8-2018

Microbial Changes in Gene Expression Level in Response to Environmental Conditions in the Delaware Bay

Margi I. Patel
Clemson University, margip@g.clemson.edu

MICROBIAL CHANGES IN GENE EXPRESSION LEVEL IN RESPONSE TO ENVIRONMENTAL CONDITIONS IN THE DELAWARE BAY

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Microbiology

by
Margi I. Patel
August 2018

Accepted by:
Dr. Barbara Campbell, Committee Chair
Dr. J. Michael Henson
Dr. Harry D. Kurtz, Jr.

ABSTRACT

Bacteria dominate in abundance, diversity and potentially metabolic activity in many environments. Our current knowledge on the influence of specific individual taxa on these processes is largely lacking. To bridge these gaps, I chose three near complete metagenome-assembled genomes (MAGs) from the Delaware Bay, phylogenetically associated with the Roseobacter clade, to compare the functional potential of the MAGs to their closest relatives. I also characterized the relative activity of one MAG by using normalized gene expression levels and differential gene expression. The normalized number of transcripts per sample revealed whether or not specific genes/pathways were being expressed at the time of sampling. In all of the different conditions that the samples were collected from, a high number of transcripts related to membrane transporters, energy metabolism and ribosomal proteins were observed for MAG 22. Differential expression was observed between environmental conditions including season, time of day and salinity. For differential gene expression, the significantly up or down regulation of gene transcription between environmental conditions was characterized to visualize any patterns in metabolism. My overall results indicate that the organism remains active throughout the year, however, the types of physiology it utilizes changes based on the conditions present.

DEDICATION

To my family

ACKNOWLEDGMENTS

I would firstly like to thank my advisor, Dr. Barbara J. Campbell, and my committee members Dr. John Michael Henson and Dr. Harry D. Kurtz, Jr. for their support and guidance throughout this project. I would also like to thank all of the members of the Campbell lab group for creating a positive and productive environemnt to work in. Thanks also to my fellow graduate students for always being willing to help. I would also like to acknowledge the NSF (OCE-1261359) and JGI CSP-1621, this work could not be done without their support. Finally, thank you to all the people that have supported me over the journey of preparing my thesis.

TABLE OF CONTENTS

Page

TITLE PAGE i

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF ABBREVIATION ix

CHAPTER

I. INTRODUCTION 1

The Delaware Bay 1
Microbial communities in estuarine environments 1
Interaction of bacteria with OM, DOM and POM 2
The Roseobacter clade 2
Current modes of measuring abundance and activity of marine bacteria 3
Meta-omics of prokaryotes 4

II. HYPOTHESIS AND OBJECTIVES OF STUDY 6

III. EXPERIMENTAL METHODS 7

Experimental design flowchart 7
Work completed by others in Campbell lab 8
Sample collection and sequencing 8
Work completed by me 9
MAG annotations 9
Phylogenomic tree construction & two-way amino acid identity 10

Table of Contents (Continued)

Page

Comparison of functional potential and transcripts 10
Sequence mapping and generating count tables 11
RNA-seq analyses 12

IV. EXPERIMENTAL RESULTS 13

General description/characteristics of MAGs 13
Phylogenomics and potential abundance of all three MAGs 15
Comparative genomics-general 19
Comparative genomics of MAG 22 21
Differential gene expression analyses of MAG 22 23

V. DISCUSSION 31

VI. CONCLUSIONS/FUTURE DIRECTIONS 35

APPENDICES 52

A: Scripts used 53
B: Unique genes in each MAG – supplemental tables 55

REFERENCES 64

LIST OF TABLES

Table Page

1. General features of samples sequenced for metagenomics and mettranscriptomic analyses. 9

2. General genomic features of the three MAGs of interest and their two closest relatives 14

3. Differentially expressed genes in MAG 22 between two seasons (summer vs. fall) 36

4. Differentially expressed genes in MAG 22 between two times points in the summer (night vs. day (11:00 pm vs. 11:00 am)) 39

5. Differentially expressed genes in MAG 22 between two salinities within spring (medium vs. high (20 vs. 30 PSU)) 42

S1a. Presence in Spr20L08 MAG 22 but not in closet relative *Planktomarina temperata* RCA23 55

S1b. Presence in Sum29NL08 MAG 147 but not in closest relative *Rhodobacteraceae bacterium* HIMB11 57

S1c. Presence in Sum29DG08 MAG 73 not in closest relative *Rhodobacteraceae bacterium* HIMB11 59

LIST OF FIGURES

Figure Page

1. Flowchart demonstrating the general experimental design and setup to analyze MAGs of interest..7

2. Phylogenomic tree showing MAGs analyzed in this study in relation to other roseobacters 15

3. Heat map of two-way amino acid identity calculated between all three MAGs assembled for this study and their two closest relatives...16

4. **A**. Coverage of all three MAGs across the metagenome samples
B. Index of replication for the three MAGs, indication of growth rates...18

5. Heat map showing the number of protein encoding genes present per MAG in each KEGG category 20

6. Metabolic potential for MAG 22 compared to its closest relative *Planktomarina temperata* RCA23...21

7. Principal Coordinates Analysis of metatranscriptome reads that mapped to MAG 22..23

8. Heatmap representation of transcripts per million base pairs for MAG 22..25

9. Heat map representation of the mean expression of MAG 22 transcripts from both summer and fall day 26

10. Heat map representation of the mean expression of MAG 22 transcripts from night and day summer samples 28

11. Heat map representation of the mean expression of MAG 22 transcripts across medium and high salinities in spring.........................30

LIST OF ABBREVIATION

AAI	–	Amino acid identity
AAnP	–	Aerobic anoxygenic phototrophy
DMSP	–	Dimethyksulfoniopropionate
DOM	–	Dissolved organic matter
HIMB11	–	*Rhosobacteraceae bacterium* HIMB11
iRep	–	Index of replication
KEGG	–	Kyoto encyclopedia of genes and genomes
MAGs	–	Metagenome assembeled genomes
Pegs	–	Protein encoding genes
POM	–	Particulate organic matter
RAST	–	Rapid annotations using subsystems technology
TPM	–	Transcripts per million

FIGURES

D	–	Day
GO8	–	Size fraction of 0.8 μm
L08	–	Size fraction of 0.2 μm
N	–	Night
PSU	–	Practical salinity unit
Spr	–	Spring
Sum	–	Summer

CHAPTER ONE

INTRODUCTION

The Delaware Bay

Estuaries are mixing zones where saline water from the open ocean mixes with fresh water from a river. The Delaware Bay is a well-studied estuary and has one of the steepest salinity gradients of the three largest urban estuaries in the U.S.A.[1] It is predominantly affected by tidal and wind action rather than by river flow. Because the estuary does not experience heavy river flow, the flushing time of the estuary is lengthy, taking anywhere from weeks to months[2]. The middle Atlantic shelf is the terminus for flushing of the Delaware Bay, and other estuaries of the northeast coast of United States, such as the Chesapeake Bay[2].

There is a high municipal sewage effluents and industrial input of organic matter and nutrients as well as terrestrial input from soil/plant detritus[1,3,4] into the bay. The Delaware Bay has an annually persistent turbidity maximum that shifts either up or down stream according to tidal and freshwater flow from the river[5]. Generally, the turbidity maximum occurs approximately 50 miles above the Delaware Bay mouth, and shifts slightly seawards with increased freshwater flow[2]. At the turbidity maximum, the estuary is light limiting owing to the high concentrations of suspended matter[1]. This light limitation in the upper reaches of the Delaware Bay limits primary production[6,7]. In contrast, decreased turbidity and increased primary production is observed in the lower bay[6,7].

Microbial communities in estuarine environments

Microbial communities vary tremendously along estuarine environmental gradients[8–10]. At the phylum level, Betaproteobacteria and Actinobacteria dominate the freshwater portion of the estuarine gradient while Alphaproteobacteria dominates the marine environment and *Bacteroidetes* seem to remain constant along the gradient[10–12]. The changes occurring in microbial communities along the estuarine transect are likely in response to variations in both biotic and abiotic factors as well as the complex interactions that microbes have with these factors[10,13]. Important factors that control microbial community composition and function in the Delaware Bay are salinity, light availability and organic matter (OM)[8].

Interaction of bacteria with OM, DOM and POM

Generally there is a high concentration of dissolved organic matter (DOM) in estuarine environments[14]. The governing factor for particulate organic matter (POM) concentration on the other hand is partly determined by the physical characteristics of the estuary[15]. For example, tidal estuaries, such as the Delaware Bay, are characterized by long residence times and in turn higher concentration of particles is present at the turbidity maximum zone[15]. POM is an important component of the total suspended matter in estuaries and consists of living biomass, as well as detrital organic and inorganic matter[16]. Dynamic exchange between DOM and POM is critical to the cycling of organic matter (OM) in coastal and inland aquatic ecosystems[17]. A significant portion of the OM synthesized by primary producers becomes DOM, about half of which is respired to CO_2 by microbes back into the atmosphere, and the rest remains dissolved in the water or is taken up again for photosynthesis[17]. There is constant cycling between DOM and POM. All bacterial taxa are not likely to equally contribute to the degradation of this matter as suggested by studies comparing at the activity of attached vs. free-living cells[18–21].

The Roseobacter clade

The Roseobacter clade is a group within the Alphaproteobacteria subclass of the *Proteobacteria*[22]. The first strain of the group was described in 1991[22], and at present there are at least 54 isolates described (see http://www.roseobase.org). Comparative analysis amongst 32 of these genomes indicate the members of this clade are ecological generalists, and can utilize a number of different pathways for carbon and energy metabolism[23].

The clade is one of the most prominent groups present in marine surface waters. Roseobacters are estimated to make up 10% of bacterial cells in the open ocean and up to 20% of bacterial cells in coastal waters[24–26]. The abundance of roseobacters varies greatly, with certain areas having a higher abundance than others. Aside from being abundant, the clade contains a diverse range of physiologies and metabolisms related to biogeochemical cycling. For example, R*uegeria pomeroyi* DSS-3, a representative of the Roseobacter clade, has been shown to be involved in dimethylsulfoniopropionate (DMSP) degradation[27]. Another member of the group, *Roseovarius* sp. TM 1035, is also able to degrade DMSP[28] as well as perform aerobic anoxygenic phototrophy (AAnP)[23]. In addition, members of the group may also perform sulfur metabolism[29], methylotrophy[30], mixotrophy,[31] carbon monoxide (CO) oxidation,[31,32] and aromatic compound degradation,[33] to name a few. This plasticity in carbon and energy metabolism allows the organisms of this clade to respond to a diverse range of environmental conditions[23].

Current modes of measuring abundance and activity of marine bacteria

Major advances occurred in the late 1970s and early 1980s in methods for quantifying the abundance of marine bacteria. Once epifluorescence microscopes became readily available, protocols based on settling cells onto membrane filters with blue-light-excited, green-fluorescing acridine orange were made available starting in 1974[34,35]. In 1980 however, UV-excited, blue-fluorescing DNA stain (DAPI), an alternate fluorochrome, became available for bacterial counts[36]. The advantage of using DAPI, versus acridine orange, is the decrease in background interference from the filter surface[36]. These advances in direct count methods revealed that the abundances of bacterial cells in seawater were orders of magnitude greater than previously estimated by colonies counts on agar plates. This discrepancy is now known as "the great plate count anomaly"[37].

Once abundance of bacteria in seawater became easily measurable, more advanced techniques of quantifying bacterial activity have since been developed. One such activity measure for the whole bacterial community is using radiolabeled leucine, a common amino acid in protein, or thymidine incorporation to quantify the rate of protein production[38,39]. Because proteins are a significant part and relatively constant proportion of bacterial cells, the rate of leucine incorporation into proteins can be directly correlated as bacterial biomass production[40]. A second method is florescent *in situ* hybridization

(FISH) that target groups via specific probe binding to rRNA in cells[41]. However, FISH does not give measures of activity. The combination of FISH with microradiography (MAR-FISH) detects specifically active cells by microradiography after leucine incorporation, and labeling with fluorescent probes[39]. All three methods, leucine/thymidine incorporation, FISH and MAR-FISH are useful in measuring activity. However, it is important to note that a bacterium may be metabolically active without incorporation of leucine/thymidine[42]. Additionally, the concentration of a substrate, such as leucine, has an impact on uptake by bacteria[43]. Another drawback to these methods is that all three rely on incubations as well as have limited sensitivity and therefore may not provide detailed information about the natural communities[41].

Another, more recent, method that is being used is to examine the rRNA:rDNA ratios from individual bacterial taxa[8,44]. Because 16S rDNA gene sequence similarity is one of the criteria used to define taxonomic groups[45], and the amount of ribosomal RNA is positively correlated with growth rates of many taxa[46–48] the 16S rRNA:rDNA ratio can provide estimates of growth rates for specific taxa. In general, there is a positive relationship between rRNA to rDNA ratio among bacterial communities[49,50]. This uncoupling of 16S rRNA and rDNA in some marine taxa reflect differences in abiotic factors such as: light, nutrient concentration, as well as other environmental parameters[8]. However, it is important to note that growth rate is not always linearly correlated to concentration of rRNA[51]. For example, under balanced growth conditions *Synechococcus* and *Prochlorococcus* strains show a three-phase relationship between growth rate and rRNA concentration: 1. during low growth rates, rRNA concentration remained constant, 2. during intermediate growth rates, rRNA concentration increased proportionally to growth rate, and 3. during higher growth rates, rRNA content seemed to decline as growth rate increased[52].

Meta-omics of prokaryotes

Recent advances in sequencing technology offer alternative tools to study microorganisms without the need for culturing[53,54]. In the last decade, DNA throughput sequencing has provided an unprecedented opportunity to obtain sequences from thousands of genomes at a time from many natural environments[53–55]. Recent studies using metagenomics data have extensively explored microbial community dynamics, including in estuaries[56–58]. Metagenomics data has provided an insight into the taxonomic and functional diversity of not only bacteria but also viruses, archaea, and other microbes[55–57,59]. Similarly, transcripts from microbes found in various environments have been sequenced (metatranscriptomics) to explore physiologically active members of the community[60–62]. The combination of both methods provides information on abundance, activity and also how this activity may differ in various environmental conditions[19,56,63]. Vast amounts of sequencing data allows researchers to infer the global distribution of phylogenetic lineages and metabolic potentials[53,54,64]; however, many bioinformatics analyses, at the individual taxa level, require meta-data to first be assembled into metagenome-assembled genomes (MAGs). MAGs are obtained by grouping assembled contiguous sequences (contigs) with similar sequence composition, coverage depth across one or more related samples, and taxonomic affiliation[65]. These assembled genomes are typically incomplete, and may contain contigs comprising multiple strains and/or species because of the challenges faced in differentiating between closely related members during both assembly and binning[66,67]. This binning approach has successfully been applied to a range of environments, including aquatic habitats[64,68–71]. One application for MAGs is to analyze activity by measuring differential gene expression (RNA-Seq) from the populations present in varied environments[40]. By mapping transcripts back to a reference genome, or MAG, the level of transcripts in the sample can be compared[72]. One of the main goals of such experiments is to identify the differentially expressed genes in multiple different conditions[73], including conditions that induce stress[74,75]. A big advantage of using RNA-Seq over other technologies, such as microarrays, is the ability to do transcriptome-wide analysis on non-model organisms since a reference genome is not required, and de novo assembly directly onto metatranscriptome reads is possible[76].

CHAPTER TWO

HYPOTHESIS AND OBJECTIVES OF STUDY

The objectives of this study are to determine and compare the functional potential of three metagenome assembled genomes (MAGs) within the *Roseobacter* clade and analyze how the activity and function of these MAGs changes by observing gene expression patterns in respect to temporal and spatial changes along the Delaware Bay using both metagenomic and metatranscriptomic methods. I hypothesize that the abundance of the three MAGs will be directly correlated to metabolic activity and, since the MAGs likely have a diverse range of physiologies, they will show signs of activity in all or most conditions but with significant differential gene expression in each. To compare the functional potential of the MAGs, KEGG categories were assigned to each protein encoding gene (peg). For one of the MAGs, MAG 22, the functional potential was compared to its closest relative, *Planktomarina temperata* RCA23[77,78]. The relative activity of MAG 22 was characterized for the diverse environmental conditions by using differential gene expression, specifically characterizing the number of transcripts related to growth and activity found in each condition. Changes were observed in activity between seasons, time of day and salinity.

CHAPTER THREE

EXPERIMENTAL METHODS

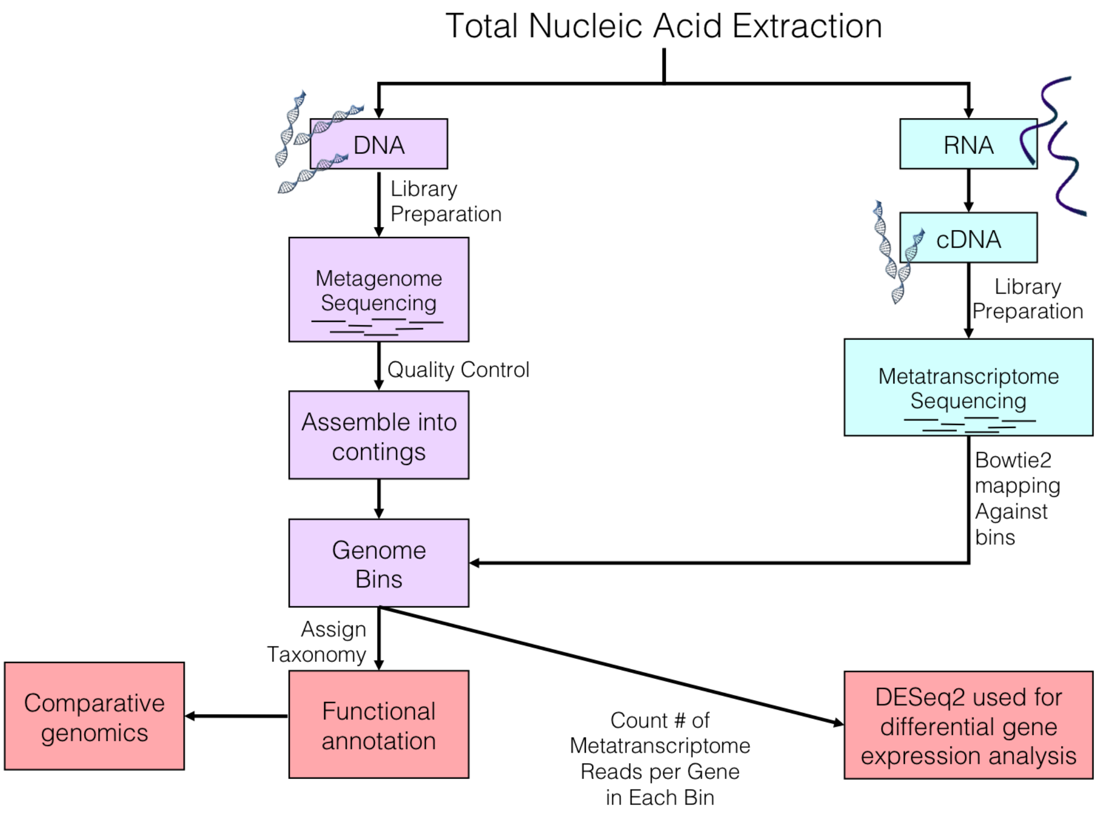
Total Nucleic Acid Extraction
DNA
Library Preparation
Metagenome Sequencing
Quality Control
Assemble into contings
Genome Bins
Assign Taxonomy
Functional annotation
Comparative genomics
RNA
cDNA
Library Preparation
Metatranscriptome Sequencing
Bowtie2 mapping Against bins
Count # of Metatranscriptome Reads per Gene in Each Bin
DESeq2 used for differential gene expression analysis

Figure 1. Flowchart demonstrating the general experimental design and setup to analyze MAGs of interest.

Work completed by others in Campbell Lab

Sample collection and sequencing

Surface water samples (~1-2 meters below surface) were collected spanning the estuarine gradient of the Delaware Bay during cruises in March, August and November, 2014 (**Table 1**). The Delaware Bay was sampled daily at four time points during the day (7:00 AM, 11:00AM, 7:00 PM and 11:00 PM). Standard oceanic properties were measured, including: water temperature, Secchi depth, salinity, light attenuation, Chlorophyll a concentration, bacterial production (Leucine incorporation) and nutrient (NO_3, NH_4, PO_4, Si) concentration as described previously[9,79]. Samples were either collected directly on 0.2 µm Durapore disk filters or first filtered through a 0.8 µm filter before collection on 0.2 µm filters in order to analyze larger cells or particle attached and free-living cells separately. All filters were frozen at -80 ^{o}C in 1 ml of RLT buffer until extraction. After both DNA and RNA were extracted, RNA was cleaned of DNA and both sent to the Joint Genome Institute for metagenomic and metatranscriptomic sequencing.

Sample preparation for sequencing, sequencing, quality trimming of sequences, and binning into MAGs was also completed by others in Campbell lab. Once the MAGs were made, their quality was assessed using CheckM. Percent completeness was estimated as the number of marker sets present in the genome, and genome contamination was measured by the number of multi-copy genes per marker gene[80]. Strain heterogeneity (SH) measured contamination, if present, that indicates whether the contamination was coming from either similar strains (SH is 100); or from more distantly related species (SH is 0)[80].

Table 1. General features of samples sequenced for metagenomic and metatranscriptomic analyses.

Month	Season	PSU[1]	Fraction	Time
March	Spring	20[2]	0.2	7:00 AM
		20[2]	0.8	
		30	0.2	
		30[2]	0.8	
August	Summer	22[2]	0.2	11:00 AM
		22[2]	0.2	11:00 PM
		22[2]	0.8	11:00 AM
		29[2]	0.2	11:00 AM
		29[2]	0.2	11:00 PM
November	Fall	30	0.2	11:00 AM
		30[2]	0.8	

[1] practical salinity unit
[2] indicates two samples are present

<u>Work completed by me</u>

<u>MAG annotations</u>

Genome sequences from three MAGs, along with their two closest relatives, were uploaded onto the Rapid Annotations using Subsystems Technology (RAST) server for annotation. RAST is an automated online annotation server that was used for annotating prokaryotic genomes[81–83]. It is built upon the framework provided by the SEED system to allow for high quality gene calling and functional annotations[81,82]. Steps utilized by the RAST annotation pipeline to annotate prokaryotic genomes are described in detail elsewhere[82]. The RAST annotations were used for most of the downstream analyses. In addition, using the BlastKOALA (KEGG (Kyoto Encyclopedia of Genes and Genomes) Orthology and Links Annotation)[84] server, I assigned K0 numbers (described below) to the MAGs. These K0 numbers were used to categorize genes expression that may be either up or downregulated in pathways during the different environmental conditions.

Phylogenomic tree construction & Two-way amino acid identity

A maximum likelihood phylogenomic tree was constructed using the concatenated alignment of 21 single-copy marker genes assigned by AMPHORA[85]. Phylogenomic tree analysis of these alignments was performed using scripts available at phylogenomics-tools (doi:10.5281/zenodo.46122)[86]. The 21 single-copy genes are as follows: *dnaG, infC, nusA, rplA, rplB, rplC, rplD, rplE, rplF, rplL, rplP, rplS, rpmA, rpoB, rpsB, rpsE, rpsJ, rpsK, rpsM, rpsS, tsf*. The tree includes 20 members from the Roseobacter clade, three MAGs and *Candidatus* Pelagibacter ubique HTCC1062 acting as the out-group.

The two-way amino acid identity (AAI)[87] scores were calculated based on the RAST "sequence based comparison tool"[81,82]. The output file contains information for each gene, marking it either unique, a unidirectional best hit or a bidirectional best hit in comparison to the reference genome[82]. Genes marked unique, compared to the reference organism were excluded from AAI calculation. For each comparison both genomes were made the reference one after another and the average of the two scores were used as the final AAI, hence the term two-way AAI.

Comparison of functional potential and transcripts

All three MAGs were submitted to BlastKOALA[84] server for assignment of K0[88] (KEGG orthology) entries. These entries characterize individual gene functions and reconstruct KEGG pathways, within BRITE hierarchies (http://www.genome.jp/kegg/kegg3b.html) and KEGG modules, to infer high-level functions of organisms[84]. The number of genes per category was counted using a bash-script (**Appendix**), prepared by Jean Lim and Jason Gholamian, and normalized per genome length before making a functional potential heat map using the heatmaply package[89], available for R (www.r-project.org).

Additionally, because *P. temperata* is the closest relative to MAG 22, the genome annotation for *P. temperata* was downloaded from the KEGG genome database to compare functional potential between the two organisms. The genome sequence of P. *temperata* was also uploaded to RAST for annotation and sequence and function based comparison was performed. Comparison of MAG 22 to *P. temperata* may also provide information about potentially missed genes during the binning process of MAG 22.

Sequence mapping to MAGs and generating count tables

Sequences from twenty metatranscriptomes and fourteen metagenomes (**Table 1**) from the Delaware Bay were mapped back to all three MAGs within the *Roseobacter* clade using Bowtie2[90]. The alignment output files created by Bowtie2 were in SAM file format[90], they were converted to compressed and sorted BAM file format using SAMTools[91] and were used for all other downstream analyses, except iRep. Two MAGs, MAG 73 and MAG 147, were originally assembled from summer samples, and one, MAG 22, was assembled from a spring sample. MAG 73 is from a summer high salinity (29 PSU) sample, collected at 11:00 am and greater than 0.8 µm size fraction. MAG 147 is from a summer high salinity sample (29 PSU), collected at 11:00 pm and less than 0.8 µm size fraction. MAG 22 is from a spring medium salinity (20 PSU), collected at 7:00 am and less than 0.8 µm size fraction.

The metagenome sequences were mapped to the MAGs using Bowtie2[90] to characterize the relative abundance of each of the MAG in the different environmental conditions that the samples were collected from. The metatranscriptome sequences were mapped to the MAG 22 using Bowtie2[90] in order to do differential gene expression analysis. Metagenome mapping allowed the characterization of potential growth rates of the MAGs by calculating an index of replication (iRep)[92]. Briefly, iRep is an algorithm that is used to estimate the population replication rate using draft-quality genome sequencing and single time-point metagenome sequencing[92]. iRep values were calculated based on the sequencing coverage depth, resulting from bi-directional replication from a single origin of replication[92]. It is important to note that iRep values are an average measure across a population of cells and in fact, some organisms may not be replicating at all where others are replicating quickly.

FeatureCounts[93] was used to generate a feature count table containing raw read counts of each gene per sample and library[93]. The program requires BAM files, created using Bowtie2/SAMTools, as well as a GFF or GTF file that was generated during RAST automated genome annotation. The GFF/GTF file was downloaded from the RAST annotation server.

RNA-Seq analyses

TPM (transcripts per million) counts were generated based on the count table produced using FeatureCounts[93]. The number of transcripts present per gene in each sample that mapped back to MAG 22 was normalized by both the gene size as well as per million base pairs. All of the genes were then assigned into KEGG categories based on K0 number. The transcripts from duplicate samples were averaged. The normalized and averaged values were log transformed before plotting on to a heat map using the heatmaply package[89] in R (www.r-project.org) to visualize the abundance of transcripts present by KEGG categories.

The R package from Bioconductor, DEseq2[94], was used to perform gene-level deferential expression with statistical analysis for MAG 22 using twenty metatranscriptomes (**Table 1**). DEseq2[94] uses non-normalized library counts, such as those generated by FeatureCounts[93], to perform gene-level differential expression analysis[94]. An increase and decrease in gene expression level was calculated as log2 fold ratios between environmental conditions. A log2 fold ratios example would be where a change from 30 to 60 would be written as 60/30, or simply a log2 fold change of 2 (2-fold increase); similarly if the value changed from 60 to 30 it would be defined as 30/60 or simply 0.5 (2-fold decrease)[95]. If the gene was differentially expressed, it was determined to be either significantly (based on p-value) increased or decreased using a negative binomial frequency distribution model present in the DEseq2 package[95], within the R environment. The tutorial for the DEseq2 package was modified as needed to analyze the MAG feature count table.

(https://bioconductor.org/packages/3.7/bioc/vignettes/DESeq2/inst/doc/DESeq2.html)

CHAPTER FOUR

EXPERIMENTAL RESULTS

General Description/characteristics of MAGs

MAGs were chosen based on a high percentage of completeness (>90%), low contamination (<1%), and phylogenetic relatedness to the Roseobacter clade. To better understand the functional potential of these three MAGs of interest, I performed comparative analysis with the MAGs and their two closest relatives, *P. temperata*[77,78] and HIMB11[96], in the subsequent sections below. The general characteristics of the three MAGs of interest and their two closest relatives are listed in **Table 2**. The three MAGs have approximately 2500 protein encoding sequences, or genes (**Table 2)**. In contrast, their closest known relatives have approximately 3200 protein encoding sequences[77,78,96], indicating the MAGs are likely missing approximately 700 genes/functions and are hence not 100% complete. Even though they are incomplete, the MAGs have value in that they contain novel genes that are not found in their closest relatives. For MAG 22 there are 23 genes that are unique and not found in *P. temperata*. MAG 147 has 17 genes that are unique and MAG 73 has 63 unique genes compared to HIMB11, the closest relative of both MAGs (**Tables S1a, S1b, S1c**). The G+C content of MAGs ranges from 40.6% (MAG 73) to 55.1% (MAG 22). The percent of each genome that is predicted as coding ranges from 87.01% (MAG 147) to 90.66% (MAG 22) (**Table 2)**.

Table 2. General genomic features of the three MAGs of interest and their two closest relatives (MAG 22 closest relative is *Planktomarina* and MAG 73 & 147 share the closest relative HIMB11) NA – indicates not applicable

MAG/Closest Relative	Genome Size	Coding Sequence Length (bp)	% DNA Coding Region	GC Content	# of Subsystems	# of Coding Sequences	# of Contigs	L50	N50	Avg. Coverage	% Complete	Contami-nation	Strain Heter-geneity
Planktomarina	3,288,122	2,929,177	89.08	53.6	404	3204	1	1	NA	NA	NA	NA	NA
HIMB11	3,098,747	2,774,072	89.52	49.7	423	3220	34	5	282310	NA	NA	NA	NA
MAG 22	2,524,490	2,288,716	90.66	55.1	389	2522	117	25	33103	4.11	98.50	0.06	50
MAG 73	2,338,765	2,076,694	88.79	40.6	347	2319	133	27	32774	1.14	91.05	0.15	0
MAG 147	2,418,783	2,104,687	87.01	50.5	371	2546	175	36	17093	3.57	93.26	0.30	100

Phylogenomics and potential abundance of all three MAGs

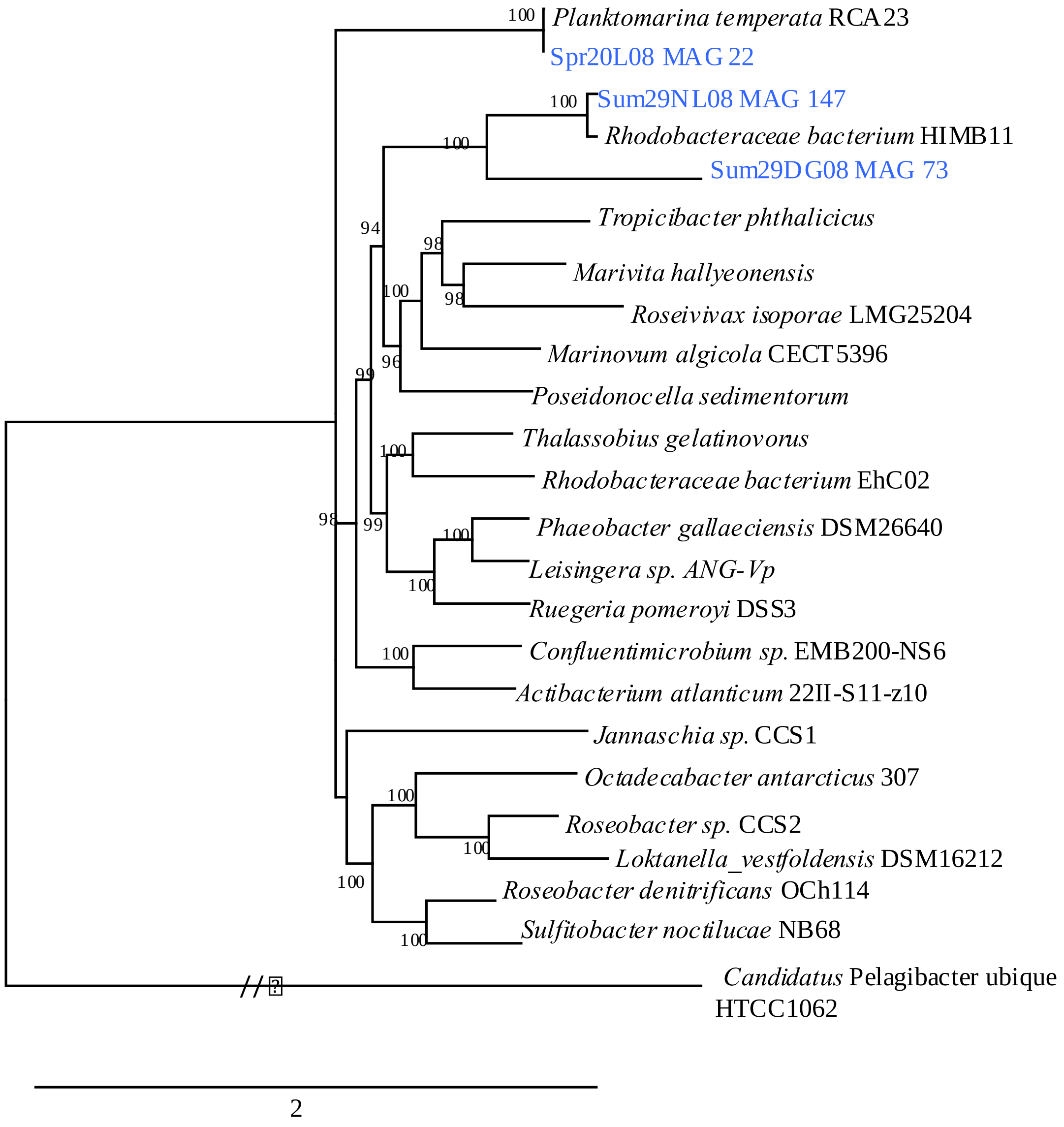
100
Planktomarina temperata RCA23
Spr20L08 MAG 22
100
Sum29NL08 MAG 147
Rhodobacteraceae bacterium HIMB11
100
Sum29DG08 MAG 73
94
Tropicibacter phthalicicus
98
Marivita hallyeonensis
100
98
Roseivivax isoporae LMG25204
Marinovum algicola CECT5396
96
99
Poseidonocella sedimentorum
Thalassobius gelatinovorus
100
Rhodobacteraceae bacterium EhC02
98
99
Phaeobacter gallaeciensis DSM26640
100
Leisingera sp. ANG-Vp
100
Ruegeria pomeroyi DSS3
100
Confluentimicrobium sp. EMB200-NS6
Actibacterium atlanticum 22II-S11-z10
Jannaschia sp. CCS1
Octadecabacter antarcticus 307
100
Roseobacter sp. CCS2
100
Loktanella_vestfoldensis DSM16212
100
Roseobacter denitrificans OCh114
Sulfitobacter noctilucae NB68
100
Candidatus Pelagibacter ubique HTCC1062
2

Figure 2. Phylogenomic tree showing MAGs analyzed in this study (indicated in blue) in relation to other roseobacters. The tree was inferred based on the concatenated alignment of the following 21 single-copy marker genes: *dnaG, infC, nusA, rplA, rplB, rplC, rplD, rplE, rplF, rplL, rplP, rplS, rpmA, rpoB, rpsB, rpsE, rpsJ, rpsK, rpsM, rpsS, tsf* assigned by AMPHORA2[85]. Approximate likelihood-ratio test (aLRT) SH-like support values[97] greater than 70% are shown beside each node. *Candidatus* Pelagibacter ubique HTCC1062 is acting as the out-group. Scale bar indicates nucleotide substitutions per site.

The phylogenomic tree based on 21 genes indicates that all three MAGs group into the Rosobacter clade (**Figure 2)**. The two-way AAI score between *P. temperata* and MAG 22 was 95.7%; between HIMB11 and MAG 147 the score was 90.6%; between HIMB11 and MAG 73 it was 58.6% (**Figure 3**). The results from both the phylogenomic tree and two-way AAI (**Figure 3**) were in agreement with one another. *P. temperata RCA23* (herein referenced as *P. temperata*) was found to be the closest known relative to MAG 22 as indicated phylogenomically and by AAI (**Figure 2, 3**). *Rhodobacteraceae bacterium HIMB11* (herein referenced as HIMB11) was found to be the closest known relative of both MAG 73 and MAG 147, where MAG 147 was more closely related to HIMB11 than MAG 73.

Two-way AAI

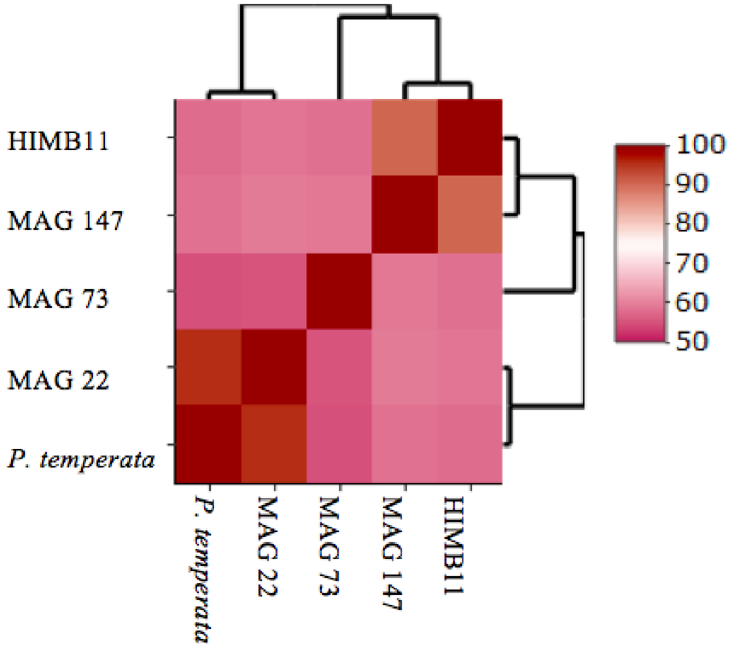

Figure 3. Heat map of two-way amino acid identity (AAI) calculated between all three MAGs assembled for this study, and their two closest relatives. The heatmap is calculated based on all shared proteins across the genomes, and was generated using the heatmaply package[89]in R (www.r-project.org).

Index of replication

Metagenomic samples were mapped to each of the three MAGs and the average coverage and iRep values were graphed (**Figure 4)**. MAG 22 has the highest coverage in spring, indicating it is most abundant during this time, whereas, MAGs 147 and 73 have the highest abundance in the summer (Figure 4A). An iRep value was assigned for samples with ≥5X coverage (**Figure 4B**). An iRep value of 1.5, for example, would indicate that approximately half of the cells are undergoing replication[92]. For the samples that were assigned an iRep value, roughly three quarters of the cells for MAG 22 (iRep average of 1.91 with a standard deviation of 0.08), three quarters for MAG 147 (iRep average of 1.87 with a standard deviation of 0.08), and all or most of the cell for MAG 73 (iRep average of 2.48 with a standard deviation of 0.13), were undergoing replication at the time of sampling. However, conclusions about the activity, specifically growth rates, cannot be made since iRep values were not assigned to the samples with low coverage (**Figure 4B**).

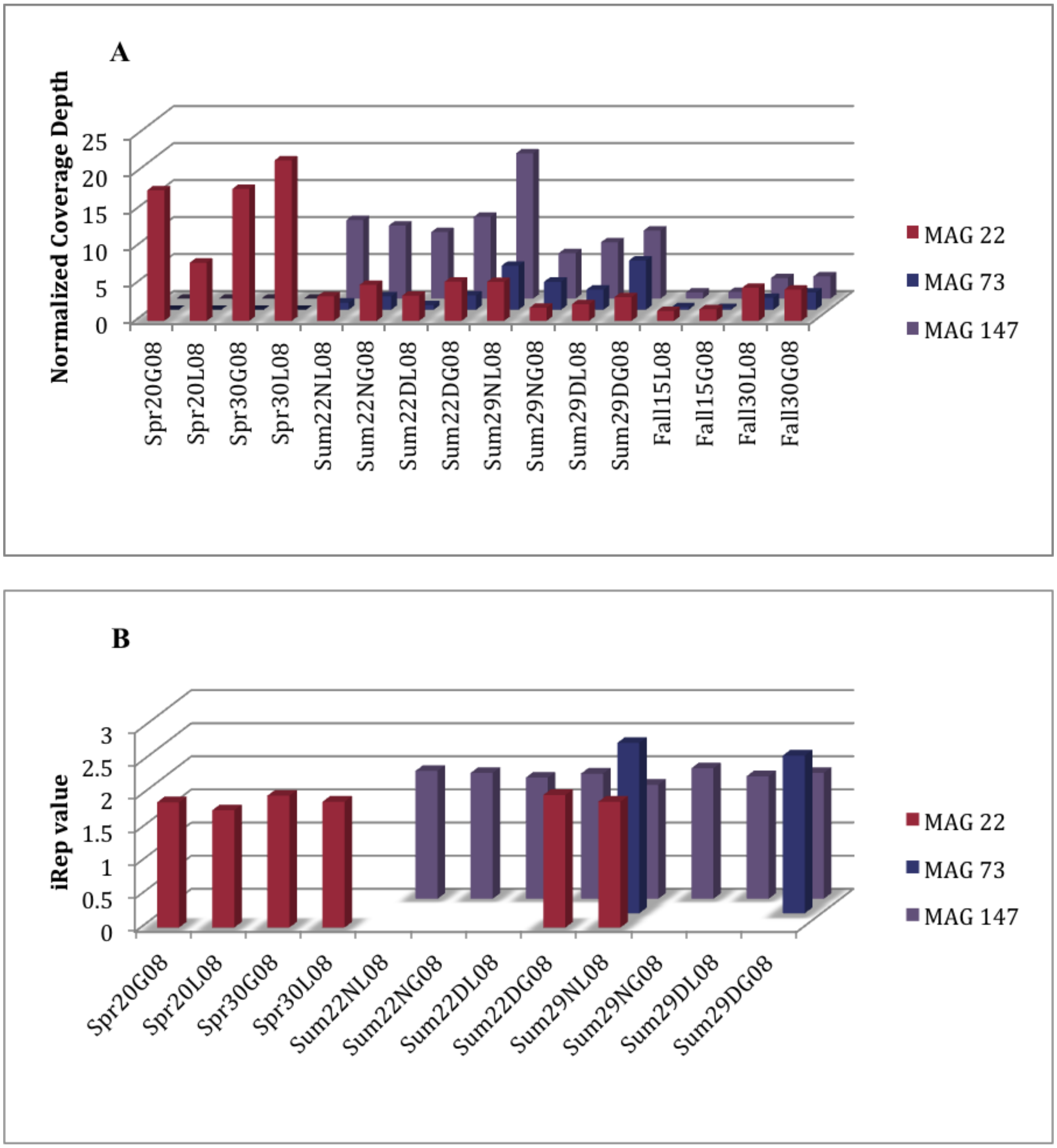

Figure 4. Coverage depths of the indicated MAG normalized to 50 million base pairs (50Mb) of the indicated metagenome sample (**A**) and iRep values from the subset of metagenome samples with coverage ≥5X (**B**). (Spr = spring, sum = summer; D= day, N=night; # = salinity in PSU (**Table 2**); L08 = size fraction of 0.2 µm, G08 – size fraction of 08 µm).

Comparative Genomics - General

Pathways/Functional potential comparisons between genomes

To predict the functional potential of the three MAGs in relation to biogeochemical activity, K0 numbers were assigned to each peg within each MAG. The pegs were next grouped into KEGG categories according to their K0 number. The number of genes per KEGG category was plotted in a heatmap to visualize the genome differences between MAGs (**Figure 5**). For all three MAGs the number of genes present per KEGG category were similar, with all three genomes having the highest number of genes for membrane transport followed by amino acid metabolism, translation and carbohydrate metabolism (**Figure 5**). For this reason, detailed analysis of only MAG 22 was performed to characterize the specific metabolic properties that may be important to biogeochemical cycling. In addition, MAG 22 has the highest percent of completeness (**Table 2**) and coverage across all the seasons (**Figure 4A**), which makes comparisons across seasons possible.

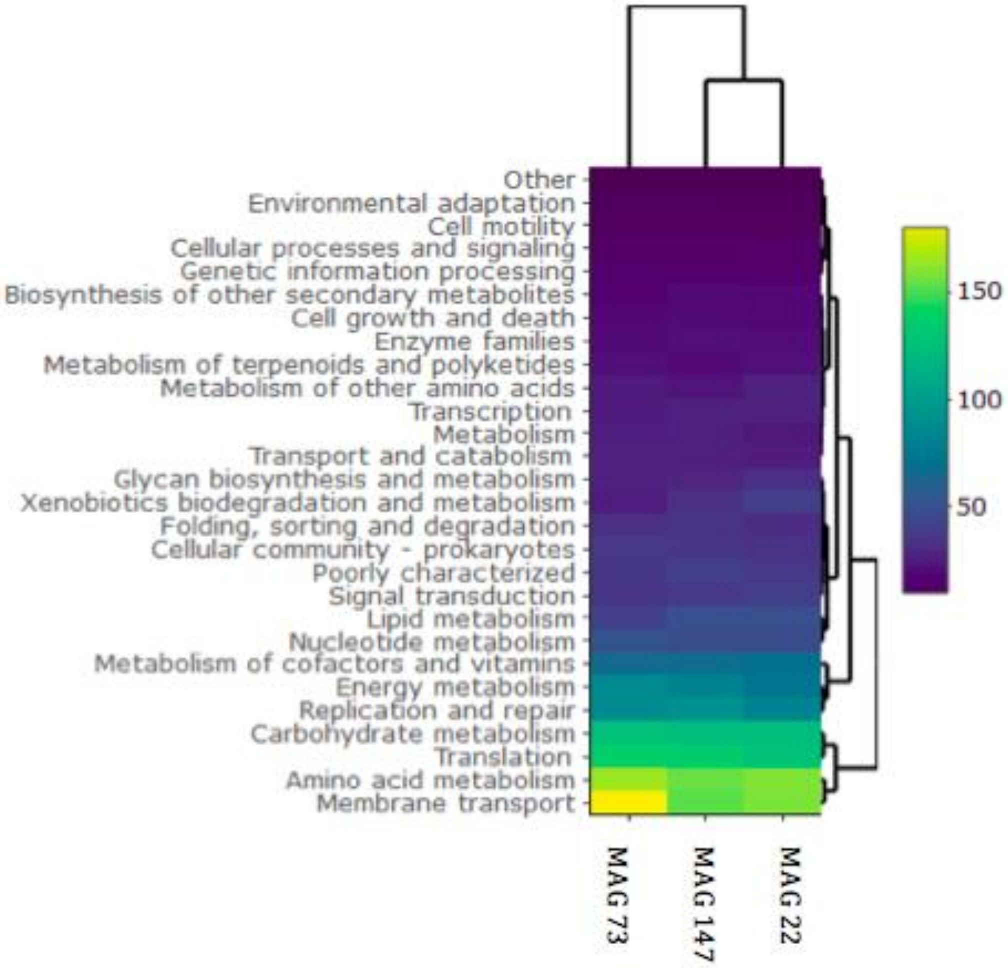

Figure 5. Heat map based on KEGG categories, showing the number of protein encoding genes present per MAG in each category, normalized to genome (MAG) size. Heatmap generated using the heatmaply package[89]in R (www.r-project.org).

Comparative Genomics of MAG 22

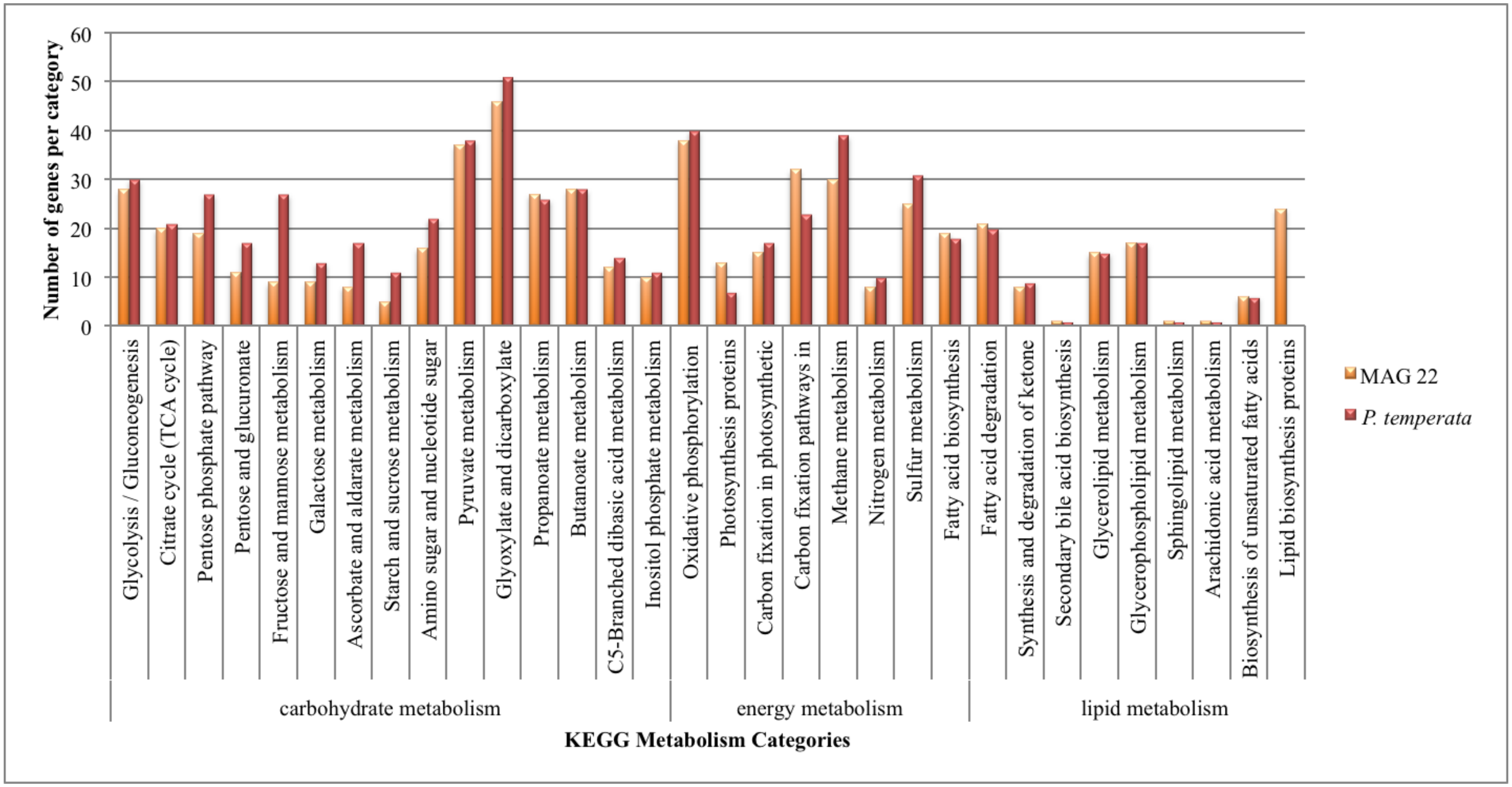

Figure 6. Metabolic potential for MAG 22. Bar graph showing the number of genes found in MAG 22 and *P. temperata* per KEGG biogeochemical cycling categories.

I. Carbon and energy acquisition

The main primary metabolic pathways predicted in MAG 22 include glycolysis, pentose phosphate pathway, the citrate cycle (TCA) and oxidative phosphorylation, among others (**Figure 6**). Additionally, MAG 22 contains most of the genes (6 out of 9) within the *cox* cluster that have been shown experimentally to mediate carbon monoxide oxidation at low concentrations, typical of ocean surface waters (≤5 nM)[31]. Many genes (46) for carbon fixation in prokaryotes and photosynthetic organisms are also found in MAG 22. Finally, based on RAST and SeedViewer analyses[81-83], MAG 22 has the required genes for a complete aerobic anoxygenic phototrophy (AAnP) pathway.

II. Nitrogen acquisition

MAG 22 is capable of assimilating amino acids (general, branched-chain and polar), polyamines, glycine-betaine, and other nitrogen-rich organic compounds, as evidenced by the presence of >250 genes for these processes. It was difficult to find the exact number as some of the genes found were listed as putative. There were no genes found for the assimilation of nitrite, nitrate or urea; however, there were 13 genes for urea degradation, indicating that urea assimilation genes may have been missed during the binning process.

III. Sulfur acquisition

There were a total of 25 genes associated with sulfur metabolism in the sequences of MAG 22. From these, no definitive conclusion can be made because only a few genes from each pathway are present indicating that they may have been missed during the binning process. There are genes associated with taurine assimilation (9), sulfite dehydrogenase (3), sulfite reductase (2), sulfur oxidizing proteins (*sox* cluster) (6), and the demethylation pathway of dimethylsulfoniopropionate (DMSP) catabolism (1).

IV. Transporters

Membrane transport proteins are particularly important for cells, as they are responsible for providing the cell with nutrients as well as discarding toxic molecules. There were a total of 366 genes for transporters found in MAG 22. Out of these, 26% are ABC transporters. The majority of the rest are for the transport of amino acids, sugars and other micronutrients.

Differential Gene Expression Analyses of MAG 22

Samples from diverse environmental conditions were examined for expression of MAG 22 pegs; these samples are from three different seasons (spring, summer, and fall), two salinities (medium and high), two size fractions (0.2 µm and 0.8 µm) and the summer samples are from two times (day and night) (**Table 1)**. To better understand and characterize the differences in transcript abundance in these environmental conditions, sequences from all twenty metatranscriptome samples were mapped to all 2,522 MAG 22 pegs.

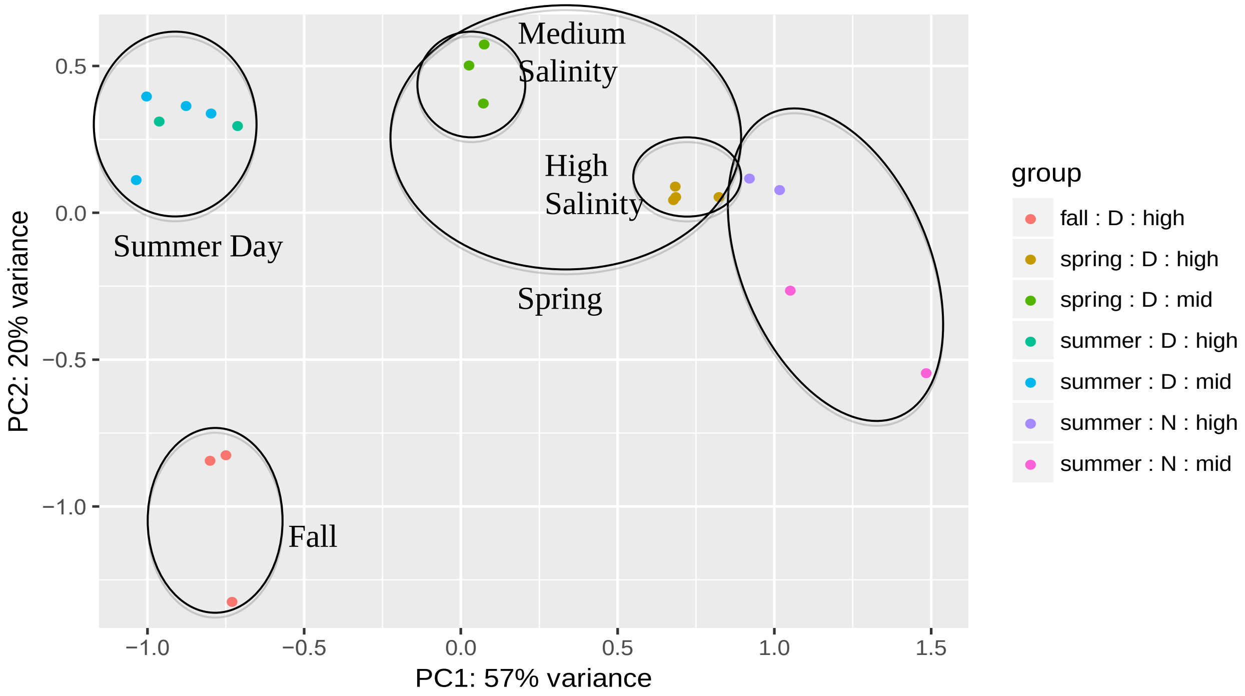

Figure 7. Principal Coordinates Analysis (PCA) plot showing how the metatranscriptome reads that mapped to MAG 22 genes cluster in relation to

environmental conditions. (D= day; N=night; high & mid (medium) refers to salinity in PSU (**Table 2**)). Circles are drawn for clarification.

Based on the PCA plot (**Figure 7**), there was a clear separation in gene expression between samples from different seasons (summer vs. fall vs. spring). Additionally, within the summer samples, there was a separation of gene expression patterns between the time of day that the sample was collected (day vs. night), but not by salinity. In contrast, gene expression in spring samples separated based on salinities where gene expression from medium salinity (20 PSU) samples and high salinity (30 PSU) samples were distinct. Based on these separations, in depth gene expression analysis was performed for the three distinct comparisons mentioned above.

The number of transcripts in MAG 22, calculated as TPM was analyzed (**Figure 8**). The heat map aids in visualizing the categories in which the highest number of transcripts were being expressed at the time of sampling. The highest number of expressed genes for MAG 22 were associated with translation, membrane transport, energy metabolism and signal transduction (**Figure 8**). A similar pattern was also observed with the two other MAGs 73 & 147 (data not shown).

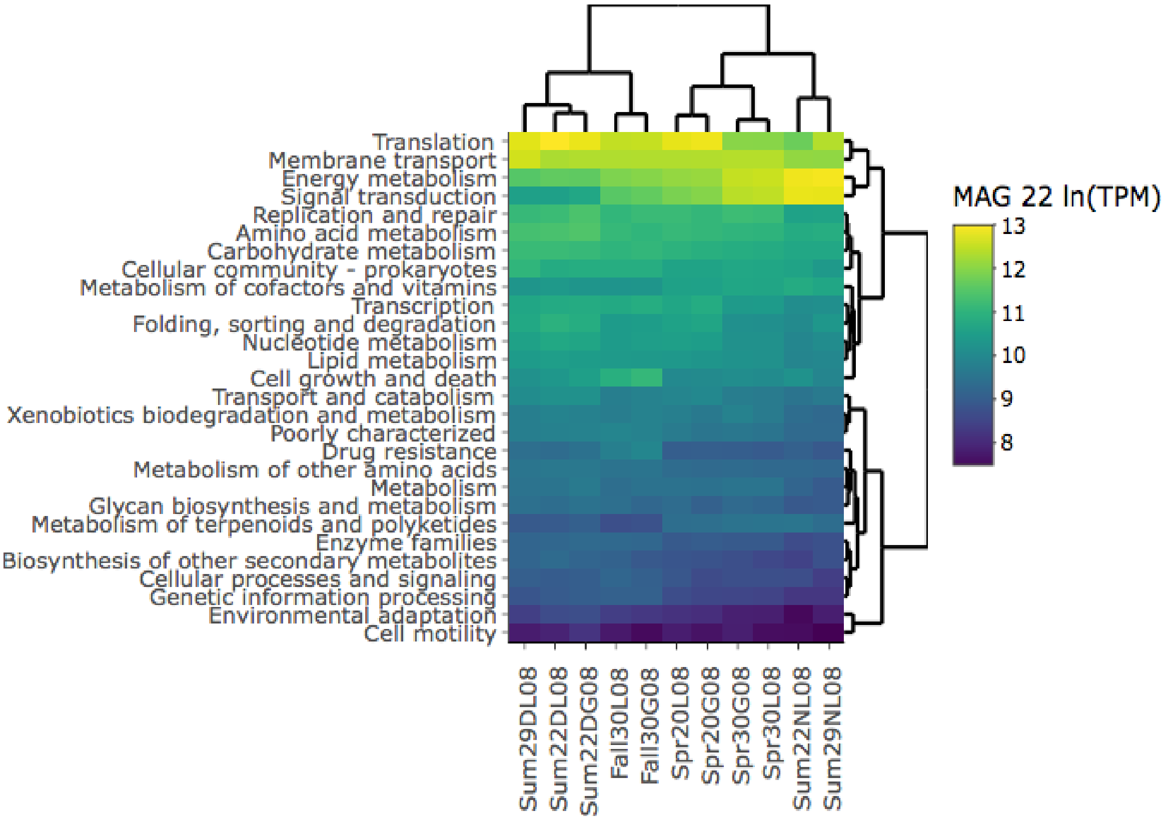

Figure 8. Heatmap representation of normalized TPM counts for MAG 22. TPM values were averaged between duplicate samples and log transformed prior to generation of the heat map. Sum = summer; spr = spring; # = salinity in PSU; G08= size fraction of 0.8 µm; L08 = size fraction of 0.2 µm. Heatmap generated using the heatmaply package[89]in R (www.r-project.org).

Gene expression analysis between seasons was performed with nine fall and summer samples collected at 11:00 am because both fall and summer samples were collected at the same time. Out of the top 20 most highly expressed genes in both seasons, 14 of them (70%) are membrane transporters (**Figure 9**). Two of 20 (10%) are related to transcription/translation of DNA and RNA. Another two of 20 are related to energy metabolism (sulfur oxidation and ATP synthesis). The remaining 10% consist of a hypothetical protein and a heat shock protein that is expressed higher in the summer compared to fall (**Figure 9**). Ninety percent of the 20 most highly expressed genes in both summer and fall are indicative of cell growth and metabolic activity.

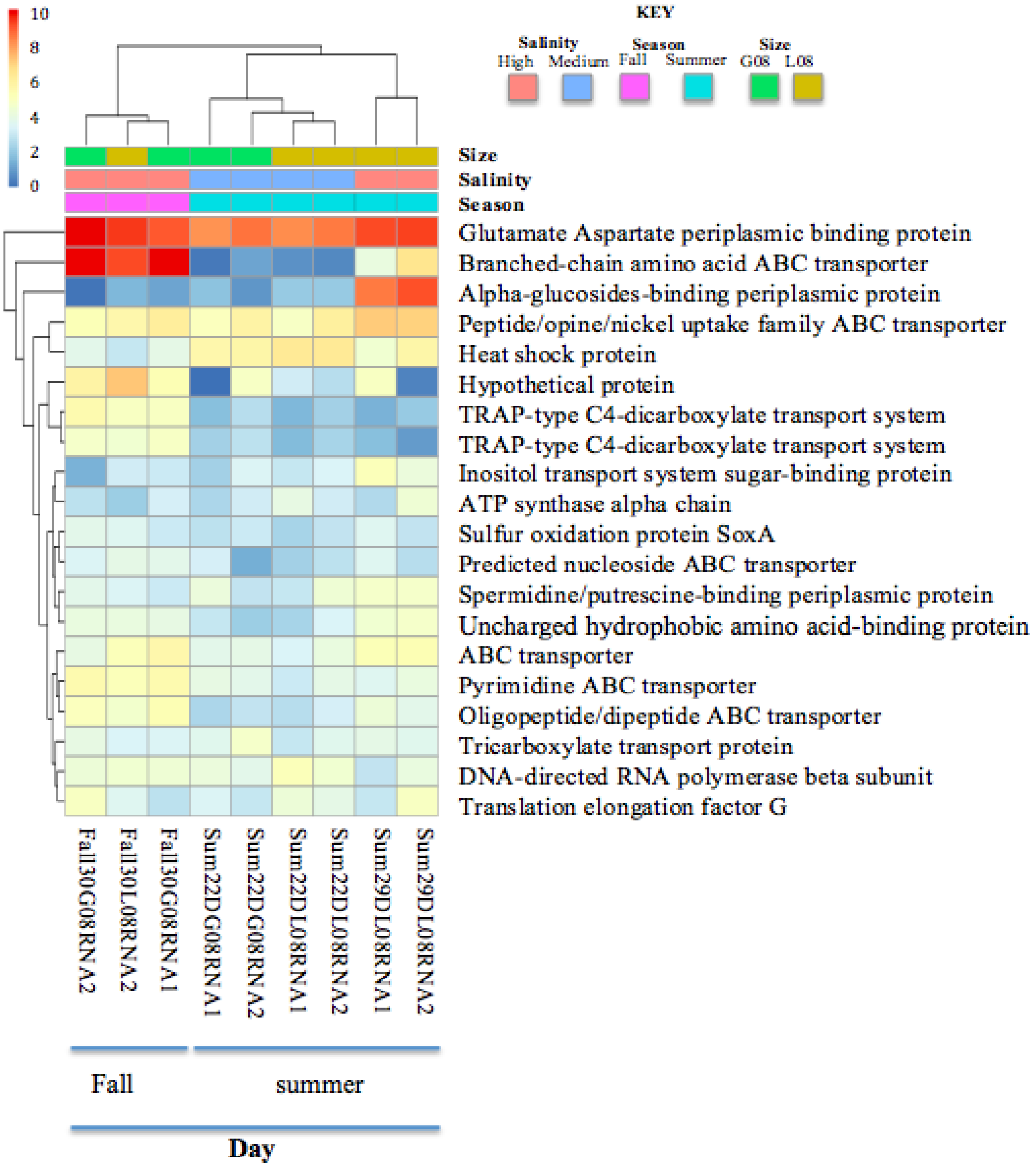

Figure 9. Heat map representation of the mean expression of MAG 22 transcripts in day samples collected from fall and summer seasons. The top 20 most expressed genes are shown. Sum = summer; # = salinity in PSU; G08= size fraction of 0.8 µm; L08 = size fraction of 0.2 µm.

While the heat map shows genes that were highly expressed, regardless of significance, **Table 3** lists genes that are significantly up or downregulated between the two seasons. In total, 61 genes were differentially expressed between fall and summer, of which 46 (75%) were upregulated during fall and 15 (25%) upregulated during summer. It is important to note that even though the differential expression was higher in fall, the overall number of transcripts was higher during the summer. Eighty-five percent of the non-hypothetical genes that were upregulated during summer are involved in transcription/translation, amino acid biosynthesis and cell wall recycling. Similarly, 93% of the non-hypothetical genes that were upregulated during fall are involved in carbohydrate and energy metabolism, membrane transport, DNA replication and protein folding.

Gene expression analyses of the 20 highest expressed MAG 22 transcripts during both day and night indicate that 4/20 (20%) of the transcripts are related to acquiring light energy, 12/20 (60%) are membrane transporters, and 2/20 (10%) are related to cell growth and activity (DNA replication & translation) (**Figure 10**). Out of the top twenty highest expressed transcripts only four are upregulated during the day (average log2 fold change of 0.42), from which two are related to growth and one is a heat shock protein.

Differential gene expression analysis between day and night samples was performed with the ten summer samples collected at 11:00 am and 11:00 pm from medium and high salinities (22 & 29 PSU). In total, there were 52 genes that were differentially expressed between the two times. Most of the genes (45/52) were upregulated during night compared to day (**Table 4).** Seventy-three percent of the non-hypothetical genes upregulated at night are related to acquiring light energy (carotenoid biosynthesis, chlorophyll biosynthesis and photosystem II synthesis). In contrast, the seven genes upregulated during the day do not fall into a broad category, and are scattered. Additionally, a high percentage (43%) of the upregulated genes during the day are hypothetical.

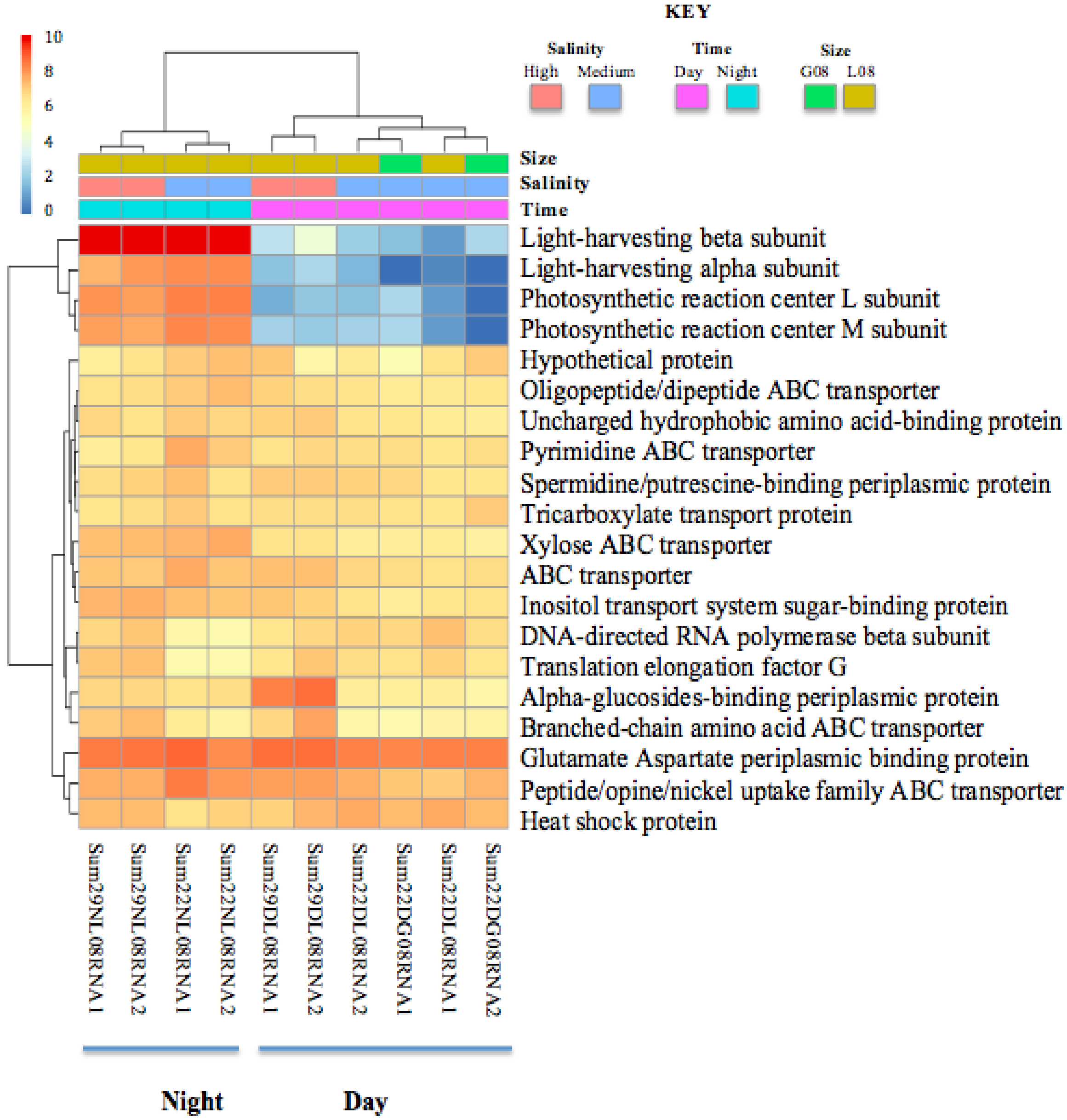

Figure 10. Heat map representation of mean expression of MAG 22 transcripts from day and night summer samples. The top 20 most expressed genes are shown. Sum = summer; # = salinity in PSU; N= night, D= day; G08= size fraction of 0.8 µm; L08 = size fraction of 0.2 µm.

Lastly, the medium and high salinities from spring (20 & 30 PSU) samples were analyzed for differential gene expression patterns. The heat map of the 20 highest expressed transcripts within medium and high salinity samples (**Figure 11**) shows that 13/20 (65%) of the highest expressed genes were membrane transporters. There were more transcripts related to transporters upregulated in high salinity samples compared to medium salinity samples (34 vs. 7). Interestingly, the type of transporters were similar in both conditions with the majority of them being ABC transporters and involved in either amino acid or sugar transport (**Table 5**).

Out of all three comparisons (summer vs. fall, summer day vs. summer night, spring medium vs. spring high salinity) salinity had the highest number of genes that were differentially expressed between the two conditions (235 genes) (**Table 5**). Fifty-three percent of the differentially expressed genes were upregulated in medium salinity (20 PSU), and 47% were upregulated in high salinity (30 PSU). However, while the other two comparisons had most of the genes that were differentially expressed with a log2 fold change value greater than 1.0, most of the genes differentially expressed between the two salinities had log2 fold change values of less than 1.0.

Out of the 53% of the MAG 22 transcripts upregulated in medium salinity (20 PSU) samples, most (94%) were indicative of activity and were categorized into carbon, energy, lipid, sulfur, amino acid, and nitrogen metabolisms. Additionally, in the medium salinity samples, a large number of upregulated genes (70 genes) were related to growth (transcription/translation, DNA replication, cell division, and ribosomal proteins). In contrast, MAG 22 transcripts upregulated in the higher salinity samples consisted of a large number of transporters and photosynthetic proteins (54 genes total).

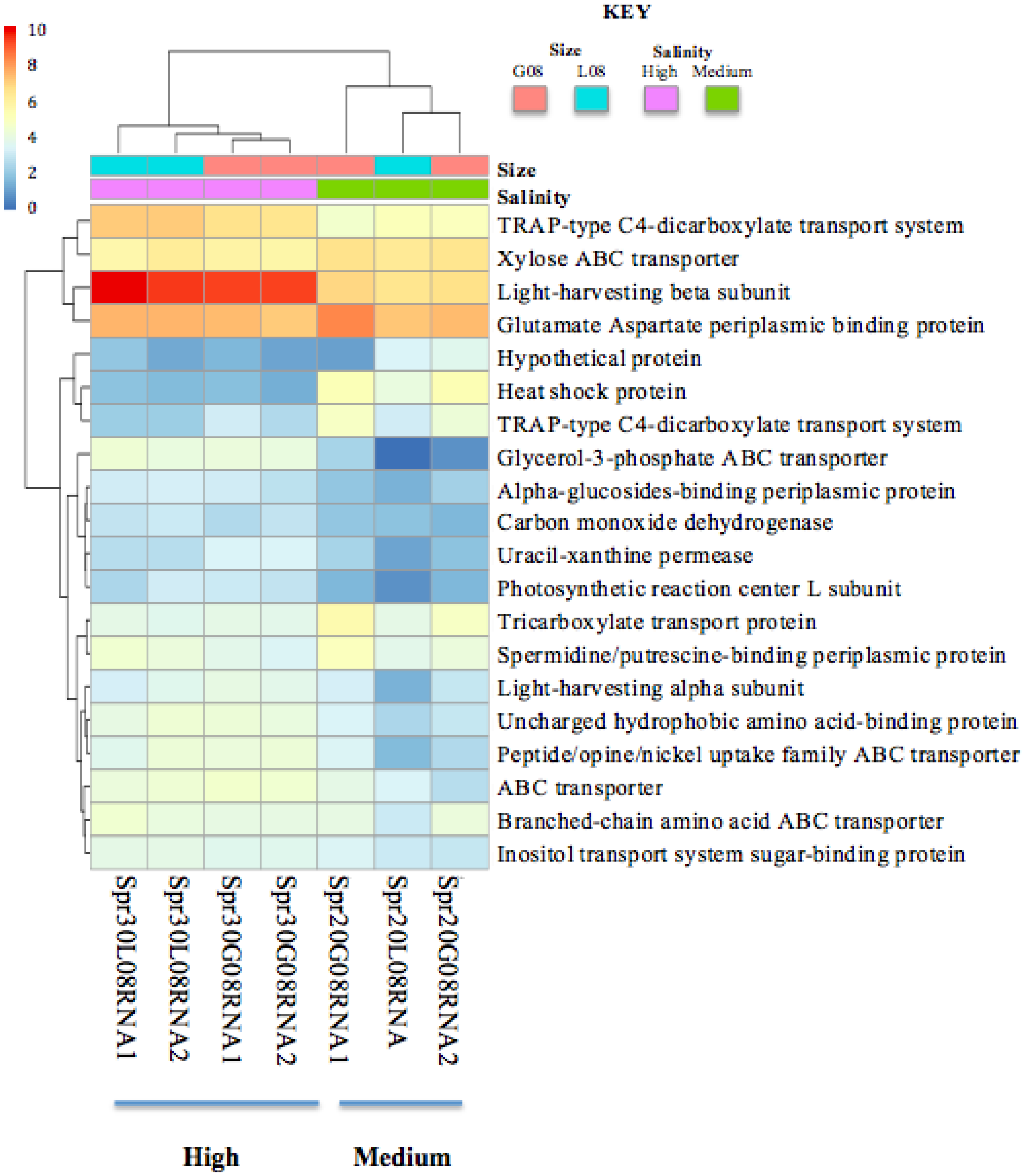

Figure 11. Heat map representation of mean expression of MAG 22 transcripts across medium and high salinity (20 & 30 PSU) in spring samples. The top 20 most expressed genes are shown. Spr = spring; # = salinity in PSU; G08= size fraction of 0.8 µm; L08 = size fraction of 0.2 µm.

CHAPTER FIVE

DISCUSSION

Bacteria dominate in abundance, diversity and potentially metabolic activity in many environments thus contributing significantly to biogeochemical cycling. My work elucidates the significance of species within the Roseobacter clade for biogeochemical cycling in the Delaware Bay, based on sequenced metagenomes that were assembled into MAGs and their corresponding metatranscriptome data. The three MAGs assembled here are phylogenetically associated with the Roseobacter clade, one of the most prominent groups present in coastal surface waters[24–26]. One of the reasons attributed to roseobacter success is the wide range of physiologies they are capable of utilizing[22]. The closest relative of MAG 22 was found to be *P. temperata.* As with *P. temperata*, MAG 22 is able to perform AAnP, CO oxidation, sulfur transformations, aromatic compound degradation and has implications for a diel cycle for turnover of organic matter. Therefore, it is most likely an important contributor to biogeochemical cycling of carbon and sulfur in the estuarine environment, especially during spring when it is most abundant. Similarly, the other two MAGs have similar physiological capabilities and likely contribute more to biogeochemical cycling during summer when their abundance is higher. All of the genes involved in the processes mentioned above were being transcribed at the time of sampling for MAG 22, but were not necessarily differentially expressed. Expression of these genes/pathways in all of the conditions indicates that they are an important part of the organism’s metabolism.

The abundance of all three MAGs was determined using iRep analysis. MAG 22 had the highest coverage in spring, where both MAG 73 and MAG 147 had higher coverage in the summer. All three MAGs seem to be least abundant during the fall. This difference in abundance between the seasons reflects the temperature preferred by their closest relatives, with *P. temperata* having optimal growth at 25 ^{o}C[77] and although the optimal temperature is not known for HIMB11, its range is higher than *P. temperata*[96]. The closest relatives of all three MAGs are found in coastal oceans[77,96] and estuaries[98]. Many studies have also associated roseobacter abundance with phytoplankton blooms and increased nutrients[25,99].

The average size of the Roseobacter clade analyzed this far is around 4.4Mb[22]. Interestingly though, all three of the MAGs have genome sizes significantly smaller than the average even with a high percentage of completeness. Out of the three nearly complete MAGs at least one represents a potentially novel species within the Roseobacter clade, MAG 73. MAG 73 has only slightly above 55% two-way AAI with its closest relative, HIMB11. Furthermore, MAG 73 has a total of 63 unique genes that are not found in HIMB11 related to carbohydrate metabolism, vitamins and cofactor metabolism, membrane transporters and cell signaling, potentially indicating the types of physiologies this organism is able to utilize. MAG 22 was the only genome analyzed in detail. The closest relative of MAG 22 is *P. temperata,* with a two-way AAI of >95%, indicating it is likely within the same genus and possibly the same species[87]. RAST and KEGG comparisons of MAG 22 to *P. temperata* indicated that there were several genes present in MAG 22 that were not found in its closest relative. These genes included several lipid biosynthesis proteins, based on KEGG category analysis, and genes involved in carbohydrate metabolism, cofactors/vitamins, stress response and RNA metabolism, based on RAST subsystem analysis.

The growth rates of the all three MAGs around the time of sampling were estimated using iRep. iRep values are only assigned when the coverage depths is greater than five times[92]. Since the coverage was too low for many samples, accurate conclusions about activity cannot be drawn based on this analysis. However, it does provide insight into potential growth rates of MAGs for which iRep values were assigned. Roughly three quarters of the cells for MAG 22 and MAG 147, and all or most of the cell for MAG 73 were likely undergoing replication at the time of sampling.

Gene expression analyses were performed only for MAG 22. The highest number of transcripts, in all conditions, was attributed to light harvesting proteins, transporters, ribosomal proteins and a few stress related genes. Additionally, within the top 50 genes being expressed there were a few genes involved in sulfur metabolism, mostly sulfur oxidation, indicating that this organism may play an important role in the cycling of sulfur, and that sulfur oxidation is an important metabolism in this organism.

Many significantly differentially expressed genes were observed between season, time and salinity but not between size fractions. A study that analyzed the transcriptomic data for *P. temperata* suggests that this organism undergoes intense metabolic reconstruction during night[78], and a significantly higher number of normalized transcriptomic reads were mapping back to categories such as protein synthesis and stress response with enhanced flagella protein synthesis during the night[78]. For MAG 22, the greatest number of normalized transcript reads that were differentially expressed and upregulated at night were related to light harvesting protein synthesis. This observation is in agreement with previous studies[78,100,101], because light has a negative effect on pigment formation in AAnP bacteria[100]. Unlike *P. temperata*, MAG 22 did not have any significant differential expression of genes associated with stress response or flagellar motility at night, despite the genes being transcribed under both conditions. The metabolic reconstruction that occurs at night is proposed to be a mode of energy conservation amongst many aerobic anoxygenic phototrophy (AAnP) bacteria, including *P. temperata* and MAG 22.

Differences in gene expression observed between summer and fall were related to growth and energy metabolisms with 75% of all of the differentially expressed genes upregulated in fall compared to summer. Despite these similarities, the type of growth and activity observed between the seasons was different. The genes upregulated during fall are involved in carbohydrate and other energy metabolism, membrane transport, DNA replication and protein folding; whereas, in the summer genes upregulated are related to transcription/translation, amino acid biosynthesis and cell wall recycling. Interestingly, more genes related to membrane transporters were expressed in fall compared to summer, potentially indicating that there are less nutrients available and therefore more transporters are needed during fall to take up available nutrients than during summer.

Differential gene expression analysis between the two different salinities (20 & 30 PSU) during spring had the maximum number of genes that were differentially expressed. However, little to no differential expression was observed between the different salinities during summer. Interestingly, the other two comparisons (summer vs. fall and day vs. night) had most of the genes being differentially expressed with a log2 fold change value greater than 1.0, most of the genes differentially expressed between the two salinities had log2 fold change values of less than 1.0. This indicates that even though the number of genes either up or downregulated between the two salinities within spring was the largest, the expression level does not drastically change. Despite this, there were 62 genes upregulated in the medium salinity (20 PSU) related to transcription/translation and ribosomal proteins all of which are indicative of growth and activity. The iRep numbers for the two different salinities in spring do not reflect the results observed using transcriptome data; potentially indicating that the iRep analysis is not sensitive enough to detect these types of changes and that transcriptomic analysis may be a better indicator of activity.

CHAPTER SIX

CONCLUSIONS/FUTURE DIRECTIONS

The three MAGs assembled here likely represent a significant percentage of the organisms present in the surface waters of the Delaware Bay. Additionally, the functional potential of MAG 22 compared to *P. temperata*, its closest relative, revealed the potential significance this organism has on processes such as biogeochemical cycling; specifically, on carbon and sulfur cycling. The abundance of all three MAGs changes in response to season, suggesting that seasons plays an important role in shaping the bacterial community. In addition, the time of day that the sample was collected from affected the types of genes being expressed, especially the light harvesting proteins, indicating that this organism uses extra energy generated from light energy when available. Lastly, differences in growth and activity related transcripts were also observed between the two different salinities, but only within the spring season, potentially indicative of the organism’s preferred saline range for growth. In the future, more in depth analysis of all three MAGs 22, 73 and 147 will prove useful in quantifying the participation of these organisms in nutrient cycling in the Delaware Bay.

Tables – Results

Table 3. Summer versus fall comparison, only day samples used. Table shows 61 significantly (p-adj <0.05) differentially expressed protein encoding genes (peg) between the two seasons (summer vs. fall). The grey boxes indicate a negative log2fold change and hence are upregulated during the fall compared to summer. (Hypothetical genes, either up or downregulated are not shown).

Category	Peg #	Name	baseMean	log2FoldChange
Carbohydrate metabolism				
Carbohydrate metabolism - Acyl-CoA dehydrogenase	768	Acyl-CoA dehydrogenase	20.28	-1.70
Carbohydrate metabolism - Glycolysis / Gluconeogenesis	1497	Acetyl-coenzyme A synthetase	61.29	-2.70
Aromatic Amin Catabolism	1245	4-hydroxyphenylacetate 3-monooxygenase	15.98	-1.59
Coenzyme A Biosynthesis	1539	Phosphopantothenoylcysteine decarboxylase	29.72	-1.58
Energy metabolism				
Energy metabolism - ammonia assimilation	1434	Glutamate synthase [NADPH] large chain	107.64	-1.17
Energy metabolism - light harvesting proteins	13	Light-harvesting LHI, beta subunit	188.45	-4.15
Energy metabolism - light harvesting proteins	12	Light-harvesting LHI, alpha subunit	20.05	-3.74
Regulator for photosystem formation	2245	PpaA, regulator for photosystem formation	36.58	-1.47
Energy metabolism - Soluble cytochromes and functionally related electron carriers	3	Cytochrome c2	167.88	-1.88
Transport				
Sugar transporter	1203	FIG097052: Sugar transporter	15.95	-1.47
Branched-chain amino acid ABC transporter	1492	Branched-chain amino acid ABC transporter	655.78	-2.28
Branched-chain amino acid ABC transporter	1494	Branched-chain amino acid transport system LivM	48.25	-2.10
Membrane transport	1496	Branched-chain amino acid ABC transporter	30.65	-3.40
Membrane transport	1495	InterPro IPR001687:IPR003439:IPR003593 COGs COG0411	22.47	-3.27
Membrane transport	302	ABC transporter, permease protein, putative	17.63	-1.79
Membrane transport	2085	Oligopeptide ABC transporter, periplasmic	165.30	-1.29

		protein OppA		
Membrane transport	1915	TRAP-type C4-dicarboxylate transport system	236.26	-1.28
Membrane transport	1120	Pyrimidine ABC transporter, substrate-binding component	317.79	-0.70
TRAP Transporter collection	1338	TRAP-type C4-dicarboxylate transport system, periplasmic component	120.54	-1.39
TRAP Transporter collection	1577	TRAP-type C4-dicarboxylate transport system, small permease component	14.92	-1.30
TRAP Transporter collection	1576	TRAP-type C4-dicarboxylate transport system, periplasmic component	222.54	-1.11
TRAP Transporter collection	1131	TRAP-type C4-dicarboxylate transport system, periplasmic component	85.58	-0.90
Stress response				
Multidrug efflux pump	418	Membrane fusion protein of RND family multidrug efflux pump	38.65	-1.45
Cold shock protein	2156	Cold shock protein CspC	107.62	-1.21
DNA replication				
Folate Biosynthesis	1645	GTP cyclohydrolase I	47.01	-1.14
Transcription	1538	RNA polymerase sigma factor RpoH-related protein	154.90	-1.69
Replication and repair	860	Excinuclease ABC subunit A	29.71	-1.33
Protein folding/assembly				
Scaffold proteins for [4Fe-4S] cluster assembly (MRP family)	1658	HflC protein	52.99	-1.44
Scaffold proteins for [4Fe-4S] cluster assembly (MRP family)	1656	HtrA protease/chaperone protein	76.94	-1.36
Proteolysis in bacteria	1735	ATP-dependent Clp protease ATP-binding subunit ClpA	98.39	-0.97
Growth				
Translation	1856	Translation elongation factor LepA	41.92	1.31
Translation	125	LSU ribosomal protein L2p (L8e)	202.69	0.82
Translation	319	LSU ribosomal protein L7/L12 (P1/P2)	102.09	0.89

Translation	2374	Heat shock protein 60 family chaperone GroEL	396.72	0.89
Translation	317	LSU ribosomal protein L1p (L10Ae)	50.53	0.96
Translation	2022	SSU ribosomal protein S13p (S18e)	91.64	0.98
Translation	2373	Heat shock protein 60 family co-chaperone GroES	103.97	1.11
Cell wall recycling	380	Protein often near L-alanine-DL-glutamate epimerase (cell wall recycling)	81.23	1.90
Carbohydrate metabolism				
Glyoxylate and dicarboxylate metabolism	898	Aminomethyl transferase family protein	18.85	2.22
Amino acid biosynthesis				
Amino acid biosynthesis	83	Anthranilate synthase, aminase component	17.74	1.59
Amino acid biosynthesis	141	Argininosuccinate synthase	19.67	1.73
Transport				
Membrane transport	1976	Various polyols ABC transporter, periplasmic protein	98.23	1.16
Membrane transport	1350	Polyamine ABC transporter, permease protein	24.29	1.34

Table 4. Summer night versus day, table shows 52 significantly (p-adj <0.5) differentially expressed genes between the time points within the summer samples (11:00am & 11:00pm). The grey boxes indicate a negative log2fold change and are upregulated during the day compared to night. (Hypothetical genes, either up or downregulated are not shown).

Categories	Peg #	Protein name	baseMean	log2FoldChange
Light energy related proteins				
Bacterial light-harvesting proteins	13	Light-harvesting beta subunit	1742.89	7.62
Carotenoid biosynthesis	22	Hydroxyneurosporene dehydrogenase	7.19	4.24
Carotenoid biosynthesis	23	Phytoene synthase	21.61	4.79
Carotenoid biosynthesis	25	Spheroidene monooxygenase	12.69	4.79
Carotenoid biosynthesis	19	Hydroxyneurosporene methyltransferase	9.33	5.00
Carotenoid biosynthesis	21	Methoxyneurosporene dehydrogenase	12.42	5.01
Carotenoid biosynthesis	24	Phytoene dehydrogenase	38.74	5.82
Chlorophyll Biosynthesis	26	Protoporphyrin IX Mg-chelatase subunit I	12.04	3.13
Chlorophyll Biosynthesis	27	Protoporphyrin IX Mg-chelatase subunit D	13.61	4.33
Chlorophyll Biosynthesis	6	Geranylgeranyl hydrogenase BchP Geranylgeranyl reductase	45.65	5.01
Chlorophyll Biosynthesis	2260	Mg protoporphyrin IX monomethyl ester oxidative cyclase (aerobic)	68.56	5.22
Chlorophyll Biosynthesis	4	Chlorophyll a synthase ChlG	66.39	5.57
Chlorophyll Biosynthesis	2253	Light-independent protochlorophyllide reductase ChlL	82.10	5.63
Chlorophyll Biosynthesis	2254	Mg-protoporphyrin O-methyltransferase	40.03	6.11
Chlorophyll Biosynthesis	2251	Light-independent protochlorophyllide reductase subunit B	59.52	6.52
Chlorophyll Biosynthesis	2252	Protoporphyrin IX Mg-chelatase subunit H	169.88	6.72
Chlorophyll Biosynthesis	16	Chlorophyllide reductase subunit BchY	138.15	6.99
Chlorophyll Biosynthesis	15	Chlorophyllide reductase subunit BchZ	121.65	7.01
Chlorophyll Biosynthesis	2250	Light-independent protochlorophyllide reductase subunit N	65.98	7.03
Chlorophyll Biosynthesis	5	Bacteriochlorophyll synthase 44.5 kDa chain	41.50	4.34

Chlorophyll Biosynthesis	2249	2-vinyl bacteriochlorophyllide hydratase BchF	72.76	7.19
Chlorophyll Biosynthesis	18	2-desacetyl-2-hydroxyethyl bacteriochlorophyllide A dehydrogenase BchC	189.20	7.81
Chlorophyll Biosynthesis	17	Chlorophyllide reductase subunit BchX	171.01	8.05
Photosystem II	2257	Putative photosynthetic complex assembly protein	65.94	5.31
Photosystem II	2256	Photosynthetic reaction center H subunit	149.87	5.62
Photosystem II	10	Photosynthetic reaction center M subunit	344.76	7.17
Photosystem II	12	Light-harvesting alpha subunit	337.42	7.43
Photosystem II	11	Photosynthetic reaction center L subunit	396.87	7.90
Carbohydrate metabolism				
Citrate cycle (TCA cycle)	1173	Aconitate hydratase	138.35	1.16
Sulfur & iron metabolism				
Energy metabolism - sulfur metabolism	3	Cytochrome c2	130.57	0.93
Heme and Siroheme Biosynthesis	2	Uroporphyrinogen III decarboxylase	18.48	2.65
No group				
Amino acid metabolism	2262	5-aminolevulinate synthase	75.29	5.24
Thiamine metabolism	8	1-deoxy-D-xylulose 5-phosphate synthase	100.94	7.12
Growth				
Transcription	2177	RNA polymerase sigma factor RpoH	67.66	0.96
Transcription	2209	Xylose-responsive transcription regulator ROK family	19.91	1.60
Membrane transport				
Membrane transport	2208	Xylose ABC transporter XylF	334.83	1.32
Membrane transport	2255	PucC protein	84.06	6.28
Terpenoid biosynthesis				
Terpenoid backbone biosynthesis	7	Isopentenyl-diphosphate delta-isomerase	11.67	4.21
Terpenoid backbone biosynthesis	20	Octaprenyl diphosphate synthase/ Dimethylallyltransferase/ (2E,6E)-farnesyl diphosphate synthase/ Geranylgeranyl	38.16	6.73

		pyrophosphate synthetase		
Carbohydrate metabolism				
Carbohydrate metabolism - Glycolysis / Gluconeogenesis	870	NADPH-dependent glyceraldehyde-3-phosphate dehydrogenase	4.98	-2.75
Carbohydrate metabolism - Glycolysis / Gluconeogenesis	2199	Phosphoenolpyruvate carboxykinase [ATP]	28.03	-1.52
Amino acid metabolism				
Histidine Biosynthesis	1481	Imidazole glycerol phosphate synthase amidotransferase subunit	6.21	-2.21
Stress response				
multidrug efflux transporter	417	RND multidrug efflux transporter Acriflavin resistance protein	52.88	-1.10

Table 5. Spring medium (20PSU) versus high (30PSU), table shows 235 significantly (p-adj <0.5) differentially expressed genes between the medium and high salinity within spring samples. The grey boxes indicate a negative log2fold change and hence are upregulated in the high salinity compared to medium salinity. (Hypothetical genes, either up or downregulated are not shown).

Categories	Peg #	Protein name	baseMean	log2 FoldChange
Light energy related proteins				
Photosystem II	11	Photosynthetic reaction center L subunit	664.02	-0.69
Photosystem II	10	Photosynthetic reaction center M subunit	547.01	-0.81
Photosystem II	2257	Putative photosynthetic complex assembly protein	105.99	-0.89
Photosystem II	2256	Photosynthetic reaction center H subunit	254.68	-1.13
Bacterial light-harvesting proteins	12	Light-harvesting alpha subunit	877.18	-0.46
Bacterial light-harvesting proteins	13	Light-harvesting beta subunit	3969.01	-1.16
Photosystem formation	2245	heme-binding SCHIC domain regulator for photosystem formation	210.50	-0.96
Carotenoid biosynthesis	25	Spheroidene monooxygenase	54.61	-0.78
Carotenoid biosynthesis	19	Hydroxyneurosporene methyltransferase	53.85	-0.90
Carotenoid biosynthesis	24	Phytoene dehydrogenase	130.11	-0.96
Chlorophyll Biosynthesis	2252	Protoporphyrin IX Mg-chelatase subunit H	452.73	-0.64
Chlorophyll Biosynthesis	17	Chlorophyllide reductase subunit BchX	452.96	-0.73
Chlorophyll Biosynthesis	15	Chlorophyllide reductase subunit BchZ	367.68	-0.78
Chlorophyll Biosynthesis	2254	Mg-protoporphyrin O-methyltransferase	124.39	-0.84
Chlorophyll Biosynthesis	2260	Mg protoporphyrin IX monomethyl ester oxidative cyclase (aerobic)	187.68	-0.87
Chlorophyll Biosynthesis	6	Geranylgeranyl hydrogenase BchlB Geranylgeranyl reductase	116.66	-0.92
Chlorophyll Biosynthesis	2250	Light-independent protochlorophyllide reductase subunit N	137.69	-1.01
Chlorophyll Biosynthesis	2249	2-vinyl bacteriochlorophyllide hydratase BchF	151.81	-1.03
Chlorophyll Biosynthesis	2251	Light-independent protochlorophyllide reductase subunit B	185.36	-1.10

Chlorophyll Biosynthesis	2253	Light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein ChlL	286.62	-1.22
Carbon Metabolism				
Carbon monoxide oxidation	1906	Carbon monoxide dehydrogenase large chain	684.59	-0.42
Carbon monoxide oxidation	1905	Carbon monoxide dehydrogenase small chain	191.18	-0.58
Pentose and glucuronate interconversions	43	Multiple polyol-specific dehydrogenase	13.29	-1.35
Glycolysis / Gluconeogenesis	1348	Aldehyde dehydrogenase	33.32	-1.06
Citrate cycle (TCA cycle)	2172	HpcH/HpaI aldolase	22.14	-1.10
2,4-dienoyl-CoA reductase	655	2,4-dienoyl-CoA reductase [NADPH]	102.08	-1.13
Acyl-CoA dehydrogenase	256	Acyl-CoA dehydrogenase	64.21	-1.05
Amino acid metabolism				
Amino acid metabolism	2262	5-aminolevulinate synthase	175.10	-0.80
Amino acid metabolism	945	Sarcosine dehydrogenase	109.01	-0.91
Amino acid metabolism	1268	Phenylacetic acid degradation protein ring-opening aldehyde dehydrogenase	32.06	-0.94
Amino acid metabolism	1269	Enoyl-CoA hydratase	18.59	-1.14
Amino acid metabolism	600	Sarcosine oxidase beta subunit	28.49	-1.15
Amino acid metabolism	1137	N-methylhydantoinase A	19.71	-1.19
Amino acid metabolism	598	Sarcosine oxidase alpha subunit	44.73	-1.31
Amino acid metabolism	1954	Glutathione peroxidase family protein	17.33	-2.16
Lipid metabolism				
Lipid metabolism	894	Acetyl-coenzyme A synthetase	180.29	-0.59
Lipid metabolism	1186	Long-chain-fatty-acid-CoA ligase	289.48	-0.59
Lipid metabolism	1077	Phosphatidylcholine synthase	38.76	-0.99
Lipid metabolism	1499	Biotin carboxyl carrier protein of acetyl-CoA carboxylase	29.71	-1.16
Nitrogen fixation				
Nitrogen fixation	1526	NifU-like domain protein	194.90	-0.55
Sulfur metabolism				

Sulfur metabolism	2369	Peptide methionine sulfoxide reductase MsrA	15.07	-1.17
Sulfur metabolism	2414	rhodanese domain protein	32.72	-1.28
Sulfur metabolism	944	Homocysteine S-methyltransferase	25.29	-1.28
No group				
Protein kinase	2292	Two component sensor kinase	16.01	-1.20
Exoenzymes	2087	Exoenzymes regulatory protein AepA	26.94	-1.07
Terpenoid biosynthesis	20	Octaprenyl diphosphate synthase	217.48	-0.73
Xenobiotics biodegradation and metabolism	2488	homoprotocatechuate 2,3-dioxygenase	18.00	-1.80
Metabolism of cofactors and vitamins	8	1-deoxy-D-xylulose 5-phosphate synthase	229.51	-0.66
Growth				
Nucleotide metabolism	1088	Uracil-xanthine permease	712.48	-0.49
Nucleotide metabolism	1553	Xanthine dehydrogenase,C molybdenum binding subunit	111.91	-0.71
Nucleotide metabolism	1554	Xanthine dehydrogenase, iron-sulfur cluster and FAD-binding subunit A	71.76	-0.81
Nucleotide metabolism	266	5'-nucleotidase	110.96	-0.83
DNA replication	1410	DNA-binding protein HU	324.29	-0.43
DNA replication	121	Chromosome partition protein smc	72.10	-0.82
DNA replication	2386	Integration host factor beta subunit	147.82	-0.88
Transcription	1982	Maltose operon transcriptional repressor MalR, LacI family	27.74	-1.02
Transcription	2177	RNA polymerase sigma factor RpoH	64.09	-1.09
Transcription	29	Transcriptional regulator, ArsR family	20.77	-1.18
Transcription	257	Predicted transcriptional regulator LiuR of leucine degradation pathway MerR family	48.65	-1.36
Transcription	1538	RNA polymerase sigma factor RpoH-related protein	26.90	-2.40
Transporters				
Tricarboxylate transporter	2294	Tricarboxylate transport protein TctC	297.81	-0.42
Membrane transporter	1510	L-proline glycine betaine binding ABC transporter protein ProX	619.27	-0.35

Membrane transporter	1022	TRAP-type C4-dicarboxylate transport system, periplasmic component	474.31	-0.40
Membrane transporter	1548	Nucleoside ABC transporter, periplasmic nucleoside-binding protein	476.91	-0.42
Membrane transporter	1048	ABC transporter substrate binding protein	1072.10	-0.44
Membrane transporter	1803	Alpha-glucosides-binding periplasmic protein AglE precursor	728.11	-0.50
Membrane transporter	2458	Leucine, isoleucine, valine, threonine and alanine-binding protein	986.49	-0.51
Membrane transporter	1315	TRAP-type C4-dicarboxylate transport system large permease component	191.10	-0.53
Membrane transporter	842	Oligopeptide/dipeptide ABC transporter periplasmic substrate-binding protein	553.82	-0.56
Membrane transporter	1339	TRAP-type transport system predicted N-acetylneuraminate transporter	139.04	-0.57
Membrane transporter	442	Peptide/opine/nickel uptake family ABC transporter	945.12	-0.63
Membrane transporter	2343	N-Acetyl-D-glucosamine ABC transport system sugar-binding protein	590.46	-0.74
Membrane transporter	1120	Pyrimidine ABC transporter substrate-binding component	541.08	-0.75
Membrane transporter	1260	TRAP transporter solute receptor unknown substrate 6	146.10	-0.76
Membrane transporter	1338	TRAP-type C4-dicarboxylate transport system periplasmic component	1972.60	-0.78
Membrane transporter	2255	PucC protein	190.12	-0.80
Membrane transporter	1123	Pyrimidine ABC transporter ATP-binding protein	43.99	-1.05
Membrane transporter	2457	Branched-chain amino acid transport system permease protein LivM	68.32	-1.06
Membrane transporter	1706	Glycerol-3-phosphate ABC transporter permease protein UgpE	40.60	-1.08
Membrane transporter	663	Glycerol-3-phosphate ABC transporter periplasmic	222.79	-1.09
Membrane transporter	664	Glycerol-3-phosphate ABC transporter ATP-binding protein UgpC	22.08	-1.09
Membrane transporter	1708	Glycerol-3-phosphate ABC transporter periplasmic	852.60	-1.24
Membrane transporter	1976	Various polyols ABC transporter periplasmic substrate-binding protein	345.87	-1.26
Membrane transporter	1705	Glycerol-3-phosphate ABC transporter UgpC	43.72	-1.27

Membrane transporter	1131	TRAP-type C4-dicarboxylate transport system periplasmic component	361.84	-1.29
Membrane transporter	1978	Maltose/maltodextrin ABC transporter permease protein MalG	31.70	-1.33
Membrane transporter	1977	binding-protein-dependent transport systems inner membrane component	61.47	-1.40
Membrane transporter	41	Various polyols ABC transporter permease component 2	13.31	-1.55
Membrane transporter	2084	Oligopeptide transport system permease protein OppB	28.43	-1.62
Membrane transporter	2085	ABC transporter oligopeptide-binding protein OppA	235.23	-1.62
Membrane transporter	44	Maltose/maltodextrin transport ATP-binding protein MalK	14.02	-1.73
Membrane transporter	1707	Glycerol-3-phosphate ABC transporter permease protein UgpA	91.26	-1.78
Membrane transporter	1570	High-affinity leucine-specific transport system LivK	59.58	-2.00
Membrane transporter	39	Various polyols ABC transporter substrate-binding protein	462.70	-3.27
Stress response				
SOS-response	815	SOS-response repressor and protease LexA	47.37	-0.84
Cold shock	1610	Cold shock protein CspA	51.91	-0.94
Carbon metabolism				
Citrate cycle (TCA cycle)	1173	Aconitate hydratase	160.17	0.99
Citrate cycle (TCA cycle)	1295	Isocitrate dehydrogenase [NADP]	63.07	0.79
Glycolysis / Gluconeogenesis	546	Aldehyde dehydrogenase	97.73	0.73
Glycolysis / Gluconeogenesis	373	Pyruvate kinase	110.24	0.65
Inositol catabolism	606	5-keto-2-deoxygluconokinase	45.23	0.89
Methane metabolism	533	Sulfopyruvate decarboxylase - alpha subunit	56.72	1.25
Methane metabolism	532	Sulfopyruvate decarboxylase - beta subunit	72.36	0.90
Pentose phosphate pathway	872	Transketolase	109.44	0.80
Energy metabolism				
Oxidative phosphorylation	1741	ATP synthase beta chain	228.71	0.93
Oxidative phosphorylation	1739	ATP synthase alpha chain	345.38	0.90
Oxidative phosphorylation	619	Cytochrome c oxidase polypeptide III	62.33	0.89

Oxidative phosphorylation	1740	ATP synthase gamma chain	105.70	0.76
Oxidative phosphorylation	1738	ATP synthase delta chain	103.84	0.70
Oxidative phosphorylation	700	NADH-ubiquinone oxidoreductase chain M	85.54	0.64
Oxidative phosphorylation	977	Ubiquinol--cytochrome c reductase, cytochrome B subunit	93.82	0.61
Sulfur metabolism				
Sulfur metabolism	1806	Aliphatic sulfonate monooxygenase family FMNH2- or F420-dependent	51.27	1.88
Sulfur metabolism	1621	Granule-associated protein	36.35	1.19
Sulfur metabolism	728	Cysteine desulfurase	57.80	0.99
Sulfur metabolism	978	ubiquinol cytochrome C oxidoreductase cytochrome C1 subunit	70.38	0.87
Lipid metabolism				
Lipid biosynthesis	2474	2-Keto-3-deoxy-D-manno-octulosonate-8-phosphate synthase	19.75	1.16
Lipid metabolism	1152	Acyl carrier protein	172.51	1.05
Lipid metabolism	484	Phosphate:acyl-ACP acyltransferase PlsX	75.36	0.72
No group				
Metabolism of cofactors and vitamins	94	CobW GTPase involved in cobalt insertion for B12 biosynthesis	27.47	0.98
Xenobiotics biodegradation and metabolism	1968	Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases	26.53	2.11
Quorum sensing	552	Signal recognition particle subunit Ffh SRP54	43.52	1.06
Amino acid metabolism				
Amino acid metabolism	850	Proline dehydrogenase	61.65	2.91
Amino acid metabolism	848	Arginase	21.83	2.65
Amino acid metabolism	2416	Aminomethyltransferase (glycine cleavage system T protein)	34.37	1.31
Amino acid metabolism	656	Acetolactate synthase large subunit	58.36	1.05
Amino acid metabolism	2417	Glycine cleavage system H protein	29.93	1.00
Amino acid metabolism	1383	Aspartate-semialdehyde dehydrogenase	63.19	0.96
Amino acid metabolism	276	Ketol-acid reductoisomerase	152.18	0.94

Amino acid metabolism	2418	Glycine dehydrogenase [decarboxylating]	98.64	0.87
Amino acid metabolism	334	Acetolactate synthase large subunit	95.91	0.81
Amino acid metabolism	83	Anthranilate synthase, aminase component	44.44	0.81
Amino acid metabolism	1371	3-isopropylmalate dehydratase large subunit	93.96	0.73
Nitrogen metabolism				
Ammonia-lyases	849	Ornithine cyclodeaminase	20.15	3.38
Growth				
Transcription	2024	DNA-directed RNA polymerase alpha subunit	330.78	1.43
Transcription	2518	Cold shock protein CspC	36.09	1.30
Transcription	315	Transcription antitermination protein NusG	133.43	1.24
Transcription	1044	Transcription termination protein NusA	134.15	0.73
Transcription	748	DNA-directed RNA polymerase omega subunit	96.19	0.67
Transcription	320	DNA-directed RNA polymerase beta subunit	578.85	0.50
Translation	1471	Translation initiation factor 1	42.56	1.40
Translation	547	Ribosomal large subunit pseudouridine synthase A	17.48	1.11
Translation	292	Translation elongation factor G	614.51	1.10
Translation	2053	Methionine aminopeptidase	37.38	1.09
Translation	1831	Aspartyl-tRNA synthetase	54.51	1.01
Translation	2517	Glutamyl-tRNA(Gln) synthetase	35.78	1.01
Translation	2152	Translation elongation factor Ts	113.14	0.97
Translation	270	Glycyl-tRNA synthetase beta chain	47.45	0.96
Translation	1770	Lysyl-tRNA synthetase (class I)	42.29	0.83
Translation	207	Ribonuclease E	287.30	0.43
Ribosome LSU bacterial	318	LSU ribosomal protein L10p (P0)	467.43	1.35
Ribosome LSU bacterial	1883	LSU ribosomal protein L27p	80.52	1.34
Ribosome LSU bacterial	317	LSU ribosomal protein L1p (L10Ae)	120.19	1.26
Ribosome LSU bacterial	129	LSU ribosomal protein L16p (L10e)	108.86	1.22

Ribosome LSU bacterial	2025	LSU ribosomal protein L17p	135.56	1.22
Ribosome LSU bacterial	319	LSU ribosomal protein L7/L12 (P1/P2)	253.98	1.21
Ribosome LSU bacterial	111	LSU ribosomal protein L13p (L13Ae)	168.56	1.21
Ribosome LSU bacterial	2424	LSU ribosomal protein L34p	17.76	1.18
Ribosome LSU bacterial	1157	LSU ribosomal protein L9p	148.04	1.17
Ribosome LSU bacterial	2007	LSU ribosomal protein L23p (L23Ae)	98.81	1.13
Ribosome LSU bacterial	2010	LSU ribosomal protein L14p (L23e)	181.40	1.12
Ribosome LSU bacterial	316	LSU ribosomal protein L11p (L12e)	84.73	1.12
Ribosome LSU bacterial	2379	LSU ribosomal protein L25p	216.82	1.11
Ribosome LSU bacterial	2005	LSU ribosomal protein L3p (L3e)	188.44	1.07
Ribosome LSU bacterial	125	LSU ribosomal protein L2p (L8e)	300.23	1.04
Ribosome LSU bacterial	2016	LSU ribosomal protein L18p (L5e)	141.12	1.02
Ribosome LSU bacterial	2012	LSU ribosomal protein L5p (L11e)	174.57	1.01
Ribosome LSU bacterial	2006	LSU ribosomal protein L4p (L1e)	119.14	0.97
Ribosome LSU bacterial	2011	LSU ribosomal protein L24p (L26e)	72.89	0.97
Ribosome LSU bacterial	2015	LSU ribosomal protein L6p (L9e)	146.13	0.97
Ribosome LSU bacterial	1884	LSU ribosomal protein L21p	222.77	0.96
Ribosome LSU bacterial	370	LSU ribosomal protein L20p	94.42	0.90
Ribosome LSU bacterial	2009	LSU ribosomal protein L29p (L35e)	130.39	0.85
Ribosome LSU bacterial	2018	LSU ribosomal protein L30p (L7e)	56.98	0.80
Ribosome LSU bacterial	2019	LSU ribosomal protein L15p (L27Ae)	174.06	0.75
Ribosome LSU bacterial	371	LSU ribosomal protein L35p	86.40	0.70
Ribosome LSU bacterial	558	LSU ribosomal protein L19p	132.60	0.66
Ribosome LSU bacterial	127	LSU ribosomal protein L22p (L17e)	91.49	0.64
Ribosome SSU bacterial	362	SSU ribosomal protein S21p	309.62	2.50
Ribosome SSU bacterial	2153	SSU ribosomal protein S2p (SAe)	154.58	1.64
Ribosome SSU bacterial	554	SSU ribosomal protein S16p	57.29	1.42

Ribosome SSU bacterial	110	SSU ribosomal protein S9p (S16e)	109.92	1.36
Ribosome SSU bacterial	1155	SSU ribosomal protein S6p	195.03	1.32
Ribosome SSU bacterial	1726	SSU ribosomal protein S4p (S9e)	169.42	1.32
Ribosome SSU bacterial	2013	SSU ribosomal protein S14p (S29e), zinc-independent	65.06	1.27
Ribosome SSU bacterial	2004	SSU ribosomal protein S10p (S20e)	142.66	1.19
Ribosome SSU bacterial	1156	SSU ribosomal protein S18p, zinc-independent	42.22	1.18
Ribosome SSU bacterial	2017	SSU ribosomal protein S5p (S2e)	186.48	1.18
Ribosome SSU bacterial	128	SSU ribosomal protein S3p (S3e)	184.71	1.15
Ribosome SSU bacterial	2022	SSU ribosomal protein S13p (S18e)	228.60	1.06
Ribosome SSU bacterial	542	SSU ribosomal protein S15p (S13e)	272.41	1.06
Ribosome SSU bacterial	2387	SSU ribosomal protein S1p	436.65	1.01
Ribosome SSU bacterial	2014	SSU ribosomal protein S8p (S15Ae)	74.75	0.89
Ribosome SSU bacterial	293	SSU ribosomal protein S7p (S5e)	95.74	0.88
Ribosome SSU bacterial	126	SSU ribosomal protein S19p (S15e)	67.86	0.74
Ribosome SSU bacterial	294	SSU ribosomal protein S12p (S23e)	99.93	1.02
Nucleotide metabolism	1463	GMP synthase [glutamine-hydrolyzing]	45.09	1.05
Nucleotide metabolism	1317	Nucleoside diphosphate kinase	40.33	1.02
Nucleotide metabolism	544	Polyribonucleotide nucleotidyltransferase	333.86	1.00
Nucleotide metabolism	1166	Adenylosuccinate lyase	32.36	0.99
Nucleotide metabolism	1639	Adenylosuccinate synthetase	38.77	0.97
DNA replication	138	Chaperonin 33 kDa	17.97	1.17
Cell division trigger factor	1158	Cell division trigger factor	211.63	0.97
Cell wall recycling	380	Protein often near L-alanine-DL-glutamate epimerase (cell wall recycling)	38.58	1.64
Transporters				
Membrane transporter	1785	putative ABC transporter solute-binding protein	130.81	1.51
Membrane transporter	1966	High-affinity branched-chain amino acid transport system permease protein LivH	12.76	1.36

Membrane transporter	1243	ABC transporter ATP-binding protein uup	23.93	1.10
Membrane transporter	1967	Branched-chain amino acid ABC transporter amino acid-binding protein	79.06	0.92
Membrane transporter	626	TRAP-type C4-dicarboxylate transport system large permease component	156.39	0.65
Membrane transporter	625	TRAP-type C4-dicarboxylate transport system periplasmic component	889.02	0.64
Tricarboxylate transporter	1998	Tricarboxylate transport protein TctC	1151.75	0.36
Stress response				
Multidrug and toxin efflux pump	1601	Multidrug and toxin extrusion (MATE) family efflux pump YdhE/NorM,C homolog	21.66	1.03
RNA degradation	2054	ATP-dependent RNA helicase RhlE	21.56	1.47
Heat shock protein	2374	Heat shock protein 60 family chaperone GroEL	905.13	1.28
Heat shock protein	2373	Heat shock protein 60 family co-chaperone GroES	221.35	0.85
Heat shock protein	2233	Chaperone protein DnaK	294.92	0.69

APPENDICES

Appendix A

Scripts used

Assigning K0 numbers and KEGG categories to pegs – script by Jason Gholamian

```
#!/bin/bash

#Usage example: ./peg2k0.sh -p p2k_pathways.json -r ./rast_k0.txt -l ./peg_ids_list.txt -o ./output

while getopts p:r:l:o: option
do
 case "${option}"
 in
 p) pathway_file+=${OPTARG};;
 r) rast_file+=${OPTARG};;
 l) list_file+=${OPTARG};;
 o) output_dir+=${OPTARG};;
 esac
done

if [ ${pathway_file} == "p2k_pathways.json" ] ; then
 echo "Since the default value for option -p was used, the most recent version of K0 Table (pathways) will be downloaded and used."
 echo "Downloading http://www.genome.jp/kegg-bin/download_htext?htext=ko00000.keg&format=json&filedir="
 wget -O p2k_pathways.json "http://www.genome.jp/kegg-bin/download_htext?htext=ko00000.keg&format=json&filedir="
 pathway_file="./p2k_pathways.json"
fi

mkdir -p ${output_dir}

python ./file2tsv.py ${pathway_file} ${rast_file} ${output_dir}

echo "Creating the database..."

rm ${output_dir}/p2k.sqlite3 2> /dev/null

sqlite3 ${output_dir}/p2k.sqlite3 <<EOF
.mode tabs
.import ${output_dir}/p2k_K0_table.tsv K0_table
.import ${output_dir}/p2k_RAST_K0.tsv RAST_K0
.quit
EOF

python ./pegcat.py ${list_file} ${output_dir}

echo "Done!"
```

Counting number of genes per category – script by Jean Lim

```
VAR1=$1
OUT=$2

for i in `seq 1 3`
do
cat $VAR1 | awk -F "\t" -v i=$i '{print $i}' | sort | uniq | tr -d "\"" | grep -v "Category" | sed "s/\[.*\]//g"
>$OUT.Category"$i".header
HEAD=`head -1 $VAR1 | awk -F "\t" '{for(i=7;i<=NF;i++){print $i}}' | tr "\n" "#" | sed "s/^/Category
"$i#"/g"`
echo $HEAD >>$OUT.Category"$i".tab.txt

while read line; do
grep "$line" $VAR1 | awk -F "\t" '{for(i=7;i<=NF;i++){sum[i]+=$i}}END{for(i=7;i<=NF;i++)  {print
sum[i]}}' | tr "\n" "#" >>$OUT.Category"$i".count
echo "" >>$OUT.Category"$i".count
done < $OUT.Category"$i".header

paste -d "#" $OUT.Category"$i".header $OUT.Category"$i".count >>$OUT.Category"$i".tab.txt

done

rm $OUT.Category*count
```

Appendix B

Unique genes in each MAG – supplemental tables

Table S1a. Presence in Spr20L08 MAG 22 but not in closest relative *Planktomarina temperata* RCA23 (23 total)

Category	**Subcategory**	**Subsystem**	**Role**
Carbohydrates	Central carbohydrate metabolism	Entner-Doudoroff Pathway	Phosphoglycerate kinase (EC 2.7.2.3)
Carbohydrates	Sugar alcohols	Inositol catabolism	Epi-inositol hydrolase (EC 3.7.1.-)
Clustering-based subsystems	alpha-proteobacterial cluster of hypotheticals	CBSS-52598.3.2843	FIG017823: ATPase, MoxR family
Clustering-based subsystems	alpha-proteobacterial cluster of hypotheticals	CBSS-52598.3.2843	FIG139612: Possible conserved membrane protein
Clustering-based subsystems	no subcategory	CBSS-211586.1.2832	Protein-export membrane protein SecF (TC 3.A.5.1.1)
Clustering-based subsystems	no subcategory	CBSS-292414.1.563	FIG119243: hypothetical protein
Clustering-based subsystems	no subcategory	ClpAS cluster	ATP-dependent Clp protease adaptor protein ClpS
Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	5-FCL-like protein	Butyryl-CoA dehydrogenase (EC 1.3.8.1)
Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	5-FCL-like protein	Phosphoribosylglycinamide formyltransferase (EC 2.1.2.2)
DNA Metabolism	DNA repair	DNA Repair Base Excision	Formamidopyrimidine-DNA glycosylase (EC 3.2.2.23)
Miscellaneous	No subcategory	Broadly distributed proteins not in subsystems	YciL protein

Nucleosides and Nucleotides	Detoxification	Nucleoside triphosphate pyrophosphohydrolase MazG	Nucleoside triphosphate pyrophosphohydrolase MazG (EC 3.6.1.8)
Protein Metabolism	Protein processing and modification	G3E family of P-loop GTPases (metallocenter biosynthesis)	Urease beta subunit (EC 3.5.1.5)
Nucleosides and Nucleotides	Purines	Purine conversions	Adenylosuccinate synthetase (EC 6.3.4.4)
RNA Metabolism	RNA processing and modification	RNA methylation	Ribosomal RNA small subunit methyltransferase C (EC 2.1.1.52)
RNA Metabolism	RNA processing and modification	RNA methylation	tRNA:Cm32/Um32 methyltransferase
RNA Metabolism	RNA processing and modification	mnm5U34 biosynthesis bacteria	tRNA 5-methylaminomethyl-2-thiouridine synthase TusA
Regulation and Cell signaling	No subcategory	LysR-family proteins in Escherichia coli	Hydrogen peroxide-inducible genes activator
Respiration	No subcategory	Biogenesis of cytochrome c oxidases	Heme A synthase, cytochrome oxidase biogenesis protein Cox15-CtaA
Stress Response	Oxidative stress	Cluster containing Glutathione synthetase	Ribosomal RNA small subunit methyltransferase E (EC 2.1.1.-)
Stress Response	Oxidative stress	Glutathione: Non-redox reactions	Uncharacterized glutathione S-transferase-like protein
Virulence, Disease and Defense	Bacteriocins, ribosomally synthesized antibacterial peptides	Colicin V and Bacteriocin Production Cluster	Dihydrofolate synthase (EC 6.3.2.12)
Virulence, Disease and Defense	Bacteriocins, ribosomally synthesized antibacterial peptides	Colicin V and Bacteriocin Production Cluster	Folylpolyglutamate synthase (EC 6.3.2.17)

Table S1b. Presence in Sum29NL08 MAG 147 but not in closest relative *Rhodobacteraceae bacterium* HIMB11 (17 total)

Category	Subcategory	Subsystem	Role
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Arginine decarboxylase (EC 4.1.1.19)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	Ornithine decarboxylase (EC 4.1.1.17)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	S-adenosylmethionine decarboxylase proenzyme (EC 4.1.1.50), prokaryotic class 1B
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Putrescine utilization pathways	Gamma-aminobutyrate:alpha-ketoglutarate aminotransferase (EC 2.6.1.19)
Carbohydrates	Central carbohydrate metabolism	Glyoxylate bypass	Isocitrate lyase (EC 4.1.3.1)
Carbohydrates	Sugar alcohols	Glycerol and Glycerol-3-phosphate Uptake and Utilization	Glycerol-3-phosphate regulon repressor GlpR
Cofactors, Vitamins, Prosthetic Groups, Pigments	NAD and NADP	NAD and NADP cofactor biosynthesis global	Nicotinamidase/isochorismatase family protein
DNA Metabolism	DNA repair	DNA repair system including RecA, MutS and a hypothetical protein	Protein Implicated in DNA repair function with RecA and MutS
Fatty Acids, Lipids, and Isoprenoids	Phospholipids	Glycerolipid and Glycerophospholipid Metabolism in Bacteria	Aldehyde dehydrogenase B (EC 1.2.1.22)
Membrane Transport	ABC transporters	ABC transporter dipeptide (TC 3.A.1.5.2)	Dipeptide transport system permease protein DppB (TC 3.A.1.5.2)
Membrane Transport	TRAP transporters	TRAP Transporter unknown substrate 5	TRAP dicarboxylate transporter, DctM subunit, unknown substrate 5
Membrane Transport	TRAP transporters	TRAP Transporter	TRAP dicarboxylate transporter, DctQ subunit,

		unknown substrate 5	unknown substrate 5
Membrane Transport	TRAP transporters	TRAP Transporter unknown substrate 5	TRAP transporter solute receptor, unknown substrate 5
Miscellaneous	no subcategory	Phosphoglycerate mutase protein family	Phosphoglycerate mutase family 4
Respiration	Electron donating reactions	Succinate dehydrogenase	Fumarate reductase flavoprotein subunit (EC 1.3.99.1)
Stress Response	Cold shock	Cold shock, CspA family of proteins	Cold shock protein CspC
Sulfur Metabolism	no subcategory	Sulfur oxidation	Sulfur oxidation protein SoxZ

Table S1c. Presence in Sum29DG08 MAG 73 but not in closest relative *Rhodobacteraceae bacterium* HIMB11 (63 total)

Category	**Subcategory**	**Subsystem**	**Role**
Amino Acids and Derivatives	Alanine, serine, and glycine	Glycine and Serine Utilization	L-serine dehydratase (EC 4.3.1.17)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	ABC transporter, periplasmic spermidine putrescine-binding protein PotD (TC 3.A.1.11.1)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Polyamine Metabolism	Putrescine transport ATP-binding protein PotA (TC 3.A.1.11.1)
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Putrescine utilization pathways	Gamma-aminobutyrate:alpha-ketoglutarate aminotransferase (EC 2.6.1.19)
Amino Acids and Derivatives	Aromatic amino acids and derivatives	Chorismate: Intermediate for synthesis of Tryptophan, PAPA antibiotics, PABA, 3-hydroxyanthranilate and more.	Isochorismatase (EC 3.3.2.1)
Amino Acids and Derivatives	Glutamine, glutamate, aspartate, asparagine; ammonia assimilation	Glutamine, Glutamate, Aspartate and Asparagine Biosynthesis	Glutamine amidotransferase class-I (EC 6.3.5.2)
Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Methionine Biosynthesis	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase (EC 2.1.1.14)
Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Methionine Biosynthesis	Methionine ABC transporter ATP-binding protein
Amino Acids and Derivatives	no subcategory	Creatine and Creatinine Degradation	Cytosine deaminase (EC 3.5.4.1)
Carbohydrates	Central carbohydrate metabolism	Pyruvate metabolism I: anaplerotic reactions, PEP	Phosphoenolpyruvate carboxylase (EC 4.1.1.31)

Carbohydrates	Central carbohydrate metabolism	TCA Cycle	Aconitate hydratase (EC 4.2.1.3)
Carbohydrates	Central carbohydrate metabolism	TCA Cycle	Fumarate hydratase class II (EC 4.2.1.2)
Carbohydrates	Monosaccharides	D-galactonate catabolism	Galactonate dehydratase (EC 4.2.1.6)
Carbohydrates	Monosaccharides	D-ribose utilization	Ribose ABC transport system, ATP-binding protein RbsA (TC 3.A.1.2.1)
Carbohydrates	Monosaccharides	D-ribose utilization	Ribose ABC transport system, permease protein RbsC (TC 3.A.1.2.1)
Carbohydrates	Monosaccharides	Xylose utilization	Xylose isomerase (EC 5.3.1.5)
Carbohydrates	Organic acids	2-methylcitrate to 2-methylaconitate metabolism cluster	2-Methylcitrate dehydratase AcnD
Carbohydrates	Organic acids	2-methylcitrate to 2-methylaconitate metabolism cluster	2-methylaconitate racemase
Carbohydrates	Sugar alcohols	Glycerol and Glycerol-3-phosphate Uptake and Utilization	Glycerol-3-phosphate ABC transporter, permease protein UgpE (TC 3.A.1.1.3)
Carbohydrates	Sugar alcohols	Glycerol and Glycerol-3-phosphate Uptake and Utilization	Glycerol-3-phosphate regulon repressor GlpR
Carbohydrates	Sugar alcohols	Inositol catabolism	5-keto-2-deoxy-D-gluconate-6 phosphate aldolase (EC 4.1.2.29)
Carbohydrates	Sugar alcohols	Inositol catabolism	Inosose isomerase (EC 5.3.99.-)
Carbohydrates	Sugar alcohols	Inositol catabolism	Myo-inositol 2-dehydrogenase 2 (EC 1.1.1.18)
Clustering-based subsystems	Cytochrome biogenesis	CBSS-196164.1.461	Glutamate-1-semialdehyde aminotransferase (EC 5.4.3.8)
Clustering-based subsystems	no subcategory	KDO2-Lipid A biosynthesis cluster 2	Lipid A export ATP-binding/permease protein MsbA (EC 3.6.3.25)

Clustering-based subsystems	no subcategory	LMPTP YfkJ cluster	Low molecular weight protein tyrosine phosphatase (EC 3.1.3.48)
Clustering-based subsystems	no subcategory	Llipid A biosynthesis cluster	UDP-2,3-diacylglucosamine pyrophosphatase
Cofactors, Vitamins, Prosthetic Groups, Pigments	Biotin	Biotin biosynthesis	Biotin-protein ligase (EC 6.3.4.15)
Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	Molybdenum cofactor biosynthesis	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family
Cofactors, Vitamins, Prosthetic Groups, Pigments	Tetrapyrroles	Cobalamin synthesis	Alpha-ribazole-5'-phosphate phosphatase (EC 3.1.3.73)
Cofactors, Vitamins, Prosthetic Groups, Pigments	Tetrapyrroles	Heme and Siroheme Biosynthesis	Uroporphyrinogen III decarboxylase (EC 4.1.1.37)
DNA Metabolism	DNA repair	DNA repair, bacterial	DNA polymerase IV-like protein ImuB
DNA Metabolism	DNA repair	DNA repair system including RecA, MutS and a hypothetical protein	Protein Implicated in DNA repair function with RecA and MutS
Fatty Acids, Lipids, and Isoprenoids	Phospholipids	Glycerolipid and Glycerophospholipid Metabolism in Bacteria	Aldehyde dehydrogenase B (EC 1.2.1.22)
Fatty Acids, Lipids, and Isoprenoids	Triacylglycerols	Triacylglycerol metabolism	Lysophospholipase (EC 3.1.1.5)
Fatty Acids, Lipids, and Isoprenoids	Triacylglycerols	Triacylglycerol metabolism	Monoglyceride lipase (EC 3.1.1.23)
Fatty Acids, Lipids, and Isoprenoids	no subcategory	Polyhydroxybutyrate metabolism	Acetoacetyl-CoA synthetase (EC 6.2.1.16)
Membrane Transport	ABC transporters	ABC transporter dipeptide	Dipeptide transport system permease protein DppB

		(TC 3.A.1.5.2)	(TC 3.A.1.5.2)
Membrane Transport	ABC transporters	ABC transporter dipeptide (TC 3.A.1.5.2)	Dipeptide-binding ABC transporter, periplasmic substrate-binding component (TC 3.A.1.5.2)
Membrane Transport	ABC transporters	Periplasmic-Binding-Protein-Dependent Transport System for α-Glucosides	Transcriptional regulator AglR, LacI family
Metabolism of Aromatic Compounds	Peripheral pathways for catabolism of aromatic compounds	Biphenyl Degradation	Acetaldehyde dehydrogenase (EC 1.2.1.10)
Metabolism of Aromatic Compounds	Peripheral pathways for catabolism of aromatic compounds	Biphenyl Degradation	Large subunit naph/bph dioxygenase
Metabolism of Aromatic Compounds	Peripheral pathways for catabolism of aromatic compounds	Biphenyl Degradation	biphenyl-2,3-diol 1,2-dioxygenase III-related protein
Nitrogen Metabolism	no subcategory	Ammonia assimilation	Ferredoxin-dependent glutamate synthase (EC 1.4.7.1)
Nitrogen Metabolism	no subcategory	Ammonia assimilation	Glutamate synthase [NADPH] putative GlxC chain (EC 1.4.1.13)
Nitrogen Metabolism	no subcategory	Ammonia assimilation	Glutamate-ammonia-ligase adenylyltransferase (EC 2.7.7.42)
Nitrogen Metabolism	no subcategory	Ammonia assimilation	Glutamine amidotransferase protein GlxB (EC 2.4.2.-)
Nucleosides and Nucleotides	Purines	Purine Utilization	Xanthine dehydrogenase iron-sulfur subunit (EC 1.17.1.4)
Nucleosides and Nucleotides	Purines	Purine Utilization	Xanthine dehydrogenase, FAD binding subunit (EC 1.17.1.4)
Nucleosides and Nucleotides	Pyrimidines	Pyrimidine utilization	Pyridine nucleotide-disulphide oxidoreductase associated with reductive pyrimidine catabolism

Phosphorus Metabolism	no subcategory	Phosphate metabolism	Probable low-affinity inorganic phosphate transporter
Protein Metabolism	Protein biosynthesis	Ribosome LSU bacterial	LSU ribosomal protein L36p
RNA Metabolism	RNA processing and modification	Queuosine-Archaeosine Biosynthesis	Epoxyqueuosine (oQ) reductase QueG
Regulation and Cell signaling	no subcategory	LysR-family proteins in Escherichia coli	Chromosome initiation inhibitor
Regulation and Cell signaling	no subcategory	cAMP signaling in bacteria	3',5'-cyclic-nucleotide phosphodiesterase (EC 3.1.4.17)
Regulation and Cell signaling	no subcategory	cAMP signaling in bacteria	ElaA protein
Respiration	no subcategory	Carbon monoxide dehydrogenase maturation factors	Aerobic carbon monoxide dehydrogenase molybdenum cofactor insertion protein CoxF
Respiration	no subcategory	Carbon monoxide dehydrogenase maturation factors	Carbon monoxide oxidation accessory protein CoxE
Stress Response	Cold shock	Cold shock, CspA family of proteins	Cold shock protein CspC
Stress Response	Osmotic stress	Synthesis of osmoregulated periplasmic glucans	Glucans biosynthesis protein C (EC 2.1.-.-)
Stress Response	no subcategory	Dimethylarginine metabolism	NG,NG-dimethylarginine dimethylaminohydrolase 1 (EC 3.5.3.18)
Sulfur Metabolism	no subcategory	Sulfur oxidation	Sulfur oxidation protein SoxZ
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Copper homeostasis	Cu(I)-responsive transcriptional regulator

REFERENCES

1. Sharp, J. H. *et al.* A biogeochemical view of estuarine eutrophication: Seasonal and spatial trends and correlations in the Delaware Estuary. *Estuaries and Coasts* **32,** 1023–1043 (2009).
2. Church, T. M. Biogeochemical factors influencing the residence time of microconstituents in a large tidal estuary, Delaware Bay. *Mar. Chem.* **18,** 393–406 (1986).
3. Mannino, A. & Harvey, H. R. Black carbon in estuarine and coastal ocean dissolved organic matter. *Limnol. Oceanogr.* **49,** 735–740 (2004).
4. Hoch, M. P. & Kirchman, D. L. Seasonal and Interannual Variability in Bacterial Production and Biomass in a Temperate Estuary. *Mar. Ecol. Ser.* **98,** 283–295 (1993).
5. Biggs, R. B., Sharp, J. H., Church, T. M. & Tramontano, J. M. Optical Properties, Suspended Sediments, and Chemistry Associated with the Turbidity Maxima of the Delaware Estuary. *Can. J. Fish. Aquat. Sci.* **40,** s172–s179 (1983).
6. Pennock, J. & Sharp, J. Phytoplankton production in the Delaware Estuary: temporal and spatial variability . *Mar. Ecol. Prog. Ser.* **34,** 143–155 (1986).
7. Pennock, J. R. & Sharp, J. H. Temporal alternation between light-limitation and nutrient-limitation of phytoplankton production in a coastal plain estuary. *Mar. Ecol. Prog. Ser.* **111,** 275–288 (1994).
8. Campbell, B. B. J. & Kirchman, D. L. D. D. L. Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *ISME J.* **7,** 210–220 (2012).
9. Kirchman, D. L., Dittel, A. I., Malmstrom, R. R. & Cottrell, M. T. Biogeography of major bacterial groups in the Delaware Estuary. *Limnol. Oceanogr.* **50,** 1697–1706 (2005).
10. Crump, B. C., Hopkinson, C. S., Sogin, M. L. & Hobbie, J. E. Microbial Biogeography along an Estuarine Salinity Gradient: Combined Influences of Bacterial Growth and Residence Time. *Appl. Environ. Microbiol.* **70,** 1494–1505

(2004).

11. Alonso, C. *et al.* Multilevel analysis of the bacterial diversity along the environmental gradient Río de la Plata–South Atlantic Ocean. *Aquat. Microb. Ecol.* **61,** 57–72 (2010).

12. Elifantz, H., Malmstrom, R. R., Cottrell, M. T. & Kirchman, D. L. Assimilation of polysaccharides and glucose by major bacterial groups in the Delaware Estuary. *Appl. Environ. Microbiol.* **71,** 7799–7805 (2005).

13. Walsh, D. A., Papke, R. T. & Doolittle, W. F. Archaeal diversity along a soil salinity gradient prone to disturbance. *Environ. Microbiol.* **7,** 1655–1666 (2005).

14. Hansell, D. a & Carlson, C. a. Biogeochemistry of marine dissolved organic matter. *Biogeochemistry of marine dissolved organic matter* 774 (2002). doi:10.1016/B978-012323841-2/50014-2

15. Middelburg, J. J. & Herman, P. M. J. Organic matter processing in tidal estuaries. *Mar. Chem.* **106,** 127–147 (2007).

16. Simon, M., Grossart, H. P., Schweitzer, B. & Ploug, H. Microbial ecology of organic aggregates in aquatic ecosystems. *Aquatic Microbial Ecology* **28,** 175–211 (2002).

17. Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nature Reviews Microbiology* **5,** 782–791 (2007).

18. Crump, B. C. & Baross, J. A. Characterization of the bacterially-active particle fraction in the Columbia River estuary. *Mar. Ecol. Prog. Ser.* **206,** 13–22 (2000).

19. Ghiglione, J. F. *et al.* Diel and seasonal variations in abundance, activity, and community structure of particle-attached and free-living bacteria in NW Mediterranean Sea. *Microb. Ecol.* **54,** 217–231 (2007).

20. Rösel, S. & Grossart, H. P. Contrasting dynamics in activity and community composition of free-living and particle-associated bacteria in spring. *Aquat. Microb. Ecol.* **66,** 169–181 (2012).

21. Rieck, A., Herlemann, D. P. R., Jürgens, K. & Grossart, H. P. Particle-associated differ from free-living bacteria in surface waters of the baltic sea. *Front.*

Microbiol. **6,** (2015).

22. Buchan, A., González, J. M. & Moran, M. A. Overview of the marine Roseobacter lineage. *Applied and Environmental Microbiology* **71,** 5665–5677 (2005).

23. Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *ISME J.* **4,** 784–798 (2010).

24. Rappé, M. S. & Giovannoni, S. J. The Uncultured Microbial Majority. *Annu. Rev. Microbiol.* **57,** 369–394 (2003).

25. González, J. M. & Moran, M. A. Numerical dominance of a group of marine bacteria in the α-subclass of the class Proteobacteria in coastal seawater. *Appl. Environ. Microbiol.* **63,** 4237–4242 (1997).

26. DeLong, E. F. Microbial community genomics in the ocean. *Nature Reviews Microbiology* **3,** 459–469 (2005).

27. González, J. M. *et al.* Silicibacter pomeroyi sp. nov. and Roseovarius nubinhibens sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *Int. J. Syst. Evol. Microbiol.* **53,** 1261–1269 (2003).

28. Miller, T. R. & Belas, R. Dimethylsulfoniopropionate Metabolism by Pfiesteria - Associated Roseobacter spp. *Appl. Environ. Microbiol.* **70,** 3383–3391 (2004).

29. Moran, M. A., González, J. M. & Kiene, R. P. Linking a bacterial taxon to sulfur cycling in the sea: Studies of the marine Roseobacter group. *Geomicrobiol. J.* **20,** 375–388 (2003).

30. Holmes, A. J. *et al.* Methylsulfonomonas methylovora gen. nov., sp. nov., and Marinosulfonomonas methylotropha gen. nov., sp. nov.: Novel methylotrophs able to grow on methanesulfonic acid. *Arch. Microbiol.* **167,** 46–53 (1997).

31. Moran, M. A. *et al.* Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment. *Nature* **432,** 910–913 (2004).

32. King, G. M. Molecular and Culture-Based Analyses of Aerobic Carbon Monoxide Oxidizer Diversity. *Appl. Environ. Microbiol.* **69,** 7257–7265 (2003).

33. Buchan, A., Collier, L. S., Neidle, E. L. & Moran, M. A. Key aromatic-ring-cleaving enzyme, protocatechuate 3,4-dioxygenase, in the ecologically important

marine Roseobacter lineage. *Appl. Environ. Microbiol.* **66,** 4662–4672 (2000).

34. Hobbie, J. E., Daley, R. J. & Jasper, S. Use of nuclepore filters for counting bacteria by fluorescence microscopy. *Appl. Environ. Microbiol.* **33,** 1225–1228 (1977).

35. Zimmerman, R. & Meyer-Reil., L.-A. A new method for fluorescence staining of bac- terial populations on membrane filters. *Kieler Meeresforsch* **30,** 24–27 (1974).

36. Porter, K. G. & Feig, Y. S. The use of DAPI for identifying and counting aquatic microflora. *Limnology and Oceanography* **25,** 943–948 (1980).

37. Staley, J. Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annu. Rev. Microbiol.* **39,** 321–346 (1985).

38. Kirchman, D., K'nees, E. & Hodson, R. Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. *Appl. Environ. Microbiol.* **49,** 599–607 (1985).

39. Cottrell, M. T. & David, K. L. Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. *Limnol. Oceanogr.* **48,** 168–178 (2003).

40. Kirchman, D. L. *Microbial Ecology of the Oceans: Second Edition*. *Microbial Ecology of the Oceans: Second Edition* (2008). doi:10.1002/9780470281840

41. Kirchman, D. L. Measuring bacterial biomass production and growth rates from leucine incorporation in natural aquatic environments. In , Methods in microbiology. *Methods Microbiol.* **30,** 227–237 (2001).

42. Pollard, P. C. & Moriarty, D. J. W. Validity of the tritiated thymidine method for estimating bacterial growth rates: Measurement of isotope dilution during DNA synthesis. *Appl. Environ. Microbiol.* **48,** 1076–1083 (1984).

43. Alonso, C. & Pernthaler, J. Concentration-dependent patterns of leucine incorporation by coastal picoplankton. *Appl. Environ. Microbiol.* **72,** 2141–2147 (2006).

44. Campbell, B. J., Yu, L., Straza, T. R. A. & Kirchman, D. L. Temporal changes in

bacterial rRNA and rRNA genes in Delaware (USA) coastal waters. *Aquat. Microb. Ecol.* **57,** 123–135 (2009).

45. DeLong, E., Wickham, G. & Pace, N. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science (80-.).* **243,** 1360–1363 (1989).

46. Kemp, P. F., Lee, S. & LaRoche, J. Estimating the growth rate of slowly growing marine bacteria from RNA content. *Appl. Environ. Microbiol.* **59,** 2594–2601 (1993).

47. Fegatella, F., Lim, J., Kjelleberg, S. & Cavicchioli, R. Implications of rRNA operon copy number and ribosome content in the marine oligotrophic ultramicrobacterium Sphingomonas sp. strain RB2256. *Appl. Environ. Microbiol.* **64,** 4433–4438 (1998).

48. Deutscher, M. P. Degradation of RNA in bacteria: Comparison of mRNA and stable RNA. *Nucleic Acids Res.* **34,** 659–666 (2006).

49. Campbell, B. J., Yu, L., Heidelberg, J. F. & Kirchman, D. L. Activity of abundant and rare bacteria in a coastal ocean. *Proc. Natl. Acad. Sci.* **108,** 12776–12781 (2011).

50. Hunt, D. E. *et al.* Relationship between abundance and specific activity of bacterioplankton in open ocean surface waters. *Appl. Environ. Microbiol.* **79,** 177–184 (2013).

51. Blazewicz, S. J., Barnard, R. L., Daly, R. A. & Firestone, M. K. Evaluating rRNA as an indicator of microbial activity in environmental communities: Limitations and uses. *ISME Journal* **7,** 2061–2068 (2013).

52. Binder, B. J. & Liu, Y. C. Growth rate regulation of rRNA content of a marine Synechococcus (cyanobacterium) strain. *Appl. Environ. Microbiol.* **64,** 3346–3351 (1998).

53. Venter, J. C. *et al.* Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science (80-.).* **304,** 66–74 (2004).

54. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome.

Science (80-.). **348,** (2015).

55. Meng, J. *et al.* Genetic and functional properties of uncultivated MCG archaea assessed by metagenome and gene expression analyses. *ISME J.* **8,** 650–659 (2014).
56. Helton, R. R. & Wommack, K. E. Seasonal dynamics and metagenomic characterization of estuarine viriobenthos assemblages by randomly amplified polymorphic DNA PCR. *Appl. Environ. Microbiol.* **75,** 2259–2265 (2009).
57. Cai, L., Zhang, R., He, Y., Feng, X. & Jiao, N. Metagenomic analysis of Virioplankton of the subtropical Jiulong river estuary, China. *Viruses* **8,** (2016).
58. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: Genomic Analysis of Microbial Communities. *Annu. Rev. Genet.* **38,** 525–552 (2004).
59. Kembel, S. W., Eisen, J. A., Pollard, K. S. & Green, J. L. The phylogenetic diversity of metagenomes. *PLoS One* **6,** (2011).
60. Hewson, I., Poretsky, R. S., Tripp, H. J., Montoya, J. P. & Zehr, J. P. Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ. Microbiol.* **12,** 1940–1956 (2010).
61. Baker, B. J. *et al.* Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *ISME J.* **7,** 1962–1973 (2013).
62. de Menezes, A., Clipson, N. & Doyle, E. Comparative metatranscriptomics reveals widespread community responses during phenanthrene degradation in soil. *Environ. Microbiol.* **14,** 2577–2588 (2012).
63. Wemheuer, B. *et al.* The green impact: Bacterioplankton response towards a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches. *Front. Microbiol.* **6,** (2015).
64. Dupont, C. L. *et al.* Functional tradeoffs underpin salinity-driven divergence in microbial community composition. *PLoS One* **9,** (2014).
65. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31,** 533–

538 (2013).

66. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2,** e603 (2014).

67. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33,** 1045–1052 (2015).

68. Herlemann, D. P. R. *et al.* Metagenomic De Novo assembly of an aquatic representative of the verrucomicrobial class Spartobacteria. *MBio* **4,** (2013).

69. Baker, B. J., Lazar, C. S., Teske, A. P. & Dick, G. J. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome* **3,** 14 (2015).

70. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6,** 1186–1199 (2012).

71. Baker, B. J., Lesniewski, R. A. & Dick, G. J. Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J.* **6,** 2269–2279 (2012).

72. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63 (2009).

73. Creecy, J. P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Current Opinion in Microbiology* **23,** 133–140 (2015).

74. Harke, M. J. & Gobler, C. J. Global Transcriptional Responses of the Toxic Cyanobacterium, Microcystis aeruginosa, to Nitrogen Stress, Phosphorus Stress, and Growth on Organic Matter. *PLoS One* **8,** (2013).

75. Pinto, A. C. *et al.* Differential transcriptional profile of Corynebacterium pseudotuberculosis in response to abiotic stresses. *BMC Genomics* **15,** 14 (2014).

76. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nature Reviews Genetics* **12,** 671–682 (2011).

77. Giebel, H. A. *et al.* Planktomarina temperata gen. nov., sp. nov., belonging to the globally distributed RCA cluster of the marine Roseobacter clade, isolated from the German Wadden Sea. *Int. J. Syst. Evol. Microbiol.* **63,** 4207–4217 (2013).

78. Voget, S. *et al.* Adaptation of an abundant Roseobacter RCA organism to pelagic systems revealed by genomic and transcriptomic analyses. *ISME J.* **9,** 371–384 (2015).

79. Cottrell, M. T., Mannino, A. & Kirchman, D. L. Aerobic anoxygenic phototrophic bacteria in the mid-atlantic bight and the north pacific gyre. *Appl. Environ. Microbiol.* **72,** 557–564 (2006).

80. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).

81. Aziz, R. K. *et al.* The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* **9,** (2008).

82. Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42,** (2014).

83. Brettin, T. *et al.* RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5,** (2015).

84. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* **428,** 726–731 (2016).

85. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9,** R151 (2008).

86. Seah, B. Phylogenomics tools Online: https://github.com/kbseah/phylogenomics-tools. *DOI 10.5281/zenodo.46122* 1 (2014).

87. Rodriguez-R, L. M. & Konstantinidis, K. T. Bypassing Cultivation To Identify Bacterial Species. *Microbe Mag.* **9,** 111–118 (2014).

88. Kanehisa, M. & Goto, S. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

89. Galili Tal, O'Callaghan Alan, Sidi Jonathan, S. C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing.

https://doi.org/10.1093/bioinformatics/btx657 doi:https://doi.org/10.1093/bioinformatics/btx657

90. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

91. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

92. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34,** 1256–1263 (2016).

93. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general-purpose read summarization program. *arXiv.org* **1305,** 3347 (2013).

94. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** (2014).

95. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** (2010).

96. Durham, B. P. *et al.* Draft genome sequence of marine alphaproteobacterial strain HIMB11, the first cultivated representative of a unique lineage within the Roseobacter clade possessing an unusually small genome. *Stand. Genomic Sci.* **9,** 632–645 (2015).

97. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55,** 539–552 (2006).

98. Henriques, I. S., Almeida, A., Cunha, Â. & Correia, A. Molecular sequence analysis of prokaryotic diversity in the middle and outer sections of the Portuguese estuary Ria de Aveiro. *FEMS Microbiol. Ecol.* **49,** 269–279 (2004).

99. Gonzalez, J. M. *et al.* Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl. Environ. Microbiol.* **66,** 4237–4246 (2000).

100. Koblížek, M. *et al.* Isolation and characterization of Erythrobacter sp. strains from the upper ocean. *Arch. Microbiol.* **180,** 327–338 (2003).

101. Tang, K. H., Feng, X., Tang, Y. J. & Blankenship, R. E. Carbohydrate metabolism and carbon fixation in Roseobacter denitrificans OCh114. *PLoS One* **4,** (2009).