic	147.8817945	-1.384624891	Cytosine/uracil/thiamine/allantoin_permeases	3.11E-05
COG0376	oxic	74.39333801	-1.381095211	Catalase__peroxidase_I	3.43E-02
COG5135	oxic	31.82656426	-1.366967171	Uncharacterized_conserved_protein	3.72E-02
COG3489	oxic	15.32295356	-1.319586772	Predicted_periplasmic_lipoprotein	4.63E-02
COG4240	oxic	55.80466612	-1.314532507	Predicted_kinase	3.54E-03
COG5524	oxic	108.9713356	-1.271498692	Bacteriorhodopsin	2.15E-04
COG2249	oxic	32.41772427	-1.251691104	Putative_NADPH-quinone_reductase__modulator_of_drug_activity_B	3.72E-02
COG2941	oxic	65.01396516	-1.19782361	Ubiquinone_biosynthesis_protein_COQ7	7.58E-03
COG4341	oxic	143.031502	-1.154204475	Predicted_HD_phosphohydrolase	1.06E-02
COG4365	oxic	34.06363455	-1.149949508	Uncharacterized_protein_conserved_in_bacteria	4.44E-02
COG2175	oxic	196.9267431	-1.13850384	Probable_taurine_catabolism_dioxygenase	5.95E-03
COG2820	oxic	53.65622034	-1.135630429	Uridine_phosphorylase	7.03E-03
COG0346	oxic	24.74381319	-1.122170372	Lactoylglutathione_lyase_and_related_lyases	4.77E-02
COG1562	oxic	75.72129148	-1.118995878	Phytoene/squalene_synthetase	4.32E-02
COG3492	oxic	70.12517138	-1.088919887	Uncharacterized_protein_conserved_in_bacteria	3.23E-03
COG0266	oxic	109.365358	-1.051690517	Formamidopyrimidine-DNA_glycosylase	1.49E-02
COG3491	oxic	148.9145457	-1.041185554	Isopenicillin_N_synthase_and_related_dioxygenases	1.54E-02
COG3104	oxic	137.2305561	-0.978681835	Dipeptide/tripeptide_permease	3.53E-02
COG2072	oxic	212.7532155	-0.96555752	Predicted_flavoprotein_involved_in_K+_transport	1.06E-02
COG2076	oxic	121.9908127	-0.937651822	Membrane_transporters_of_cations_and_cationic_drugs	6.18E-03
COG3000	oxic	194.4879837	-0.932666914	Sterol_desaturase	4.41E-02
COG0386	oxic	153.0790618	-0.903062248	Glutathione_peroxidase	4.14E-02
COG1794	oxic	82.22762088	-0.892828258	Aspartate_racemase	1.38E-02
COG2154	oxic	95.72597631	-0.854659638	Pterin-4a-carbinolamine_dehydratase	2.24E-02
COG1194	oxic	272.5556876	-0.845816982	A/G-specific_DNA_glycosylase	3.06E-05

Table 5: continued

COG2130	oxic	161.7892751	-0.772690569	Putative_NADP-dependent_oxidoreductases	3.31E-02
COG0408	oxic	283.3632177	-0.763722383	Coproporphyrinogen_III_oxidase	7.21E-03
COG0678	oxic	101.2833765	-0.719479088	Peroxiredoxin	3.72E-02
COG0232	oxic	201.3670187	-0.718042872	dGTP_triphosphohydrolase	3.56E-02
COG4147	oxic	392.9028765	-0.674447737	Predicted_symporter	3.22E-02
COG0235	oxic	156.0032756	-0.665976004	Ribulose-5-phosphate_4-epimerase_and_related_epimerases_and_aldolases	3.05E-02
COG0785	oxic	120.580467	-0.629922123	Cytochrome_c_biogenesis_protein	4.94E-02
COG2609	oxic	913.9550409	-0.626472821	Pyruvate_dehydrogenase_complex__dehydrogenase__E1__component	2.04E-02
COG0694	oxic	200.7499083	-0.611303236	Thioredoxin-like_proteins_and_domains	2.38E-02
COG4215	oxic	143.7527902	-0.605076648	ABC-type_arginine_transport_system__permease_component	4.96E-02
COG1494	oxic	229.7747878	-0.589773758	Fructose-1_6-bisphosphatase/sedoheptulose_1_7-bisphosphatase_and_related_proteins	1.46E-02
COG0489	oxic	316.0303291	-0.57185993	ATPases_involved_in_chromosome_partitioning	1.54E-02
COG0423	oxic	321.5372874	-0.569880783	Glycyl-tRNA_synthetase__class_II	2.29E-02
COG2352	oxic	369.7466633	-0.539840749	Phosphoenolpyruvate_carboxylase	6.22E-03
COG4642	oxic	290.6401361	-0.508153353	Uncharacterized_protein_conserved_in_bacteria	1.83E-02
COG0114	oxic	593.3703787	-0.498181937	Fumarase	8.81E-04
COG1192	oxic	354.4559971	-0.472409421	ATPases_involved_in_chromosome_partitioning	2.78E-02
COG3288	oxic	475.0244479	-0.450077929	NAD/NADP_transhydrogenase_alpha_subunit	2.08E-02
COG2021	oxic	430.4294555	-0.442556986	Homoserine_acetyltransferase	2.93E-02
COG0667	oxic	507.0632302	-0.434042231	Predicted_oxidoreductases__related_to_aryl-alcohol_dehydrogenases	3.74E-02
COG1282	oxic	672.0045554	-0.424380069	NAD/NADP_transhydrogenase_beta_subunit	3.34E-03
COG0206	oxic	507.6555745	-0.421264162	Cell_division_GTPase	1.28E-02
COG0044	oxic	815.3592287	-0.413897777	Dihydroorotase_and_related_cyclic_amidohydrolases	3.57E-03
COG0036	oxic	455.9160584	-0.401697368	Pentose-5-phosphate-3-epimerase	2.38E-02
COG1178	oxic	639.5047681	-0.388921704	ABC-type_Fe3+_transport_system__permease_component	1.15E-02
COG4221	oxic	1137.057779	-0.37421169	Short-chain_alcohol_dehydrogenase_of_unknown_specificity	1.54E-02
COG0074	oxic	606.5618638	-0.366048226	Succinyl-CoA_synthetase__alpha_subunit	2.08E-02
COG5009	oxic	931.7567334	-0.364334253	Membrane_carboxypeptidase/penicillin-binding_protein	6.41E-03

Table 5: continued

COG0504	oxic	931.5906825	-0.34342577	CTP_synthase__UTP-ammonia_lyase	1.12E-02
COG0719	oxic	815.9843183	-0.336931365	ABC-type_transport_system_involved_in_Fe-S_cluster_assembly__permease_component	1.08E-02
COG0652	oxic	611.8840658	-0.336457607	Peptidyl-prolyl_cis-trans_isomerase__rotamase__-_cyclophilin_family	4.41E-02
COG0495	oxic	1026.64821	-0.318513053	Leucyl-tRNA_synthetase	1.06E-02
COG4770	oxic	1166.801798	-0.261357012	Acetyl/propionyl-CoA_carboxylase__alpha_subunit	2.96E-02
COG0187	oxic	1657.225141	-0.26053036	Type_IIA_topoisomerase__DNA_gyrase/topo_II__topoisomerase_IV___B_subunit	2.25E-02
COG1012	oxic	3814.07097	-0.213872678	NAD-dependent_aldehyde_dehydrogenases	3.13E-02
COG1529	suboxic	1506.963773	0.366724196	Aerobic-type_carbon_monoxide_dehydrogenase__large_subunit_CoxL/CutL_homologs	9.59E-04
COG3894	suboxic	414.2557923	0.405695755	Uncharacterized_metal-binding_protein	4.41E-02
COG0156	suboxic	617.813662	0.411701666	7-keto-8-aminopelargonate_synthetase_and_related_enzymes	4.63E-02
COG1410	suboxic	1013.598062	0.412248805	Methionine_synthase_I__cobalamin-binding_domain	1.20E-03
COG5598	suboxic	959.0046854	0.415223082	Trimethylamine_corrinoid_methyltransferase	3.09E-03
COG5557	suboxic	397.251433	0.429082532	Polysulphide_reductase	3.41E-02
COG1319	suboxic	396.4344012	0.437136582	Aerobic-type_carbon_monoxide_dehydrogenase__middle_subunit_CoxM/CutM_homologs	3.20E-02
COG0790	suboxic	320.7576996	0.452572136	FOG__TPR_repeat__SEL1_subfamily	4.63E-02
COG0146	suboxic	594.6966492	0.478104737	N-methylhydantoinase_B/acetone_carboxylase__alpha_subunit	1.06E-02
COG0635	suboxic	455.200649	0.489772632	Coproporphyrinogen_III_oxidase_and_related_Fe-S_oxidoreductases	3.67E-02
COG0243	suboxic	548.1506924	0.505698088	Anaerobic_dehydrogenases__typically_selenocysteine-containing	1.78E-02
COG2217	suboxic	598.4696138	0.510670642	Cation_transport_ATPase	2.72E-03
COG0659	suboxic	610.5423588	0.512501203	Sulfate_permease_and_related_transporters__MFS_superfamily	1.60E-02
COG1778	suboxic	252.6044303	0.512742274	Low_specificity_phosphatase__HAD_superfamily	3.53E-02

Table 5: continued

COG0145	suboxic	776.2733359	0.514511172	N-methylhydantoinase_A/acetone_carboxylase__beta_subunit	4.10E-04
COG2010	suboxic	289.4116283	0.537857005	Cytochrome_c__mono-_and_diheme_variants	3.31E-02
COG1760	suboxic	280.5414124	0.543493632	L-serine_deaminase	3.56E-02
COG4231	suboxic	399.6133853	0.563476116	Indolepyruvate_ferredoxin_oxidoreductase__alpha_and_beta_subunits	7.58E-03
COG4106	suboxic	232.2189414	0.574499114	Trans-aconitate_methyltransferase	3.70E-02
COG1066	suboxic	417.1613065	0.587199393	Predicted_ATP-dependent_serine_protease	2.08E-02
COG0502	suboxic	158.7820999	0.596337698	Biotin_synthase_and_related_enzymes	4.51E-02
COG1858	suboxic	233.5938014	0.59693419	Cytochrome_c_peroxidase	2.26E-02
COG3213	suboxic	266.4371037	0.600398635	Uncharacterized_protein_involved_in_response_to_NO	4.69E-02
COG0422	suboxic	284.5969897	0.60417653	Thiamine_biosynthesis_protein_ThiC	1.21E-02
COG5012	suboxic	167.1682998	0.607009844	Predicted_cobalamin_binding_protein	3.70E-02
COG1042	suboxic	375.3384912	0.612156472	Acyl-CoA_synthetase__NDP_forming	2.16E-03
COG0007	suboxic	268.5297405	0.621903838	Uroporphyrinogen-III_methylase	4.13E-02
COG2518	suboxic	310.6653871	0.627705086	Protein-L-isoaspartate_carboxylmethyltransferase	4.41E-02
COG0339	suboxic	399.7791648	0.633949747	Zn-dependent_oligopeptidases	7.91E-03
COG0612	suboxic	578.9170866	0.643035323	Predicted_Zn-dependent_peptidases	2.03E-02
COG2956	suboxic	148.3480687	0.64344995	Predicted_N-acetylglucosaminyl_transferase	3.56E-02
COG4145	suboxic	796.5793699	0.647016637	Na+/panthothenate_symporter	6.42E-03
COG0674	suboxic	873.255474	0.668637692	Pyruvate_ferredoxin_oxidoreductase_and_related_2-oxoacid_ferredoxin_oxidoreductases__alpha_subunit	7.58E-03
COG1030	suboxic	212.5016291	0.673449296	Membrane-bound_serine_protease__ClpP_class	1.28E-02
COG4191	suboxic	186.2750941	0.675545483	Signal_transduction_histidine_kinase_regulating_C4-dicarboxylate_transport_system	1.09E-02
COG1251	suboxic	222.1254689	0.683675436	NAD_P_H-nitrite_reductase	1.65E-02
COG0804	suboxic	248.290185	0.695323805	Urea_amidohydrolase__urease__alpha_subunit	4.63E-02
COG0043	suboxic	451.2269262	0.712500379	3-polyprenyl-4-hydroxybenzoate_decarboxylase_and_related_decarboxylases	7.59E-05
COG3401	suboxic	117.9036394	0.719235005	Fibronectin_type_3_domain-containing_protein	3.72E-02
COG0053	suboxic	248.4150535	0.725944158	Predicted_Co/Zn/Cd_cation_transporters	7.58E-03

Table 5: continued

COG3316	Suboxic	207.8142561	0.72640592	Transposase_and_inactivated_derivatives	3.87E-03
COG1013	suboxic	652.3044364	0.748813684	Pyruvate_ferredoxin_oxidoreductase_and_related_2-oxoacid_ferredoxin_oxidoreductases__beta_subunit	1.02E-02
COG0701	suboxic	182.5127819	0.772032073	Predicted_permeases	1.11E-02
COG1541	suboxic	130.8895193	0.784876551	Coenzyme_F390_synthetase	1.46E-02
COG1797	suboxic	175.7123278	0.786240911	Cobyrinic_acid_a_c-diamide_synthase	4.46E-02
COG2186	suboxic	155.4553112	0.787437696	Transcriptional_regulators	4.49E-02
COG1199	suboxic	309.7791696	0.797733103	Rad3-related_DNA_helicases	4.72E-04
COG1015	suboxic	156.1982838	0.79885878	Phosphopentomutase	3.67E-02
COG0378	suboxic	76.77630945	0.81174172	Ni2+-binding_GTPase_involved_in_regulation_of_expression_and_maturation_of_urease_and_hydrogenase	4.65E-02
COG3164	suboxic	143.7388404	0.815873822	Predicted_membrane_protein	2.82E-02
COG1951	suboxic	174.3655571	0.834743632	Tartrate_dehydratase_alpha_subunit/Fumarate_hydratase_class_I__N-terminal_domain	6.18E-03
COG4977	suboxic	127.1539037	0.837069898	Transcriptional_regulator_containing_an_amidase_domain_and_an_AraC-type_DNA-binding_HTH_domain	1.28E-02
COG0062	suboxic	147.9576459	0.84679905	Uncharacterized_conserved_protein	1.68E-02
COG0671	suboxic	83.6319824	0.848438329	Membrane-associated_phospholipid_phosphatase	2.96E-02
COG3001	suboxic	113.8486973	0.870359436	Fructosamine-3-kinase	2.56E-02
COG2210	suboxic	165.7991287	0.888940586	Uncharacterized_conserved_protein	7.58E-03
COG0641	suboxic	138.1125677	0.907343286	Arylsulfatase_regulator__Fe-S_oxidoreductase	2.69E-02
COG1148	suboxic	166.0609196	0.907905465	Heterodisulfide_reductase__subunit_A_and_related_polyferredoxins	3.72E-02
COG0651	suboxic	276.6482051	0.912883136	Formate_hydrogenlyase_subunit_3/Multisubunit_Na+/H+_antiporter__MnhD_subunit	1.61E-02
COG3243	suboxic	379.4735139	0.914183256	Poly_3-hydroxyalkanoate__synthetase	1.28E-02
COG1002	suboxic	59.44249907	0.915914671	Type_II_restriction_enzyme__methylase_subunits	4.63E-02
COG3039	suboxic	134.868933	0.924900075	Transposase_and_inactivated_derivatives__IS5_family	6.41E-03
COG1139	suboxic	237.2191588	0.926201461	Uncharacterized_conserved_protein_containing_a_ferredoxin-like_domain	1.99E-04
COG3155	suboxic	77.75538437	0.927800181	Uncharacterized_protein_involved_in_an_early_stage_of_isoprenoid_biosynthesis	4.75E-02
COG4585	suboxic	65.85110555	0.941359883	Signal_transduction_histidine_kinase	3.31E-02
COG0829	Suboxic	62.8967203	0.953136069	Urease_accessory_protein_UreH	3.56E-02

Table 5: continued

COG2374	suboxic	78.06394976	0.958304317	Predicted_extracellular_nuclease	2.56E-02
COG3696	suboxic	450.9244539	0.97366906	Putative_silver_efflux_pump	9.59E-04
COG1014	suboxic	242.9620063	0.981798185	Pyruvate_ferredoxin_oxidoreductase_and_related_2-oxoacid_ferredoxin_oxidoreductases__gamma_subunit	1.78E-02
COG0370	suboxic	303.3672546	0.991191426	Fe2+_transport_system_protein_B	2.15E-04
COG1838	suboxic	115.6404773	1.002226553	Tartrate_dehydratase_beta_subunit/Fumarate_hydratase_class_I__C-terminal_domain	5.55E-03
COG3301	suboxic	72.49782226	1.011923562	Formate-dependent_nitrite_reductase__membrane_component	3.05E-02
COG2223	suboxic	575.0062344	1.019643168	Nitrate/nitrite_transporter	4.58E-04
COG4674	suboxic	121.2999159	1.035369729	Uncharacterized_ABC-type_transport_system__ATPase_component	1.28E-02
COG3415	suboxic	94.41625548	1.03670787	Transposase_and_inactivated_derivatives	1.13E-02
COG2861	suboxic	90.83896747	1.039564212	Uncharacterized_protein_conserved_in_bacteria	1.28E-02
COG2206	suboxic	136.6817329	1.040839304	HD-GYP_domain	7.58E-03
COG0758	suboxic	155.2698765	1.049203483	Predicted_Rossmann_fold_nucleotide-binding_protein_involved_in_DNA_uptake	6.18E-03
COG2221	suboxic	191.9724321	1.075623697	Dissimilatory_sulfite_reductase__desulfoviridin___alpha_and_beta_subunits	8.19E-03
COG3850	suboxic	112.4809307	1.079756847	Signal_transduction_histidine_kinase__nitrate/nitrite-specific	2.42E-02
COG2963	suboxic	126.4690493	1.080008709	Transposase_and_inactivated_derivatives	6.78E-04
COG1648	suboxic	155.5009503	1.095081896	Siroheme_synthase__precorrin-2_oxidase/ferrochelatase_domain	1.46E-02
COG3481	suboxic	92.74974046	1.101667806	Predicted_HD-superfamily_hydrolase	1.49E-02
COG1140	suboxic	321.8660829	1.117921282	Nitrate_reductase_beta_subunit	7.65E-07
COG1964	suboxic	112.6226843	1.123464201	Predicted_Fe-S_oxidoreductases	1.21E-03
COG0622	suboxic	68.24352888	1.128567231	Predicted_phosphoesterase	6.42E-03
COG3524	suboxic	47.25687665	1.133216962	Capsule_polysaccharide_export_protein	4.17E-02
COG5013	suboxic	2361.237269	1.144758361	Nitrate_reductase_alpha_subunit	1.09E-05
COG3676	suboxic	111.971246	1.148608125	Transposase_and_inactivated_derivatives	6.32E-04
COG3328	suboxic	228.9558027	1.152886629	Transposase_and_inactivated_derivatives	6.22E-05
COG2003	suboxic	57.13911599	1.160601924	DNA_repair_proteins	1.72E-02
COG1661	Suboxic	45.07518707	1.189298432	Predicted_DNA-binding_protein_with_PD1-like_DNA-binding_motif	3.46E-02

Table 5: continued

COG3945	suboxic	22.03104666	1.223695751	Uncharacterized_conserved_protein	4.68E-02
COG2048	suboxic	79.23952308	1.228781183	Heterodisulfide_reductase__subunit_B	4.22E-03
COG4656	suboxic	241.785204	1.229024765	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfC	1.14E-03
COG1032	suboxic	576.6975848	1.23019254	Fe-S_oxidoreductase	3.06E-03
COG0095	suboxic	77.75025456	1.230560664	Lipoate-protein_ligase_A	2.16E-03
COG3323	suboxic	29.0726488	1.240801074	Uncharacterized_protein_conserved_in_bacteria	4.49E-02
COG2826	suboxic	206.822268	1.247347866	Transposase_and_inactivated_derivatives__IS30_family	2.96E-07
COG5441	suboxic	95.68145721	1.264665206	Uncharacterized_conserved_protein	2.80E-03
COG0648	suboxic	42.8376689	1.272137302	Endonuclease_IV	2.08E-02
COG4284	suboxic	25.69276313	1.275193149	UDP-glucose_pyrophosphorylase	3.46E-02
COG1896	suboxic	16.84528006	1.282161727	Predicted_hydrolases_of_HD_superfamily	4.77E-02
COG4520	suboxic	28.32527676	1.297901089	Surface_antigen	3.65E-02
COG4242	suboxic	20.06456017	1.302361186	Cyanophycinase_and_related_exopeptidases	4.17E-02
COG2362	suboxic	79.63971849	1.306790352	D-aminopeptidase	9.07E-04
COG2516	suboxic	46.27274902	1.306851518	Biotin_synthase-related_enzyme	7.83E-03
COG4659	suboxic	95.73665748	1.311396147	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfG	5.69E-03
COG3303	suboxic	49.93692765	1.314327879	Formate-dependent_nitrite_reductase__periplasmic_cytochrome_c552_subunit	8.92E-03
COG3260	suboxic	13.59829708	1.322049201	Ni_Fe-hydrogenase_III_small_subunit	4.63E-02
COG0658	suboxic	107.360061	1.339754167	Predicted_membrane_metal-binding_protein	1.81E-03
COG1413	suboxic	40.41823611	1.352684392	FOG__HEAT_repeat	7.50E-03
COG0826	suboxic	478.6845573	1.354997288	Collagenase_and_related_proteases	4.86E-06
COG3439	suboxic	36.41033064	1.356867509	Uncharacterized_conserved_protein	3.05E-02
COG3154	suboxic	27.59338471	1.357406389	Putative_lipid_carrier_protein	3.72E-02
COG2069	suboxic	12.62828972	1.370670716	CO_dehydrogenase/acetyl-CoA_synthase_delta_subunit__corrinoid_Fe-S_protein	3.72E-02
COG4657	suboxic	122.739191	1.385849016	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfA	6.32E-04
COG1882	suboxic	98.13976482	1.386247725	Pyruvate-formate_lyase	4.72E-04
COG2044	suboxic	39.18647592	1.387488208	Predicted_peroxiredoxins	6.83E-03
COG4113	suboxic	18.07121119	1.388200009	Predicted_nucleic_acid-binding_protein__contains_PIN_domain	2.77E-02
COG2414	suboxic	85.56275727	1.394289645	Aldehyde_ferredoxin_oxidoreductase	3.09E-03
COG0374	suboxic	29.07884994	1.404687724	Ni_Fe-hydrogenase_I_large_subunit	1.54E-02
COG0282	suboxic	30.22152521	1.406979951	Acetate_kinase	2.29E-02
COG3531	suboxic	70.63568242	1.409256014	Predicted_protein-disulfide_isomerase	6.21E-03

Table 5: continued

COG2333	suboxic	79.03643895	1.439971843	Predicted_hydrolase__metallo-beta-lactamase_superfamily	3.21E-03
COG1180	suboxic	119.7231802	1.496415311	Pyruvate-formate_lyase-activating_enzyme	7.59E-05
COG2703	suboxic	31.02874044	1.5109909	Hemerythrin	7.03E-03
COG0831	suboxic	55.2880772	1.518225579	Urea_amidohydrolase__urease__gamma_subunit	3.34E-03
COG1150	suboxic	75.35961457	1.519993624	Heterodisulfide_reductase__subunit_C	2.89E-04
COG3261	suboxic	23.05788963	1.55419918	Ni_Fe-hydrogenase_III_large_subunit	1.46E-02
COG1614	suboxic	21.89084621	1.573270628	CO_dehydrogenase/acetyl-CoA_synthase_beta_subunit	1.28E-02
COG3379	suboxic	48.5569514	1.590475169	Uncharacterized_conserved_protein	3.54E-03
COG2316	suboxic	47.07975878	1.616106625	Predicted_hydrolase__HD_superfamily	3.45E-03
COG1775	suboxic	358.3925684	1.616770782	Benzoyl-CoA_reductase/2-hydroxyglutaryl-CoA_dehydratase_subunit__BcrC/BadD/HgdB	1.38E-10
COG2354	suboxic	44.84232354	1.651988062	Uncharacterized_protein_conserved_in_bacteria	7.03E-03
COG1924	suboxic	316.9554153	1.654092584	Activator_of_2-hydroxyglutaryl-CoA_dehydratase__HSP70-class_ATPase_domain	3.62E-07
COG2116	suboxic	113.0338457	1.675504953	Formate/nitrite_family_of_transporters	6.32E-05
COG4658	suboxic	182.4340933	1.677706026	Predicted_NADH_ubiquinone_oxidoreductase__subunit_RnfD	3.70E-05
COG1856	suboxic	69.13946379	1.696312061	Uncharacterized_homolog_of_biotin_synthetase	1.47E-04
COG3464	suboxic	26.00858162	1.750322739	Transposase_and_inactivated_derivatives	3.57E-03
COG3581	suboxic	25.3009934	1.768534398	Uncharacterized_protein_conserved_in_bacteria	3.69E-03
COG4584	suboxic	309.7622515	1.779214372	Transposase_and_inactivated_derivatives	2.65E-18
COG0426	suboxic	71.02463753	1.820637208	Uncharacterized_flavoproteins	7.27E-05
COG3436	suboxic	101.9972679	1.831889282	Transposase_and_inactivated_derivatives	3.62E-07
COG3580	suboxic	20.88635146	1.875946284	Uncharacterized_protein_conserved_in_bacteria	2.19E-03
COG0650	suboxic	18.72650354	1.918835406	Formate_hydrogenlyase_subunit_4	1.56E-03
COG2403	suboxic	61.5800439	2.011418954	Predicted_GTPase	3.22E-04
COG2006	suboxic	73.46082711	2.024429482	Uncharacterized_conserved_protein	4.77E-06
COG3005	suboxic	47.4838218	2.141415988	Nitrate/TMAO_reductases__membrane-bound_tetraheme_cytochrome_c_subunit	7.28E-05
COG3335	suboxic	56.09394728	2.184973525	Transposase_and_inactivated_derivatives	9.08E-06

Table 6: Counts Proteobacterial Classes (columns) by lowest taxonomic classification (rows) for suboxic zone, counts are normalized by largest sample library.

Lowest classification	Alpha	Beta	delta/epsilon	Gamma	NO MATCH
Acetobacteraceae	30				
Acidovorax		2			
Alcaligenaceae		1			
Alcanivoracaceae				11	
Alcanivorax				1	
Alphaproteobacteria	29				
Alteromonadaceae				14	
Alteromonadales				1	
Anaeromyxobacter			6		
Betaproteobacteria		38			
Bradyrhizobiaceae	1				
Brucellaceae	3				
Burkholderia		30			
Burkholderiaceae		2			
Burkholderiales		2			
Burkholderiales Genera incertae sedis		0			
Caulobacteraceae	1				
Chromatiales				0	
Chromobacteriaceae		3			
Comamonadaceae		1			
Cronobacter				1	
Cupriavidus		1			
delta/epsilon subdivisions			179		
Deltaproteobacteria			75		
Desulfobacteraceae			10		
Dickeya				0	
Ectothiorhodospiraceae				527	
Enterobacter				2	
Enterobacter cloacae				1	
Enterobacteriaceae				3	
Gammaproteobacteria				27	
Geobacter			14		
Geobacteraceae			69		
Hahella				1	

Table 6: continued

Halomonas				6	
Hyphomicrobium	2				
Hyphomicrobium denitrificans	34				
Hyphomonadaceae	4				
Marinobacter				3	
Methylobacterium	1				
Microbulbifer				47	
Oxalobacteraceae		1			
Pandoraea		0			
Proteobacteria					42
Providencia				1	
Pseudomonadales				0	
Pseudomonas				3	
Ralstonia solanacearum		1			
Rhizobiales	6				
Rhodobacteraceae	5				
Rhodocyclaceae		133			
Rhodospirillaceae	5				
Rhodospirillales	48				
Roseobacter	1				
Serratia				1	
Shewanella				2	
Sutterellaceae		30			
Thioalkalivibrio				25	
Thiobacillus		256			
Thiomonas		1			
unclassified Gammaproteobacteria				221	
Vibrionaceae				4	
Xanthomonadaceae				6	