Public health applications of gene-environment interactions for Oceans and Human Health

Jesse A. Port

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2012

Reading Committee:
Prof. Elaine M. Faustman, Chair
Prof. E. Virginia Armburst
Prof. Timothy Rose

Program Authorized to Offer Degree:
Environmental and Occupational Health Sciences
School of Public Health

University of Washington

Abstract

Public health applications of gene-environment interactions for
Oceans and Human Health

Jesse A. Port

Chair of Supervisory Committee:
Professor Elaine M. Faustman
Department of Environmental and Occupational Health Sciences

The field of Oceans and Human Health (OHH) is an emerging discipline that requires novel, interdisciplinary approaches to address the ecological and public health consequences of our changing oceans. As population growth and development continue to increase in coastal zones, there is an urgent need to reduce anthropogenic impacts on marine ecosystems. This research is framed around the premise that the increasing availability of environmental genomic sequence data in tandem with the advancing bioinformatics now offers high-throughput and systems-based approaches and opportunities for environmental health monitoring. Specifically, we hypothesize that environmental genomic information can provide sensitive and functional markers of human impacts on marine ecosystems which can then be used to improve our understanding of how the composition of micro-organism communities relates to public health. We utilize a gene-environment approach whereby the complex interplay between genetic and environmental factors in marine ecosystems can provide insight into potential OHH concerns including chemical and pharmaceutical pollution. Three case studies are

presented that incorporate comparative genomic, metagenomic and decision-analysis methodologies. First, a comparative genomic approach is used to investigate the G protein-coupled receptor signaling pathway in marine diatoms. This pathway plays an important role in the genomic underpinnings of mammalian environmental response, and thus its presence and potential functionality in environmental perception and response to stressors in these crucial primary producers is of interest. Secondly, metagenomics in combination with next generation sequencing is used to profile the microbial composition and antibiotic resistance determinant signals across differentially impacted environments, including marine, nearshore and wastewater ecosystems. The environment serves as a reservoir for resistance genes that can be disseminated to pathogenic bacteria in clinical settings, but baseline data is needed for the prevalence and distribution of these genes and the genomic vectors that transfer these genes across environmental bacteria. Thirdly, this metagenomic data is conceptually framed within a public health screening framework by incorporating the data into an index for surveillance of antibiotic resistance determinants in the environment. Metagenomic screening of genomic markers relevant to public health may serve as an early risk management approach that can trigger further monitoring and analysis.

TABLE OF CONTENTS

List of Figures ii

List of Tables iii

List of Abbreviations iv

Acknowledgements v

Chapter 1: Background and Significance 1

Chapter 2: Identification of G protein-coupled receptor signaling pathway proteins in marine diatoms using comparative genomics 23

Chapter 3: Metagenomic profiling of microbial composition and antibiotic resistance determinants in Puget Sound 54

Chapter 4: Incorporating metagenomics into Oceans and Human Health decision-making: Considerations for antibiotic resistance surveillance 97

Chapter 5: Significance of research findings and implications for future Oceans and Human Health GxE investigations 131

References 134

LIST OF FIGURES

Page Number

Figure 1.1 Global map of cumulative human impacts on marine ecosystems 16
Figure 1.2 Gene-environment in OHH framework 17
Figure 1.3 OHH risk chain linking toxic algae to human exposure and disease 18
Figure 1.4 Bioinformatic pipeline and databases for OHH GxE investigations 19
Figure 1.5 Fate and transport of antibiotic resistance determinants in the environment 20

Figure 2.1 Data analysis framework for GPCR pathway in diatoms 44
Figure 2.2 Conservation of GPCR signaling proteins in diatoms 45
Figure 2.3 Sequence alignment of putative diatom GPCRs 47
Figure 2.4 Structure and size of putative diatom GPCRs 48
Figure 2.5 Phylogenetic tree of putative diatom GPCRs and known class C GPCRs 49

Figure 3.1 Locations of sampling sites in Puget Sound 76
Figure 3.2 Construction of the expanded Antibiotic Resistance Genes Database 77
Figure 3.3 Abundance of major taxa in Puget Sound, WWTP and other metagenomes 78
Figure 3.4 Over and under representation of major taxa in Puget Sound 79
Figure 3.5 Abundance of major taxa in Puget Sound and WWTP samples at the order level 80
Figure 3.6 Recruitment plot of reads to alpha-Proteobacterium HTCC2255 81
Figure 3.7 Relationship between taxa abundance and salinity and temperature 82
Figure 3.8 Rarefaction analysis for Puget Sound and WWTP metagenomes 83
Figure 3.9 Prevalence of antibiotic resistance genes in the Puget Sound metagenomes 84
Figure 3.10 Abundance of antibiotic resistance genes in environmental metagenomes 85
Figure 3.11 Prevalence of mobile genetic elements in the Puget Sound metagenomes 86

Figure 4.1 Bioinformatic framework for quantification of the ARD index 121
Figure 4.2 Metagenomic index of antibiotic resistance determinants (ARDs) 122
Figure 4.3 Proportion of index-positive sequences per metagenome 123
Figure 4.4 Proportion of index-positive sequences by index sub-category per metagenome 124
Figure 4.5 Principal component analysis of index footprints for the metagenomic samples 125
Figure 4.6 Public health management decisions associated with use of the ARD index 126
Figure 4.7 Methods for screening of public health signals in environmental samples 127

LIST OF TABLES

Page Number

Table 1.1 Next generation sequencing platforms 21
Table 1.2 Bioinformatic databases for comparative metagenomic analyses 22

Table 2.1 Description of the functions of GPCR signaling pathway proteins 50
Table 2.2 Organisms included in the custom microeukaryote database 51
Table 2.3 EST data for the putative diatom GPCRs 52
Table 2.4 Best protein sequence hits within the diatoms to human GPCR signaling proteins 53

Table 3.1 Summary statistics for the Puget Sound metagenomes 87
Table 3.2 Puget Sound sequences with similarity to antibiotic resistance genes 88
Table 3.3 NCBI plasmid sequences that match Puget Sound and WWTP contigs 89
Table 3.4 NCBI plasmid sequences that match Puget Sound and WWTP reads 90

Table 4.1 Metagenomic samples used in this study 128
Table 4.2 Sequence similarity criteria for the varying stringency classification levels 129
Table 4.3 Pearson's correlation coefficients (r) for the index sub-categories 130

LIST OF ABBREVIATIONS

ARD, antibiotic resistance determinant
ARDB, Antibiotic Resistance Genes Database
ARG, antibiotic resistance gene
BLAST, Basic Local Alignment Search Tool
DA, domoic acid
EST, expressed sequence tag
GxE, gene-environment
$GABA_B$, γ-aminobutyric acid receptor
GOS, Global Ocean Sampling Expedition
GPCR, G protein-coupled receptor
GPCRDB, G protein-coupled receptor database
HMM, hidden markov model
JGI, Joint Genome Institute
MRG, metal resistance gene
NCBI, National Center for Biotechnology Information
OHH, oceans and human health
PBP, periplasmic binding protein
PCA, principal component analysis
POP, persistent organic pollutant
PCR, polymerase chain reaction
RDP, Ribosomal Database Project
TE, transposable element
TMD, transmembrane domain
VOI, value of information
WWTP, wastewater treatment plant

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Prof. Elaine Faustman, for her guidance through the dissertation process and her flexibility in allowing me to pursue innovative and interdisciplinary research directions. I greatly appreciate your open mind and dynamic approach. I am truly indebted to Jim Wallace for his assistance with bioinformatics analyses and research frameworks, and also to Bill Griffith for statistical support. And thanks to the rest of the Institute for Risk Analysis and Risk Communication, including Alison Laing, Alison Scherer, Marissa Smith and Eric Vigoren. I would also like to recognize the Pacific Northwest Center for Human Health and Ocean Studies and our collaborators in the School of Oceanography. This Oceans and Human Health-based project would not have been possible without the assistance of the Armbrust lab. I also must thank Alison Cullen for her time and guidance in helping me apply my work within a decision-analysis context. Thanks to Stephanie Moore for serving as my mentor as part of my NOAA traineeship, and Roger Bumgarner and Rob Hall for sequencing efforts. And thanks to Ginger Armbrust, Tim Rose, Scott Meschke and Stephanie Moore for serving as members of my PhD committee. Lastly, I would like to thank the NOAA Oceans and Human Health training grant, NSF and NIEHS, which have funded this work.

CHAPTER 1: Background and Significance

1.1 Oceans and Human Health

The linkages between the oceans and human health are obvious and ancient, and in some cultures they are deeply intertwined. For example, the term 'Sato-umi' in Japanese translates to marine ecosystem health and human well-being. Only recently though have scientific, governmental and policy communities in the United States formalized these interactions into the field of Oceans and Human Health (OHH). The National Academy of Sciences seminal report "From Monsoons to Microbes: Understanding the Ocean's Role in Human Health" (NRC 1999) provided a foundational framework for understanding how marine processes and systems can both increase and reduce public health risks. This report defined OHH by five categories: Climate and weather impacts, waterborne infectious diseases, harmful algal blooms (HABs), marine-derived pharmaceuticals and aquatic organism models for biomedical research. A volume of OHH articles in *Environmental Health* was recently published (Laws 2008) that now expands this list to also include chemical pollutants, food security, marine sentinel species and biotechnological applications. Governmental agencies including the National Science Foundation (NSF), National Institute of Environmental Health Sciences (NIEHS) and National Oceanic and Atmospheric Administration (NOAA) have provided funding for OHH research through the establishment of OHH Centers at various branches and universities throughout the United States.

Much of OHH research pertains to the ecological and human health consequences of anthropogenic impacts and pressures in the coastal zone. As population growth and development continue to increase along the world's coasts, there is an urgent need to reduce these footprints. Humans have already had a profound adverse impact on global marine ecosystems, with every

square kilometer of the ocean affected by at least one of seventeen anthropogenic drivers (Halpern et al. 2008) (Figure 1.1). Coastal areas experienced the largest burden due to these drivers, which included organic and inorganic pollution, nutrient input, overfishing, shipping, invasive species and various climate change-associated impacts such as ocean acidification and sea temperature. Further cumulative assessments have shown that ocean health is negatively correlated with coastal human population (Halpern et al. 2012). As ecosystem and human health are inextricably intertwined, these impacts can feedback to pose public health concerns including contaminated seafood, harmful algal blooms, microbial and viral pathogens and decreased well-being (Kite –Powell et al. 2008; Fleming et al. 2006). Despite these obvious OHH risks, in many cases it has been challenging to effectively link exposure to disease. Epidemiological evidence for diseases associated with marine-borne exposures is limited, and especially so for diseases associated with chronic exposures to low levels of marine contaminants or toxins, such as cancer or reproductive/developmental toxicity due to consumption of seafood contaminated with organic pollutants or heavy metals, or neurodegeneration associated with seafood or drinking water supplies contaminated by harmful algal bloom (HAB) toxins (Kite-Powell et a. 2008). Consequently, there is a need for metrics and assessments that better define exposure in OHH investigations. Human dependency on and value of ocean resources continue to increase, and without adequate metrics and integrated assessments the exposure side of the exposure-disease relationship is under evaluated. As opposed to metrics such as those used by Halpern et al. (Halpern et al. 2008) that are largely based on data that is available, OHH applications would benefit from data that is derived from biologically-based and public health relevant questions. One integrating concept for evaluating exposure in and formulating metrics for OHH is to use

gene-environment (GxE) relationships. The purpose of our Center is to examine these interactions as one framing concept.

1.2 Gene-environment interactions in OHH investigations

We hypothesize that the complex interplay between genetic and environmental factors in marine ecosystems provides a conceptual framework for defining OHH by linking ocean and coastal processes to human health. The concept of GxE has been defined predominantly within the context of human disease, and describes how genetic and environmental factors jointly influence the risk of developing a disease (Hunter 2005). To implement GxE in our OHH investigations, we utilize the longstanding biomarker paradigm from the National Academy of Sciences (NAS 1994). A biomarker is broadly defined as an indicator of cellular or biochemical functioning or response that is measurable in a biological system. While traditional biomarkers of human exposure to xenobiotics are based on chemicals, metabolites, gene expression or genetic polymorphisms, this concept can be extended to include changes in gene abundance or expression in marine micro-organisms resulting from exposure to anthropogenic pressures. Figure 1.2 conceptually shows how GxE interactions may mediate an interrelated relationship between humans and marine ecosystems, whereby the genetics of both marine organisms and humans are interrogated against a backdrop of natural and human-induced environmental variation. On one side of the relationship, human impacts on marine organisms can be evaluated using an exposure, response and impact paradigm. Such a "risk chain" provides insight into marine ecosystem health and thus possible human exposures (e.g. to a chemical contaminant, pathogen, transgenic organism, genes, toxins) that are mediated by anthropogenic pressures on the genomic pools of microbial and phytoplankton communities. The types of genetic parameters

considered across the risk chain mirror one another regardless of the organism. These parameters include gene and gene expression profiles and genetic polymorphisms. There is also a shared environmental component, including factors such as temperature, nutrients (or diet), age, sex, behavior, community composition and proximity to exposure source. For example, the frequency and toxicity of HABs has been associated with changes in ocean temperature and eutrophication events (Heisler et al. 2008; Moore et al. 2008). Increases in global temperature and coastal eutrophication may be related to human-derived climate change and land-use impacts respectively. In the case of the toxic diatom *Pseudo-nitzschia*, these impacts can mediate community assemblages and dynamics (Figure 1.3). As different species and even strains of *Pseudo-nitzschia* produce different levels of the neurotoxin domoic acid (DA) (EV Armbrust, unpublished), community composition is important in determining potential human exposure risk through contaminated seafood and can therefore be thought of as a biomarker of potential exposure risk. Furthermore, DA production can be influenced by nutrient ratios associated with eutrophication. The risk chain can be extended further to evaluate human health risk post-exposure (Figure 1.3). Children, adults and certain ethnic groups may have different levels of susceptibility to DA based on genomic composition or response or dietary behavior. This leads to alterations in the dose-response curves and ultimately differential toxicity and disease. This example shows how human exposure and disease outcomes related to DA toxicity are dependent on both genetic and environmental variation influencing DA production by *Pseudo-nitzschia*.

1.3 Gene: Utility of high-throughput meta-omics for OHH GxE investigations

The vast amount of environmental genomic information currently being generated is paving the way for new approaches to environmental health investigations. Microbial, viral and

microeukaryotic communities in particular have been the focus of large-scale sequencing projects investigating community composition and functionality in marine ecosystems. The genomic repertoires of these communities are extremely important given the fact that the oceans cover 70% of the planet and that 1 ml of seawater contains 10^3-10^6 microbes (Whitman et al. 1998) and 10^7 viruses (Breitbart 2012) and that organisms such as diatoms play crucial roles in primary production and nutrient cycling and serve as a base for marine food webs (Armbrust 2009).

While significant progress has been made in generating fully sequenced genomes and gene expression libraries for model organisms, sentinel organisms and toxin producers relevant to OHH, meta-omic investigations may ultimately provide the most insight into GxE in marine micro-organism communities. Metagenomics, metatranscriptomics and metaproteomics involve the extraction of DNA, RNA or proteins directly from mixed environmental micro-organism communities, and as such allows for the study of the genomes, transcriptomes and proteomes of many organisms simultaneously (NRC 2007). As a result, a more systems-based approach focusing on the aggregate instead of the individual level can be taken to study the collective composition and activities of micro-organism communities. Meta-omics also overcomes the major limitation of unculturability that currently exists for >99% of microbes (Amann et al. 1995). These culture-independent approaches have already vastly increased our knowledge of bacterial diversity and functioning. Because of their smaller genome size, microbes and viruses have been the main focus of meta-omic sequencing projects up to this point. Of these -omic approaches, metagenomics has been most frequently used in marine microbial investigations. Metagenomic research attempts to answer the fundamental biological and ecological questions of 'who is there?', 'what are they doing?', 'who is doing what?', and 'what evolutionary processes

determine these parameters?' (Kennedy et al. 2010). Metagenomic data is available for the microbes inhabiting many different ocean niches, including surface waters, water column, benthos, open ocean, coastal ocean, hydrothermal vents, among others. The sequence data from these projects has greatly expanded genetic diversity and provided insight into how, at the genome level, a community can be shaped by its environment.

Our ability to process and analyze metagenomic and metatranscriptomic data has been made possible by the evolution of next generation sequencing technology and bioinformatics. The traditional and "gold standard" approach to whole genome sequencing has been the Sanger method pioneered in 1977. Sanger sequencing utilizes fluorescently labeled dideoxyribonucleotide triphosphates (ddNTPs) to halt DNA strand elongation (Pettersson et al. 2009). The DNA fragments can then be separated by size using electrophoresis and the order of bases can be determined. This method produces highly accurate sequences currently ranging from 700-1,000bp. The downsides to Sanger sequencing are time and cost. For example, a mammalian-sized genome is estimated to cost US$10-25 million and would take several years (Rogers and Venter 2005). A microbial genome is on the order of $20-50,000. Traditional sequencing approaches also rely on a bacterial cloning step to amplify DNA fragments before sequencing. There are a number of next generation sequencing platforms currently available, including the Roche/454 Life Sciences GS FLX, Illumina/Solexa Genome Analyzer, Applied Biosystems SOLiD System and Life Technologies Ion Torrent (Table 1.1). These platforms can generate hundreds of thousands to tens of millions of sequencing reads in the order of hours to days (Shokralla et al. 2012). These machines rely on either emulsion polymerase chain reaction (PCR) or solid-phase amplification to amplify DNA and thus there is no need for a bacterial cloning step (Metzker 2010). Sequencing of environmental metagenomic data has exploded due

to the continued advancement of these next generation sequencing platforms, and 454 technology (a.k.a pyrosequencing) has been at the forefront for OHH applications. One of the main advantages of pyrosequencing over the other next generation sequencing platforms is the considerably longer read length (400-800 bp) (Table 1.1). Longer read lengths increase our ability to assign taxonomic and functional information to these sequences and to assemble these reads into contigs and ultimately genomes. This is extremely important for environmental metagenomic data where reference genomes are not available for the majority of species and *de novo* assembly is necessary (Zhou et al. 2010). The difficulty in assembling metagenomic data from complex microbial communities has thus led to analyses based on the unassembled sequence fraction. That said, the vastly increased sequencing depth provided by the Illumina platform (~500-600 Gb vs. ~500-700 Mb for 454) is gaining traction in environmental studies. The SOLiD may be useful for studies investigating potential genetic polymorphisms such as single nucleotide polymorphisms (SNPs) that may differentiate toxic versus non-toxic strains of bacteria or phytoplankton. In general, the choice of sequencing platform is highly dependent on the research question. There are methodological limitations to next generation sequencers that should be considered such as recognition of homopolymer regions by 454 and amplification bias due to the PCR step.

1.4 Environment: A context for genomic variation identified through meta-omics

To enable this wealth of environmental genomic information to be applied to GxE investigations in marine ecosystems, the environmental context for genomic sampling must also be captured concomitantly. As OHH spans the land-sea interface, GxE information is needed from a multitude of environments from source to sink. These include rivers, estuaries, coastal

zones, beaches, surface waters, wastewater treatment plants and outfalls and aquaculture operations, among others. To standardize and coordinate the description of metadata collected from various water environments and sampling efforts, guidelines such as the minimum information about a genomic sequence (MIGS) specification (Field et al. 2008) are being developed. These specifications include geographic location, time, depth, temperature, salinity, tidal cycle, chlorophyll, dissolved oxygen and measurements of various organic and inorganic compounds. Environmental data specific to OHH applications may include distance from land, a wastewater treatment plant outfall, port, aquaculture activity, shellfishery or freshwater input, and field measurements for contaminants such as POPs, metals, pesticides or pharmaceuticals. Having robust environmental data for genomic and metagenomic samples will allow for comparative GxE analyses across samples and will elucidate whether certain environmental conditions or factors are associated with increased gene abundance or expression in micro-organism communities. This information in turn can inform the GxE risk chain by determining baseline environmental levels of public health relevant signals such as pathogens, HAB producing species, antibiotic and metal resistance genes and detoxification enzymes and their relevance for potential human exposures.

Global information systems (GIS), principal component analysis and other spatial analysis tools have allowed investigators to begin to draw correlations between genomic and environmental components in large metagenomic datasets. These analyses have shown that there are environmentally-dependent molecular pathways and species composition profiles that can be used to define footprints of distinct environments (Gianoulis et al. 2009; Parks et al. 2009; Patel et al. 2010). This information provides valuable reference or baseline data for future GxE assessments. For example, the Global Ocean Sampling (GOS) Expedition led by the J. Craig

Venter Institute has provided a large repository of environmental sequence data for micro-organism communities in the oceans and has been probed for numerous ecological and biological applications. The GOS project has sampled at over 51 stations located at approximately 200 mile intervals along the eastern North American coast through the Gulf of Mexico and into the equatorial Pacific (Rusch et al. 2007). While this dataset is thus spatially extensive, it provides limited temporal information on microbial community dynamics as each station is sampled only once, leading to uncertainty in the variability in microbial community composition and dynamics. To make this dataset or other marine sampling efforts such as the U.S. Integrated Ocean Observing System (IOOS) more relevant to OHH applications, more temporal and nearshore sampling where human and ocean interactions are greatest is required. Associations between human impacts and microbial composition or potential functionality over time and space can then be better addressed and linked to public health.

1.5 Databases for OHH GxE investigations

Integration of multiple data platforms is required for investigated GxE interactions pertaining to OHH. Fortunately, there are already a number of publicly available online databases that allow for integrated GxE analysis in marine ecosystems (Table 1.2). Metagenomic and metatranscriptomic sequence data and associated metadata are deposited and available to the scientific community in a number online databases which include: Meta Genome Rapid Annotation using Subsystem Technology (MG-RAST) (Meyer et al. 2008), Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (Sun et al. 2011), MEtaGenome Analyzer (MEGAN) (Huson et al. 2007), the Integrated Microbial Genomes system (IMG/M) (Markowitz et al. 2008) run by the Department of Energy Joint

Genome Institute and Metagenomic Reports (METAREP) through the J. Craig Venter Institute (Goll et al. 2010). These databases provide analysis tools whereby metagenomic data can be analyzed and linked to associated metadata. Taxonomic and functional analysis of selected metagenomic data can be performed using user-inputted bioinformatic criteria and tools such as heat maps and principal component analysis can then be used for comparative GxE metagenomics. While these databases are broad-based in application and are primarily used by microbial ecologists, oceanographers and environmental scientists, there is obvious potential for OHH specific workflows to be included (Figure 1.4). For example, MG-Rast currently links to the Ribosomal Database Project (RDP) to classify 16s rRNA sequences, but the 16s rRNA data can then be linked to a microbial pathogen database such as the Microbial Rosetta Stone database (Ecker et al. 2005) to identify potential pathogens. Pathogen abundances are then compared across various environments to determine if GxE interactions may be present. Environmental antibiotic resistance determinants can also be probed by cross-referencing the Antibiotic Resistance Genes Database (ARDB) (Liu and Pop 2009) and mobile genetic elements databases (e.g. NCBI plasmid and transposable element databases). The abundances of antibiotic resistance genes or mobile genetic elements can again be compared across various environmental parameters. The Pfam database (Punta et al. 2012) also provides for a high-throughput survey of protein motifs relevant to genomic composition and function. Ultimately the abundance of these genomic markers of public health importance can be linked to associated environmental metadata and then across samples to determine if GxE relationships exist.

1.6 OHH gene-environment applications

This dissertation will present three applications of gene-environment interactions relevant to OHH, including GPCR signaling in diatoms, antibiotic resistance determinant profiling in marine microbial communities and generation of a metagenomic index for antibiotic resistance determinants in environmental samples that is conceptually framed within a public health surveillance and decision analysis framework.

1.6.1. GPCR signaling in diatoms

Targeted genome sequencing of individual species important to the OHH framework and subsequent comparative cross-species genomic studies allow for initial exploration into the conservation of genes, proteins or molecular pathways across diverse taxa. Seventeen highly conserved intracellular signaling pathways have been identified and characterized across vertebrates and non-vertebrates (NAS 2000). One of these pathways, G-protein coupled receptors (GPCR) signaling, is conserved across all metazoa and some single-celled eukaryotes. GPCR signaling is a crucial pathway responsible for regulating numerous cell functions including neurotransmission, differentiation and growth, inflammatory responses, smell, taste and vision in mammalians (Qian et al. 2003). GPCRs are cell-surface receptors that function in signal transduction of messages such as neurotransmitters, calcium, odorants, proteins and other small molecules and they control the activity of enzymes and transport of vesicles (Bockaert and Pin 1999). Thus GPCRs can be thought of as essential nodes of communication between the internal and external cellular environments (Rosenbaum et al. 2009), and as such play a major role in the perception of and response to the environment. It is unknown whether this pathway is present to any extent in unicellular photosynthetic algae, but if so would shed insight into environmental response to stressors of potential relevance to human health. Diatoms are a major

class of eukaryotic phytoplankton found throughout the world's oceans which play a crucial role in primary production and nutrient cycling, serve as a basis for marine food webs and form large blooms that in some cases can be toxic. Four diatom genome projects and associated expressed sequence tag (EST) libraries have been completed. This sequence data now available for the diatoms allows for comparative and cross-species analyses relevant to signal transduction and cellular communication. If functionally conserved in diatoms, GPCRs may serve as potential biomarkers of anthropogenic or environmental stressors in marine environments and may improve our understanding of diatom bloom ecology.

1.6.2. Metagenomics and environmental antibiotic resistance

The antibiotic resistance problem has been well documented and continues to be a major public health concern worldwide. The costs of resistance are high, resulting in a doubling of hospital stays, mortality and morbidity rates (Levy and Marshall 2004). Bacterial resistant mechanisms are present for most, if not all, antibiotics. Resistance can be due to a number of factors, including impermeability of cell membranes to an antibiotic, multidrug resistant efflux pumps that pump antibiotics (and other compounds such as toxins) out of the cell, mutations that modify the antibiotic-binding site and inactivation of the antibiotic by the proteins encoded by antibiotic resistant genes (Allen et al. 2010). Antibiotic resistance determinants (ARDs), which include antibiotic resistance genes (ARGs), mobile genetic elements (e.g. plasmids or transposons) that carry ARGs and the bacterial species that carry the gene or mobile genetic element, move within and between different environments including human ecosystems and thus can impact human health (Figure 1.5). The overuse and misuse of antibiotics has directly led to

the selection of ARGs in bacterial populations via direct antibiotic exposure or horizontal gene transfer of these ARGs.

Efforts to monitor antibiotic resistance in natural environments have been infrequent and incomplete, and have relied on low-throughput approaches (e.g. culture, PCR) focusing on single species or genes. These approaches are not conducive to broad-scale monitoring or spatiotemporal coverage. Metagenomics, in combination with next generation sequencing, potentially offers a high-throughput approach for assessing ARDs in the environment. The ARD profile of a metagenome may reflect potential anthropogenic impacts indicative of GxE interactions, and thus may be informative for public health surveillance and management efforts in coastal environments. Furthermore, the prevalence and distribution of ARDs have not been well characterized in offshore or open ocean environments, and thus we know little about background levels of resistance in marine environments or the potential associations between anthropogenic pressures (i.e. environment) and ARD abundance in marine ecosystems.

1.7. Hypotheses and Specific Aims

The work presented in this document addresses our overall objective to determine whether environmental genomic information can be used to answer public health relevant questions by identifying sensitive and functional markers of both human impacts on marine ecosystems and marine impacts on human health.

H_1: Because of evolutionary conservation of signaling pathways we anticipate that similar environmental response pathways significant for humans and plants (e.g. GPCR signaling)

will be identifiable in diatoms, thus providing potential mechanisms of algal perception and response to external stressors.

Specific Aim 1: Characterize the potential for a GPCR signaling pathway repertoire in four sequenced diatoms via a comparative genomics approach to the human GPCR pathway.

Specific Aim 2: Identify putative G protein-coupled receptor (GPCR) or GPCR-like sequences and their potential functionality in these diatoms.

H_2: High-throughput analysis of marine metagenomic data will allow us to examine and compare trends in taxonomic and antibiotic resistance determinants across differentially impacted environments and furthermore establish baseline profiles for these determinants in the environment.

Specific Aim 1: Obtain snapshots of microbial community composition across Puget Sound and a proximal wastewater treatment plant using 454 pyrosequencing to determine whether composition changes along a potential gradient of impact.

Specific Aim 2: Develop a bioinformatic analysis framework for annotating antibiotic resistance determinants in metagenomic datasets.

Specific Aim 3: Determine whether an antibiotic resistance determinant signal can be detected in marine metagenomic samples using 454 pyrosequencing.

H_3: Using a high-throughput metagenomic epidemiological approach, we can construct a public health surveillance and decision-making tool related to environmental antibiotic resistance.

Specific Aim 1: Construct an index of environmental antibiotic resistance determinants that can be used to track the resistance potential across differentially impacted marine and freshwater environments.

Specific Aim 2: Quantify the index for environmental metagenomes and identify common index modalities across the samples using principal component analysis.

Specific Aim 3: Conceptually integrate the index a public health surveillance framework for environmental antibiotic resistance.

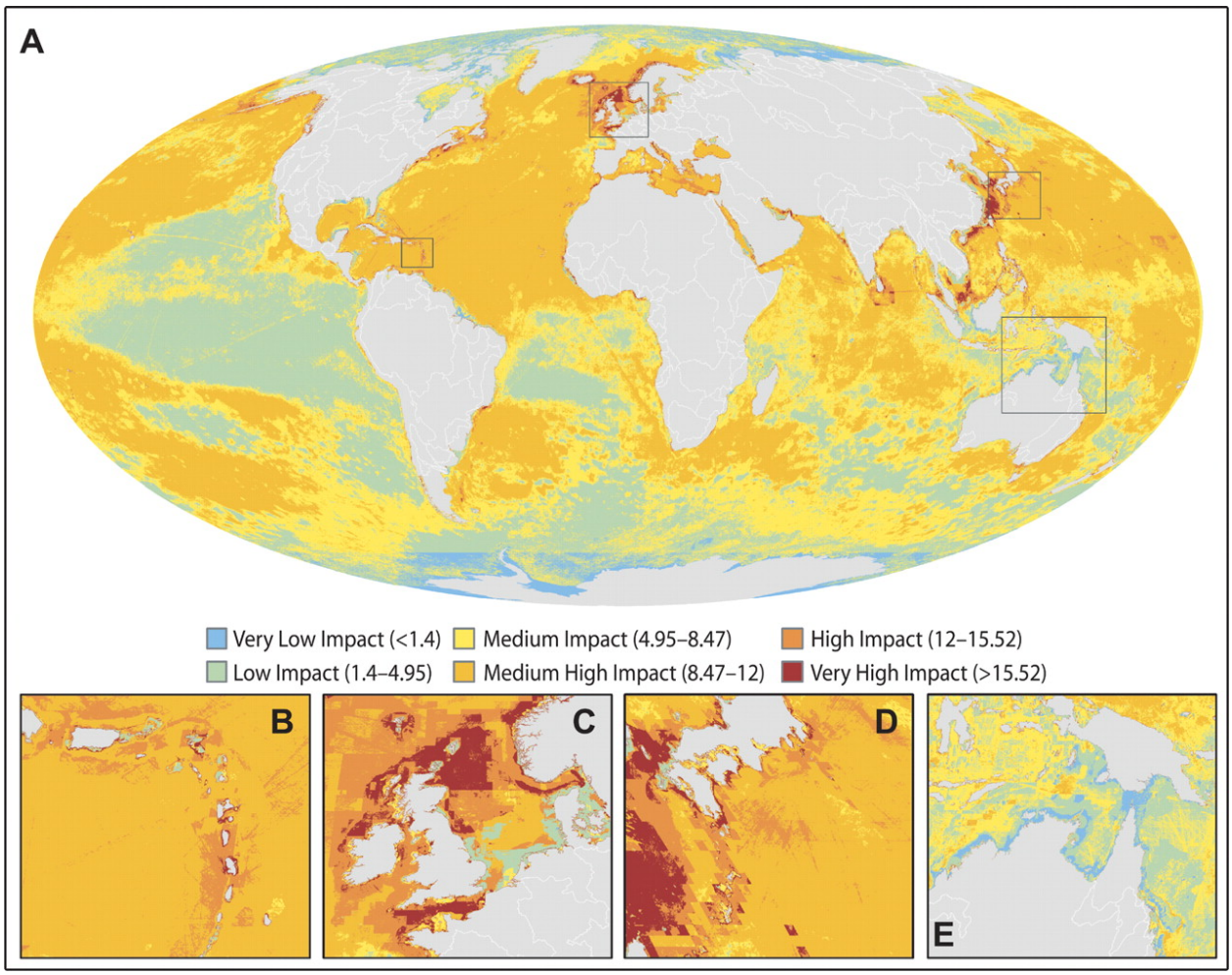

Figure 1.1. Global map (A) of cumulative human impact across 20 ocean ecosystem types. (Insets) Highly impacted regions in the Eastern Caribbean (B), the North Sea (C), and the Japanese waters (D) and one of the least impacted regions, in northern Australia and the Torres Strait (E) and the U.S. Pacific Northwest. Reproduced from (Halpern et al. 2008; Halpern et al. 2009)

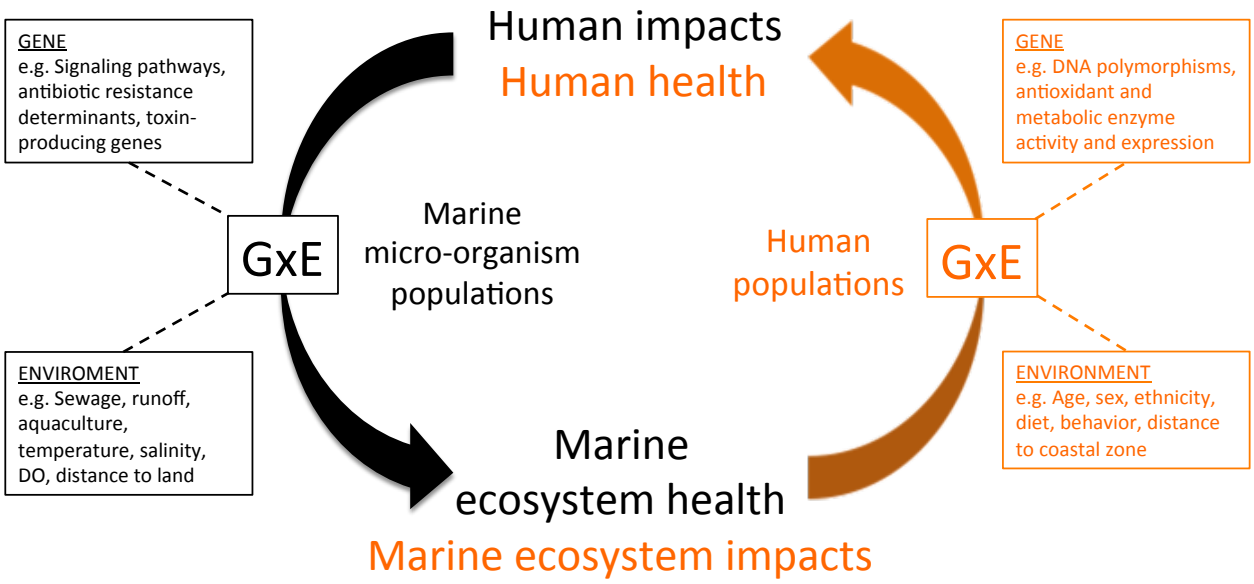

Figure 1.2. Gene-environment (GxE) framework for oceans and human health research. GxE interactions mediate human impacts on marine ecosystem health, and furthermore these impacts can feedback to affect human health also via GxE. While most public health efforts focus on human populations, GxE interactions in marine micro-organism populations allow for exploration of early risk management approaches for environmental health monitoring.

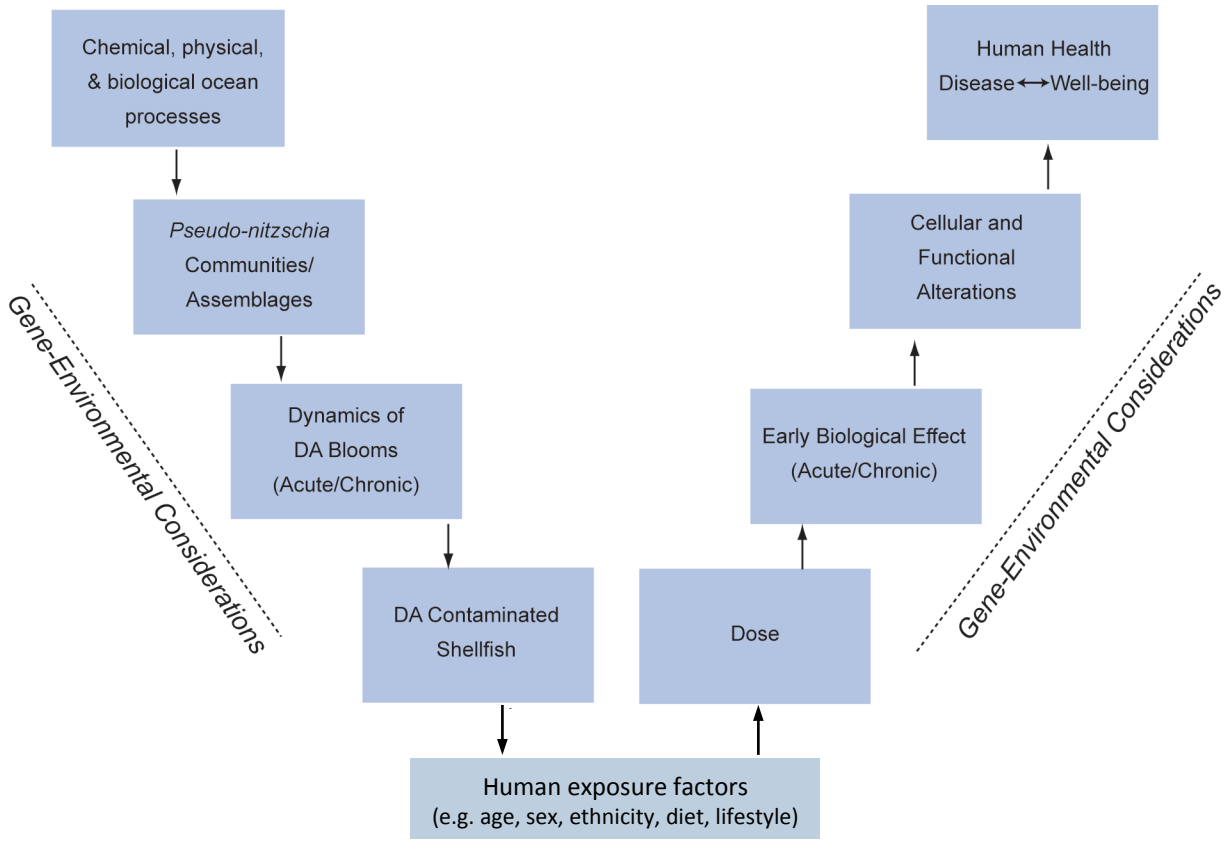

Figure 1.3. Oceans and human health (OHH) risk chain that connects the ways that ocean processes influence toxic algal blooms and how these blooms cause public health impacts and risks. The risk chain is mediated by gene-environment (GxE) interactions at both the diatom and human levels.

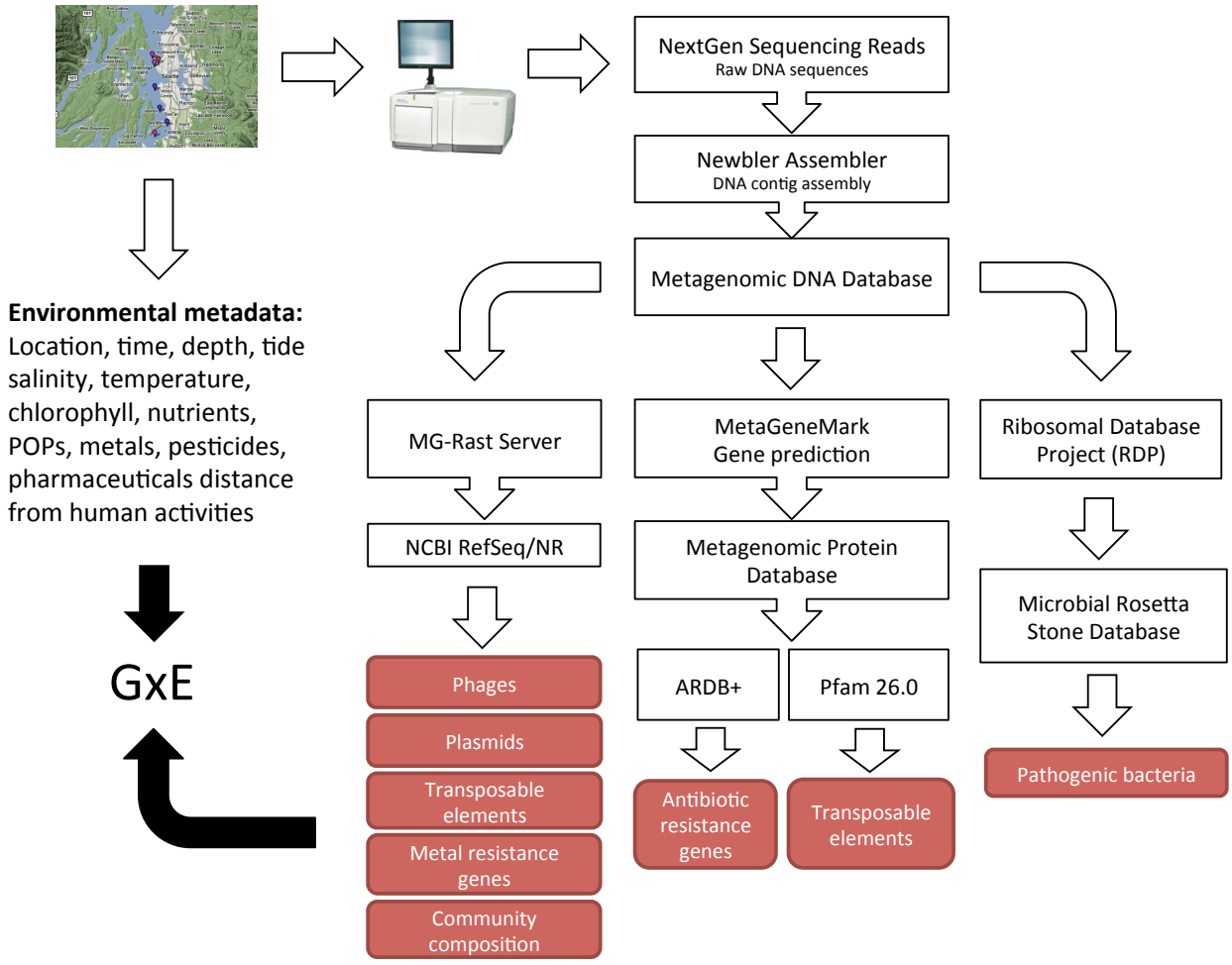

Figure 1.4. Bioinformatic pipeline for high-throughput metagenomic analysis of gene-environment (GxE) interactions relevant to Oceans and Human Health (OHH). ARDB+, expanded Antibiotic Resistance Genes Database; NR, nonredundant database.

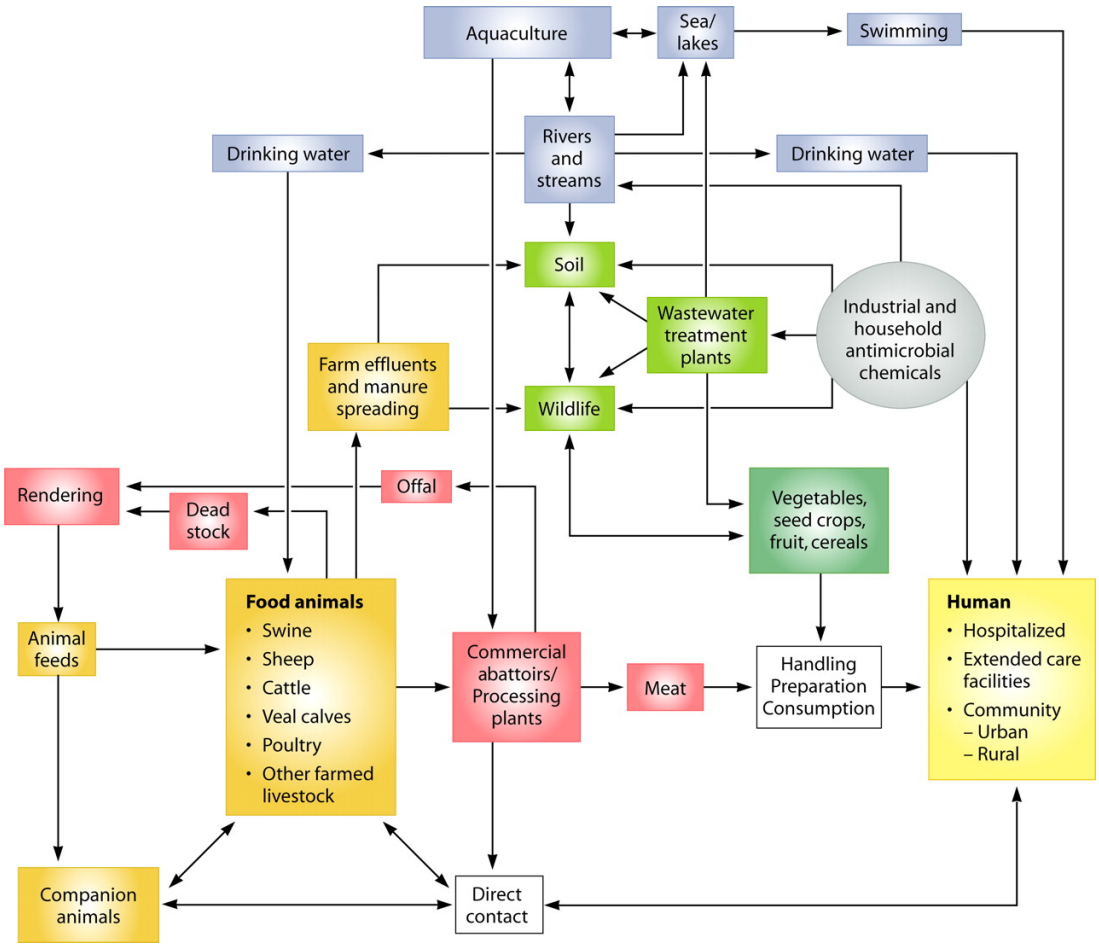

Figure 1.5. Movement and transfer of antibiotic resistance determinants in the environment involves and depends on the interactions between agricultural, community, hospital, aquatic and wastewater treatment environments. Reproduced from Davies and Davies 2010.

Table 1.1 Comparison of currently available next-generation sequencing technologies. Modified from Shokralla et al. 2012 and Metzker 2010.

Platform	Read length (bp)	Max. number of reads/run	Sequencing output/run	Run time	Benefits	Limitations	Environmental genomic/OHH applications
Roche 454 GS FLX	400-500	$1x10^6$	≤500 Mb	10 h	Longer read lengths; fast run times	Lower sequencing depth; homopolymer repeat errors; high reagent costs	De novo assembly; unassembled shotgun metagenomic read analysis; 16s metagenomics; transcriptome sequencing
Roche 454 GS FLX+	600-800	$1x10^6$	≤700 Mb	23 h			
Illumina HiSeq 2000	100-200	$6x10^9$	≤540-600 Gb	11 d	Best balance between sequencing depth and read length	Shorter read lengths than 454	Whole-genome sequencing; contig assembly; transcriptome sequencing
Illumina HiSeq1000	100-200	$3x10^9$	≤270-300 Gb	8.5 d			
ABI SOLiD 5500	35-75	$2.4x10^9$	~100 Gb	4 d	Highest depth of sequencing	Short read lengths	SNP discovery; whole genome sequencing; transcriptome sequencing
ABI SOLiD 5500 xl	35-75	$6x10^9$	~250 Gb	7-8 d			
Ion Torrent - 318 chip	100-200	$11x10^6$	≥ 1 Gb	5.5 h	Less expensive; faster run times	Weaker balance between sequencing depth and read length	Whole-genome sequencing; transcriptome sequencing
Ion Torrent - 316 chip	100-200	$6x10^6$	≥ 100 Mb	4.7 h			

Table 1.2. Bioinformatic databases for comparative metagenomic analyses.

Database	Tools	Size of data repository	Host	Website/Reference
MG-Rast	Taxonomic (LCA and best hit) and functional (COG, eggNOG, KEGG, SEED) annotation; cross-metagenome comparisons; principal component analysis; sequence QC; rarefaction analysis; heat map visualization; links to STAMP (metagenomic statistical package)	~11,000 public metagenomes; 17.66 Tbp	Argonne National Laboratory, U.S. Department of Energy	http://metagenomics.nmpdr.org
CAMERA	Taxonomic (best hit and OTU) and functional (COG, KEGG, Pfam, TIGRfam) annotation; cross-metagenome comparisons; advanced workflows including BLAST, clustering, ORF prediction and sequence QC; GIS query of metadata and genomic data	~72 public metagenomic projects; >48 Bbp; collection of marine microbial reference genomes	University of California, San Diego	http://camera.calit2.net/
MEGAN	Taxonomic (LCA) and functional (KEGG and SEED) annotation; dendrograms to display data; rarefaction analysis; microbial attribute metadata	No data repository	Tübingen University, Germany	http://ab.inf.uni-tuebingen.de/software/megan/
IMG/M	Taxonomic and functional (COG, KOG, KEGG, Pfam, TIGRfam, SEED, pathways, networks) annotation; Metagenome/genome BLAST; Metagenome to genome comparison	1271 public metagenomes; 429 Bbp	Joint Genome Institute, U.S. Department of Energy	http://img.jgi.doe.gov/cgi-bin/m/main.cgi
METAREP	Taxonomic and functional (KEGG, GO, Enzymes, Pfams, TIGRFAMs, pathways) annotation; statistical tests; clustering; heatmap visualization; Cross-link function with phylogeny	No data repository	J. Craig Venter Institute (JCVI)	http://www.jcvi.org/metarep

CHAPTER 2: Identification of G protein-coupled receptor signaling pathway proteins in marine diatoms using comparative genomics

This chapter has been submitted for publication in its entirety and is currently under review with the journal BMC Genomics. The authors of this submitted manuscript are:

Jesse A. Port[1], Micaela S. Parker[2], Robin B. Kodner[2], James C. Wallace[1], E. Virginia Armbrust[2] and Elaine M. Faustman[1]

[1]Department of Environmental and Occupational Health Sciences, School of Public Health

[2]Center for Environmental Genomics, School of Oceanography

2.1 Abstract

The G protein-coupled receptor (GPCR) signaling pathway plays an essential role in signal transmission and response to external stimuli in mammalian cells. Protein components of this pathway have been characterized in plants and simpler eukaryotes such as yeast, but their presence and role in unicellular photosynthetic eukaryotes have not been determined. We use a comparative genomics approach using whole genome sequences and gene expression libraries of four diatoms (*Pseudo-nitzschia multiseries*, *Thalassiosira pseudonana*, *Phaeodactylum tricornutum* and *Fragilariopsis cylindrus*) for evidence of GPCR signaling pathway proteins that share sequence conservation to known GPCR pathway proteins. Except for the G protein γ-subunit and a number of transcription factors, the majority of the core components of GPCR signaling were well conserved in all four diatoms, with protein sequence similarity to human G

protein α- and β-subunits, protein kinases and other downstream effectors. Phylogenetic analysis of putative diatom GPCRs indicated similarity but deep divergence to the class C GPCRs, with branches basal to the $GABA_B$ receptor subfamily. The extracellular and intracellular regions of these putative diatom GPCR sequences exhibited large variation in sequence length, and only six of these sequences contained the necessary ligand binding domain for class C GPCR activation. Transcriptional data indicates that a number of the putative GPCR sequences are expressed in diatoms under various stress conditions in culture, and that a majority of GPCR-activated signaling proteins are also transcribed. The presence of sequences in all four diatoms that code for many of the proteins required for a functional mammalian GPCR pathway highlights the highly conserved nature of this pathway and suggests a complex signaling machinery related to environmental perception and response in these unicellular organisms. The absence of the G protein γ-subunit though warrants further investigation into the structure and functionality of putative G proteins in diatoms. The high divergence of putative diatom GPCR sequences to known class C GPCRs suggests these sequences may represent another, potentially ancestral, subfamily of class C GPCRs. The presence of GPCRs in these organisms has potential implications for environmental health monitoring, drug discovery and human disease research.

2.2 Introduction

The G protein-coupled receptor (GPCR) superfamily represents one of the largest and most diverse families of proteins in mammals and is found in nearly all multicellular life (Bockaert and Pin 1999; Fredriksson and Schioth 2005). These proteins are cell-surface receptors that play a major role in signal transduction and perception of and response to the environment. GPCRs bind a diverse array of ligands including proteins, lipids, neurotransmitters, calcium,

odorants, and other small molecules (Bockaert and Pin 1999). In vertebrates, GPCR signaling networks are associated with neurotransmission, cellular metabolism, secretion, cellular differentiation and growth, inflammatory and immune responses, smell, taste and vision (Qian et al. 2003). All GPCRs share a core seven transmembrane α-helical region with an extracellular ligand binding domain that is coupled intracellularly to a G protein heterotrimer. GPCR activation leads to the exchange of GDP for GTP by a G protein, and G protein subunits then interact and regulate effector molecules (e.g. calcium, adenylyl cyclase, phospholipase C, phosphodiesterases, protein kinases), activating further downstream signaling pathways such as the mitogen-activated protein kinase (MAPK), phosphoinositide-3 kinase (PI3K)-Akt and NF-kappaB pathways that ultimately activate transcription factors that affect gene expression and regulation (Marinissen and Gutkind 2001; Ho et al. 2009). Many of these scaffolding and signaling proteins mediate signal transduction in other intracellular pathways in eukaryotes and thus are highly conserved.

GPCRs are divided into five highly diverged families: *Rhodopsin*/class A, *Secretin*/class B, *Adhesion*/class B, *Glutamate*/class C and *Frizzled/Taste2*/class F (Fredriksson et al. 2003). GPCR sequences within these families can be highly diverged between species, in some cases sharing less than 25% similarity (Moriyama et al. 2006). The importance of these receptors is exemplified by the fact that 3-4% of human genes code for GPCRs and that nearly 30% of all currently marketed drugs target these receptors (Landry and Gies 2008). Numerous endocrine and sensory-related diseases are associated with GPCR mutations in humans (Schoneberg et al. 2004).

Despite the crucial importance of GPCR signaling in metazoa, the prevalence and function of these proteins in non-model organisms such as unicellular photosynthetic eukaryotes is not

well understood. Diatoms are a major class of eukaryotic phytoplankton found throughout the world's oceans that play a crucial role in primary production and nutrient cycling and serve as a base for marine food webs (Armbrust 2009). Diatoms are also responsible for forming large phytoplankton blooms that in some cases can be toxic to humans, marine mammals and seabirds (Scholin et al. 2000; Pulido 2008). While the molecular mechanisms by which diatoms perceive and respond to their surrounding environment have not been resolved, previous findings suggest a role for cell surface receptors linked to intracellular signaling pathways. For example, exposure to osmotic, shear or nutrient (iron) stress in culture leads to changes in cytosolic Ca^{2+} concentrations in the diatom *Phaeodactylum tricornutum* (Falciatore et al. 2000). The presence of a chemical-based defense system in *P. tricornutum* and *Thalassiosira weissflogii* has also been reported in which these diatoms respond to challenge via diatom-derived aldehydes triggering Ca^{2+} and nitric oxide release (Vardi et al. 2006). This "stress surveillance system" may function in cell-cell communication across diatom populations to detect damaged or stressed cells resulting from phytoplankton competitors and other ecological or physical stressors. These findings are important when considering environmental perception and response as alterations in Ca^{2+} homeostasis are a hallmark of signal transduction activation throughout the eukaryotes (Berridge et al. 2003). Levels of the second messenger cAMP have also been shown to change in cultures of *P. tricornutum* following exposure to elevated carbon dioxide levels (Harada et al. 2006). While there is sequence evidence for putative GPCR signaling pathway proteins in the *T. pseudonana* (Armbrust et al. 2004; Montsant et al. 2007; Nordstrom et al. 2011) and *P. tricornutum* (Bowler et al. 2008) genomes, the role GPCR signaling may play in regulating environmental perception and response in diatoms warrants more detailed investigation.

Here we use an *in silico* approach to probe the genomes of *P. multiseries*

[http://genome.jgi-psf.org/Psemu1/Psemu1.home.html], *T. pseudonana* (Armbrust et al. 2004), *P. tricornutum* (Bowler et al. 2008) and *Fragilariopsis cylindrus* [http://genome.jgi-psf.org/Fracy1/Fracy1.home.html] for translated nucleotide sequences with similarity to known GPCR signaling pathway proteins. We also probe expressed sequence tag (EST) libraries for each diatom to determine whether these putative proteins are actively expressed in laboratory isolates. Our rationale for emphasizing sequence comparisons between diatoms and higher eukaryotes is three-fold. First, the GPCR signaling pathway is well-characterized in mammals compared to less well-studied, non-model organisms and thus the functions of putative homologs are better understood in this system. While model organisms such as yeast provide valuable insight into potential GPCR signaling mechanisms in mammalians, yeast and humans are found in the same eukaryotic supergroup, and thus other unicellular systems with different evolutionary histories found outside this supergroup would allow for further comparative analyses of GPCR signaling pathway diversity. Secondly, diatoms must rapidly sense and respond to multiple environmental changes, many of which are likely mediated by receptor-based signaling pathways. As major contributors to ocean productivity and carbon cycling, diatoms may play a critical role in the changing ecosystems of the future ocean, and thus understanding the breadth of their ability to sense and respond to environmental changes may be crucial to predicting their future success. Lastly, from a human health perspective, a better understanding of GPCR conservation and functionality in other organisms may provide further insight into the importance of these receptors as extracellular or environmental sensors and as pharmacological and human disease relevant targets.

The goal of this study is thus to provide a comprehensive analysis of the GPCR signaling repertoire and its potential functionality in sequenced diatoms by using a suite of bioinformatics

tools aimed at annotating the genomes of non-model organisms. We hypothesize that the conservation of this pathway in diatoms may reflect shared mechanisms of environmental response related to GPCR signaling across the eukaryotes. Conservation of this pathway has implications for the evolutionary and ecological success of these organisms in addition to informing cross-species GPCR research involving drug discovery and human disease.

2.3 Methods

The complete genomes and filtered models for the diatoms *T. pseudonana* v3.0, *P. tricornutum* v2.0, P. multiseries v1.0 and *F. cylindrus* v1.0 were obtained from the DOE Joint Genome Institute (JGI) database [http://www.jgi.doe.gov/genome-projects/], and the human proteome (~37,000 proteins) from the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database. Expressed sequence tags (ESTs) for each diatom were downloaded from GenBank and supplemented with the diatom ESTs from JGI. In total we data-based 61,913 ESTs for *T. pseudonana*, 133,807 for *P. tricornutum,* 16,535 for *P. multiseries* and 21,802 for *F. cylindrus*.

2.3.1. GPCR signaling pathway analysis

Using the Basic Local Alignment Search Tool (BLAST) program (Altschul et al. 1990), the human proteome was TBLASTN queried against the translated genomes and EST libraries of each diatom at an E-value threshold of 10^{-20} to generate protein sets for *P. multiseries*, *F. cylindrus*, *T. pseudonana* or *P. tricornutum* with similarity to known human proteins (Figure 2.1). The protein sets were categorized by Gene Ontology (GO) (Harris et al. 2004). Proteins involved in the GPCR signaling pathway (Table 2.1) were identified in the proteins sets and the

best diatom sequence hit for a human GPCR pathway protein was defined as the sequence that had the lowest E-value ($<10^{-20}$) over an alignment length > 200 amino acids and a bit score >100. For each diatom, the predicted protein set corresponding to the genome BLAST was compared to the protein list for the EST BLAST to determine whether the best hits from the genome BLASTs also exhibit transcriptional activity.

2.3.2. GPCR identification

Known and predicted eukaryotic GPCR protein sequences from all GPCR families were downloaded from the GPCRDB and TBLASTN searched against each translated diatom genome using an E-value threshold of 10^{-10} (Figure 2.1). These predicted protein sequences were entered into the TMD prediction programs HMMTOP (Tusnady and Simon 2001) and TMMHMM (Krogh et al. 2001) using default settings. Of those sequences with 7 TMDs, only those predicted to have an extracellular N-terminus (i.e. ligand binding domain) were selected for further analyses. Duplicate sequences were removed. These putative diatom GPCR sequences were further verified for evidence of GPCR motifs and conserved domains using the NCBI Conserved Domain Database (CDD) v2.21 (Marchler-Bauer et al. 2005) and Pfam v26.0 (Punta et al. 2012) using an e-value threshold of 10^{-5}.

A second approach to identify putative diatom GPCRs involved downloading sequence alignments for the GPCR families from the GPCRDB (classes A, B and C) and then converting these seed alignments to HMM profiles to search against a custom microeukaryote database (Table 2.2) and the NCBI RefSeq database using HMMER3 (E-value $<10^{-10}$).

The putative diatom GPCR sequences were also searched against their respective ESTs that were generated under different culture treatments (Table 2.3). We also searched the putative *T. pseudonana* GPCR sequences against transcriptomic datasets representing silicon repletion and

starvation (Mock et al. 2008; Shrestha et al. 2012).

2.3.3. Phylogeny of putative diatom GPCRs

To phylogenetically classify putative diatom GPCRs, a seed alignment was generated that contained the 7 TMD regions of the 16 putative diatom GPCR sequences identified through the BLAST search of the diatom genomes against the GPCRDB and subsequent motif and N- and C-terminus analyses. The diatom TMD sequences were then converted to a HMM profile using the HMMER3 sequence analysis software package (Eddy 1998) and searched against the custom microeukaryote and the NCBI RefSeq databases (E-value $<10^{-10}$). Only those recruited sequences containing 6-8 TMDs were retained. A subset of the recruited sequences from each of the represented class C subfamilies ($GABA_B$, metabotropic glutamate, calcium-sensing, vomeronasal, pheromone, odorant and taste receptors) in addition to predicted and hypothetical proteins were selected based on the lowest E-values per NCBI taxa ID and included in a single alignment using MUSCLE (Edgar 2004). This alignment was used to generate a phylogenetic tree representing all class C GPCR subfamilies. Phylogenetic and molecular evolutionary analyses were conducted using MEGA v.5 (Tamura et al. 2011). The best-fit model of protein evolution was predicted for the alignment. A maximum likelihood tree (Yang et al. 1995) was inferred using the WAG (+F) model of evolution (Whelan and Goldman 2001) and 100 bootstrap iterations.

2.4 Results

2.4.1. GPCR signaling pathway analysis

Using the analysis framework shown in Figure 2.1, we first searched the diatom genomes and ESTs for evidence of GPCR signaling pathway proteins, many of which are expected to be

highly conserved across eukaryotes (Table 2.1). Based on sequence similarity to human proteins, all four diatoms have genes coding for core components of the GPCR signaling pathway. These putative proteins share 30-58% sequence identity to human GPCR pathway proteins when considering alignment lengths >200 aa (Figure 2.2 and Table 2.4). G protein α- and β-subunits were well-conserved across the diatoms, especially within the *T. pseudonana* genome. There was no sequence similarity to the G protein γ-subunit in any of the diatoms, suggesting either the γ-subunit is sufficiently diverged it cannot be identified by sequence similarity or the diatom G protein has a heterodimeric structure. To verify the absence of the G protein γ subunit, the diatom genomes were searched for evidence of the γ subunit and γ-like subunit conserved GGL domain (Pfam accession no.: PF00631) at an e-value cutoff of 10^{-5}. The GGL domain is conserved across eukaryotic G protein γ subunits, but there were no apparent homologs in any of the diatom genomes. All four diatoms have predicted amino acid sequences with high sequence similarity ($E < 10^{-100}$) to G protein-binding protein CRFG (NP_036473). This small GTPase acts as a molecular switch in signal transduction pathways and is similar to the α-subunit of heterotrimeric G proteins.

Downstream pathways such as MAPK were represented (Figure 2.2) and may follow a similar pattern of activation as in mammals via GPCR signaling and subsequent activation of stimulatory (Gαq or Gαs) or inhibitory (Gαi) G protein α- subunits. Diatom sequences with low similarity to human NF-kappa beta (NF-kB) were further analyzed for conserved NF-kB domains to confirm the lack of homologs. NF-kB proteins are a large family of transcription factors that modulate a large number of cellular processes including stress, immune response, growth, development and apoptosis (Perkins 2007). NF-kB is found in almost all animal cells as well as more simple organisms including sea anemones, sponges, insects and the single-celled

eukaryote *Capsaspora owczarzaki* but is absent in yeast and *Caenorhabditis elegans*. These diatom sequences with low similarity to human NF-kB contained the ankyrin (ANK) repeat domain which mediates protein-protein interactions across a wide range of protein families but lacked the NF-kB transcription factor domains involved in DNA binding. Strong phospholipase C (PLC) conservation as shown in in Figure 2.2 and the presence of DAG and IP_3 kinases and phosphatases suggests signaling activity by the second messengers diacyl glycerol (DAG) and inositol 1,4,5-triphosphate (IP_3). Sequences encoding cAMP-dependent protein kinase (PKA) and adenylate cyclases in all four diatoms also support an important role for cAMP signaling. The important and multifunctional mammalian transcription factor cAMP response-element binding (CREB) is not present in the diatoms even at a much less conservative e-value threshold of 10^{-5}, but there is sequence conservation of a human CREB binding protein (NP_004371). Except for a number of the *F. cylindrus* sequences, the majority of diatom nucleotide sequences with similarity to GPCR signaling proteins have EST support. The concordance between the *F. cylindrus* genome and ESTs is most likely lower because mRNA sequencing is not as complete.

We also searched the *Viridiplantae* (land plants and green algae) for the presence of GPCR signaling components (Table 2.4). Similar to diatoms, the *Viridiplantae* also lack obvious homologs to NF-kB and cAMP response-element binding (CREB). Other components of core GPCR signaling pathway proteins appear to be present within the *Viridiplantae*, including the G protein γ-subunit.

2.4.2. GPCR Identification

Due to the divergence of GPCR sequences, the BLAST search space was expanded to include known and predicted non-human eukaryotic GPCRs within the G protein-coupled

receptor database (GPCRDB). This approach resulted in 5 *F. cylindrus,* 5 *P. multiseries*, 4 *P. tricornutum* and 2 *T. pseudonana* candidate GPCR sequences (Table 2.3) with correct transmembrane domain (TMD) orientation and the strongest similarity to the class C GPCR family. *P. tricornutum* GPCR1a and GPCR1b appear to be splice variants of the same gene. The TMD sequences contained residues conserved across the class C GPCRs (Figure 2.3). The two cysteine residues in the first and second extracellular loops are postulated to be involved in disulfide bond formation (Pin et al. 2003). The two basic residues (Lys and Arg) at the cytoplasmic face of the third TMD and the acidic residue (Glu or Asp) at the cytoplasmic end of the sixth TMD (Binet et al. 2007) have been shown to be crucial for activation of the $GABA_B$ receptor subclass of class C GPCRs. The PK motif in the seventh TMD is also conserved in class C GPCRs (Pin et al. 2003). Other conserved residues that may be important for receptor activation in the diatom sequences are highlighted in Figure 2.3.

We used a second approach to identify potential diatom GPCRs by downloading the class A, B and C GPCR alignments from the GPCRDB, and then converting these alignments to hidden markov model (HMM) profiles to search against a custom microeukaryote database (Table 2.2) and the NCBI RefSeq database. No diatom sequences were recruited to the class A and class B alignments, while two sequences that were also identified using the BLAST approach (*F. cylindrus* GPCR5 and *T. pseudonana* GPCR2) were recruited to the class C alignment. These results further indicate that putative diatom GPCRs are likely to be highly diverged from known metazoan GPCRs and that they appear to have family C GPCR-like properties.

Based on sequence similarity searches against the diatom EST libraries, a number of the putative diatom GPCR sequences are transcribed (Table 2.3). The putative *P. tricornutum* GPCR

sequences had the greatest EST coverage, likely due to the larger repository of ESTs for this diatom. *P. tricornutum* GPCR1a and GPCR1b (splice variants) had 100% EST coverage, while GPCR2 had EST coverage for the N-terminus and first 5 TMDs, GPCR3 full coverage except for ~150aa of the N-terminus and GPCR4 full coverage except for ~240aa of the N-terminus and the C-terminus. Of the *F. cylindrus* and *P. multiseries* putative GPCRs, only *F. cylindrus* GPCR4 (100% coverage), *P. multiseries* GPCR1 (TMD and C-terminus) and *P. multiseries* GPCR2 (TMD) appear to be expressed. *T. pseudonana* GPCR2 had EST coverage for the end of N-terminus, TMD and beginning of the C-terminus. The majority of these ESTs are expressed in culture under various stress conditions, including growth-limiting factors for *P. tricornutum*, domoic-acid producing conditions for *P. multiseries*, osmotic stress for *F. cylindrus* and iron limitation and silicon starvation for *T. pseudonana*.

The length of the candidate diatom GPCR sequences ranged from 300 to over 1000 amino acids. This wide range was due to considerable size and domain variation in the N- and C-termini among the candidate class C GPCR sequences (Figure 2.4). Only six of these sequences contain an N-terminus domain with similarity to prokaryotic periplasmic binding proteins (PBPs), which are involved in amino acid and nutrient transport in bacteria and are considered to be the origin of the class C GPCR ligand binding domains (Pin et al. 2003). Based on BLAST similarity, three of these six sequences (*P. multiseries* GPCR1, *P. multiseries* GPCR2 and *F. cylindrus* GPCR1) have N-termini with strongest similarity to the class C GPCRs, while the other three from *P. tricornutum* have similarity to the periplasmic component of ABC sugar transporters, which also are known to contain PBPs. There is no evidence in the extracellular region upstream of the TMD for a cysteine-rich domain. This domain is found in certain class C GPCRs including metabotropic glutamate, calcium-sensing and taste receptors but is absent in

$GABA_B$ receptors (Lagerstrom and Schioth 2008). The C-termini of the putative diatom GPCRs share no similarity to other GPCRs or known proteins in general. This is not the case for other known class C GPCRs, which share sequence similarity in the C-terminus. While the C-termini of mammalian metabotropic glutamate receptors, a subclass of the family C GPCRs, have been shown to exhibit alternative splicing (Hermans and Challiss 2001), there was no evidence of splicing in the putative diatom GPCR sequences.

2.4.3. Phylogeny of predicted diatom GPCRs

To determine the phylogenetic relationship among the putative diatom GPCRs and their relationship to known GPCRs from other domains of life, the 7 TMD regions of the putative diatom GPCRs identified by BLAST were used as the seed alignment to HMM search for other GPCRs in the custom microeukaryote and NCBI RefSeq databases (Figure 2.1). The recruited sequences were then used to generate a tree that examined the diversity of possible diatom GPCRs among other putative microeukaryote and established GPCRs. The diatom TMD HMM profile recruited an array of class C GPCRs, with the strongest sequence similarity to the $GABA_B$ receptors and to lesser extent the mGluR and calcium-sensing/vomeronasal receptors. For maximum likelihood (ML) tree construction, the wag amino acid model was found to have the highest likelihood score when comparing models with different fixed substitution rates among the different amino acid changes. We also tested a number of other amino acid models and obtained similar tree topology. Within the ML tree, the majority of diatom sequences clustered with one another and with other microeukaryotes and formed a recently diverged sister clade to the $GABA_B$ receptors albeit with weak bootstrap support (Figure 2.5). The six putative diatom class C GPCR sequences that contained the PBP1 ligand binding domain grouped within

a divergent clade. Within this clade, the sequences were differentiated by their N-termini, with the three *P. tricornutum* sequences that had similarity to bacterial ABC transporters grouping together and the remaining three sequences (*P. multiseries* GPCR1, *P. multiseries* GPCR2 and *F. cylindrus* GPCR1) clustering together with strong bootstrap support. Other than the stramenopiles (diatoms and *Phytopthora*), microeukaryotic sequences recruited to the tree included the slime mold (*Dictyostelium sp.*) and the coccolithophore *Emiliania huxleyi*.

2.5 Discussion

Using a comparative genomics approach, we elucidated a GPCR signaling repertoire in diatoms that points to the highly conserved nature of this signaling pathway across the eukaryotes. Specifically, sequence comparisons indicate the presence and expression of the G protein α- and β- subunits, protein kinases and other common downstream effectors and pathways known to be activated by GPCR signaling. There was no evidence for the G protein γ-subunit. Based on a combination of BLAST, HMM profiling and phylogenetic analyses, the putative diatom GPCRs were most similarly related to but still highly diverged from the class C GPCRs. They furthermore formed a clade basal to the $GABA_B$ receptors within the class C GPCRs. There was considerable variability in the size of the putative diatom GPCRs, and only six of these sequences had N-terminus similarity to the class C ligand binding domain. A number of these putative GPCR sequences also have EST support. These putative GPCRs may potentially represent an additional group of class C sequences recovered from diatoms and other microeukaryotes that do not fit within existing class C subfamilies.

2.5.1 The diatom GPCR signaling pathway repertoire

The majority of the proteins involved in GPCR signaling pathway proteins appeared to be well conserved in the diatom genomes based on protein sequence similarity. Furthermore, gene expression data indicates that the genes coding for many of these proteins are transcribed. The G protein γ-subunit was an exception. Despite strong conservation of the G protein α- and β-subunits in all four diatoms, there was no evidence for diatom sequences encoding the γ-subunit. This finding has also been reported elsewhere for *T. pseudonana* (Montsant et al. 2007). All identified functional G proteins in other organisms are heterotrimeric, consisting of α, β and γ-subunits. The β and γ-subunits function structurally as a monomer, as the two subunits cannot be dissociated (Clapham and Neer 1997). The G protein βγ subunits regulate the functionality of the α-subunit in addition to mediating downstream signaling pathways. While the γ-subunits of different mammalian species can share as low as 27% sequence similarity (Dupre et al. 2009), we were still not able to recover the highly conserved γ-subunit GGL domain. Thus the structure and functionality of the G protein in diatoms remains to be resolved.

The putative diatom GPCR sequences had the strongest similarity to the class C GPCRs based on TMD and in a few cases N-terminus similarity. Class C receptors are characterized by a long extracellular N-terminus (~600 aa) required for ligand binding. The diatom N-termini ranged from 40-730 aa, with the shorter domains potentially representing a truncated or absent ligand binding region and thus altered functionality. Viral GPCRs provide an example of a GPCR with a very small or absent N-terminus whose functionality is maintained through constitutive expression (Rosenkilde et al. 2001). The N-terminus ligand binding domain required for class C GPCR activation was present in 6 of the putative GPCRs. The ligand binding domain is considered to have evolved from prokaryotic periplasmic binding proteins which are involved in amino acid and nutrient transport in bacteria (Pin et al. 2003). Three of these candidate GPCRs

(from *P. tricornutum*) had stronger similarity to the periplasmic component of bacterial ABC sugar transporters, which are in the same superfamily as the GPCR ligand binding domains. Their TMDs had typical class C structure and similarity. These sequences were also shorter than typical class C receptors (especially within the N-terminus). Taken together, these results suggest that these *P. tricornutum* sequences are class C GPCR-like but may have modified or altered binding activity. The lengths of the intracellular C-termini were also variable across the diatom sequences, and the C-termini shared no sequence similarity to any other proteins. This is not that unexpected given the fact the C-terminus is the least conserved region of class C GPCRs, even among orthologs (Pin et al. 2003). The C-terminus is not required for receptor coupling to G proteins, but interactions have been shown between this region and various scaffolding proteins associated with receptor signaling, desensitization and targeting. In sum, while the TMD regions of these putative diatom GPCR sequences were conserved, further investigation into the diversity of the N- and C- termini are needed to classify these sequences. The phylogenetic analysis does indicate that there is considerable diversity even within the TMDs that distinguishes the diatom and microeukaryotic sequences from the TMDs of metazoan class C GPCRs.

The extent and diversity of microeukaryotic sequences recovered from the HMM searches suggests GPCR signal transduction is an ancient signaling cascade retained in diverse forms of life. The putative diatom and microeukaryotic GPCRs were distantly related to the metazoan class C GPCRs and formed a series of branches basal to the metazoan $GABA_B$ receptor sequences. A distant relationship between these groups of putative GPCRs is expected considering the deep evolutionary relationships of these organisms and the unlikely functional homology of the putative diatom and microeukaryote sequences to metazoan receptors. The class C GPCR subfamily profile recovered sequences from the stramenopiles (diatoms and

Phytopthora), a haptophyte (*Emiliania huxleyi*) and slime mold (*Dictyostelium* sp.). Slime molds are evolutionarily basal to the opisthokonts, which include metazoa, and thus the presence of metazoan homologs in *Dictyostelium* is not surprising. Diatoms and haptophytes were the only photosynthetic organisms recruited by these searches. Both of these groups are products of secondary endosymbiotic events (Parker et al. 2008) and therefore derive from a more recent heterotrophic host cell than green or red algae. An example of the acquisition of a novel trait due to this secondary endosymbiotic event is the presence of the urea cycle in diatoms (Allen et al. 2011). The urea cycle is typically associated with metazoan metabolism, and is absent in green algae and plants. The class C GPCR-like proteins in the diatoms may also reflect their evolutionary history and represent a more ancestral version of this protein family. The other photosynthetic chlorophyll a + c groups that were included in the custom database, such as dinoflagellates and cryptophytes, were not recruited to the class C subfamily profile with a HMMER e-value $<10^{-5}$. The diatom sequences profile also did not recruit any sequences at an e-value $<10^{-5}$ from *Aureococcus,* another microalgal stramenopile. A putative GPCR sequence was recovered from an additional slime mold but no sequence matches were recovered from the many other heterotrophic microeukaryotes for which ESTs are available or from *Naeglaria gruberii*, one of the only free living excavate-lineage genomes available. The apparent sporadic recovery of GPCR-like sequences from microeukaryotes may simply be an issue of low reference sequence availability, possibly compounded by low expression of the putative class C receptor homologs in microeukaryotes, limiting representation in EST libraries. Alternatively, the basal nature of the putative diatom and microeukaryotic class C GPCR sequences may indicate that these sequences represent a separate, possibly ancestral, family of class C GPCRs. It is possible that diatom GPCRs underwent an independent evolution after a recombination event

between an ancestral class C receptor and a GPCR-like locus, or evolved directly from an ancestral class C receptor as has been proposed for plant GPCRs (Turano et al. 2001).

2.5.2. Functionality of GPCR signaling in diatoms

Class C GPCRs are responsible for a vast array of physiological processes ranging from the modulation of synaptic transmission to the perception of sensory stimuli in the nervous system (Kuang et al. 2006). The primary ligands for class C receptors include the neurotransmitters glutamate and GABA. Sequences cloned from the sponge (Perovic et al. 1999) and the amoeba *Dictyostelium discoideum* (Taniura et al. 2006) represent the most basal class C receptors that have been identified to date. A metabotropic glutamate receptor identified in *D. discoideum* (DdmGluPR) is diverged from other characterized metabotropic glutamate receptors in multicellular organisms (Figure 2.5), as also appears to be the case for the putative diatom GPCR sequences. Functionally, DdmGluPR is involved in early development of *D. discoideum* (Taniura et al. 2006). GABA produced by *D. discoideum* also functions as an intracellular signaling molecule by regulating differentiation during development through a $GABA_B$ receptor homologue (Loomis and Anjard 2006; Fountain 2010). *D. discoideum* thus provides an example whereby class C receptors play an important functional role in the absence of neuronal synapses, and may therefore shed light on the functionality of potential diatom GPCRs.

While GABA has been directly detected in cultures of *P. tricornutum* (Bowler et al. 2006), there has been no documented role for GABA in diatoms or algae in general. In the mammalian central nervous system, GABA is the most widely distributed inhibitory neurotransmitter (Bormann 2000). It has also been found in nearly every plant and plant part examined (Kinnersley and Turano 2000). GABA levels in plants increase several-fold in

response to biotic and abiotic stresses such as heat shock, cold shock, mechanical stimulation, hypoxia, phytohormones, and water stress (Shelp et al. 1999; Bouche and Fromm 2004). There is also support for a role for GABA in plants in contributing to C:N balance, regulation of cytosolic pH, protection against oxidative stress, self-defense, osmoregulation and cell signaling (Bouche and Fromm 2004). A GPCR system in diatoms involving GABA or other extracellular or intracellular ligands may function similarly in protecting the cells against abiotic and biotic stressors such as temperature or salinity changes. The coupling of $GABA_B$ receptors with Ca^{2+} ion channels or Ca^{2+}-sensing receptors in other organisms (Ohmori et al. 1990; Cheng et al. 2007) suggests the potential for a GPCR to mediate the documented increase in intracellular Ca^{2+} following exposure to physical and chemical stressors in diatoms (Falciatore et al. 2000; Harada et al. 2006; Vardi et al. 2006).

2.5.3. Human health significance of diatom GPCR signaling

The presence of GPCR signaling in diatoms has potential implications for human health research. First, the presence of GPCRs in simple eukaryotes such as algae has relevance to drug discovery and human disease research. GPCRs are extremely important pharmacological targets, with nearly 30% of all currently marketed drugs targeting GPCRs. Furthermore, human diseases associated with endocrine disruption and sensory abnormalities have been linked to GPCR mutations in each GPCR subfamily (Schoneberg et al. 2004; Spiegel and Weinstein 2004). Identification of new model systems for GPCR research is thus important, and identification of GPCRs in eukaryotic supergroups representing different evolutionary histories would expand the sequence and function space covered by these systems. Transgenic, recombinant and transfected cell systems have commonly been used to study the cellular and phenotypic effects of GPCR

defects (Seifert and Wenzel-Seifert 2002), but unicellular organisms have been demonstrated to be successful model systems for disease investigations. Examples include *Chlamydomonas reinhardtii* (green algae) and ciliary dyskinesia and lateralization defects (El Zein et al. 2003), *Dictyostelium discoideum* (slime mold) and immune system, mitochondrial and neurodegenerative diseases (Annesley and Fisher 2009) and *Saccharomyces cerevisiae* (yeast) and mitochondrial, neurodegenerative and developmental diseases and ageing (Petranovic and Nielsen 2008). The shorter life cycle of algae and other simple eukaryotes potentially would allow for cheaper and faster preliminary studies that could then be supplemented with more traditional models if necessary. Secondly, because of the role GPCRs play in extracellular perception and response in mammalian cells, it is worth speculating that these receptors in diatoms may be involved in cell-cell communication or response to environmental stimuli. From an oceans and human health perspective, the possible roles these receptors may play in harmful algal bloom formation, toxin production (e.g. *Pseudo-nitzschia* spp.), programmed cell death or contaminant exposure is worth investigating. Linking GPCR expression in culture to the molecular mechanisms associated with these pathways would provide insight into the potential use of GPCRs as molecular markers in environmental health monitoring.

2.6 Conclusions

In summary, using a cross-species comparative genomics approach this study has found conservation at the amino acid level of many of the core proteins involved in the mammalian GPCR signaling pathway in diatoms. The G protein γ-subunit and a number of important transcription factors appear to be absent though. A number of the putative diatom GPCR sequences have transcriptional evidence under various stress conditions in culture. Phylogenetic

analyses revealed the putative diatom GPCR sequences to be most similar to but diverged from the class C GPCRs, and within this family they appear to form a unique clade basal to the $GABA_B$ receptors. These putative diatom GPCR sequences exhibit high diversity in the N- and C- termini and have a conserved TMD region that is unique from that of metazoan class C receptors. The presence of GPCRs and GPCR signaling proteins in diatoms suggests a secondary signaling mechanism that warrants further experimental investigation to better define the functional roles of these proteins in diatoms. The confirmation of GPCR functionality in diatoms would indicate that these organisms are able to perceive and respond to their surrounding environment in a more complex manner.

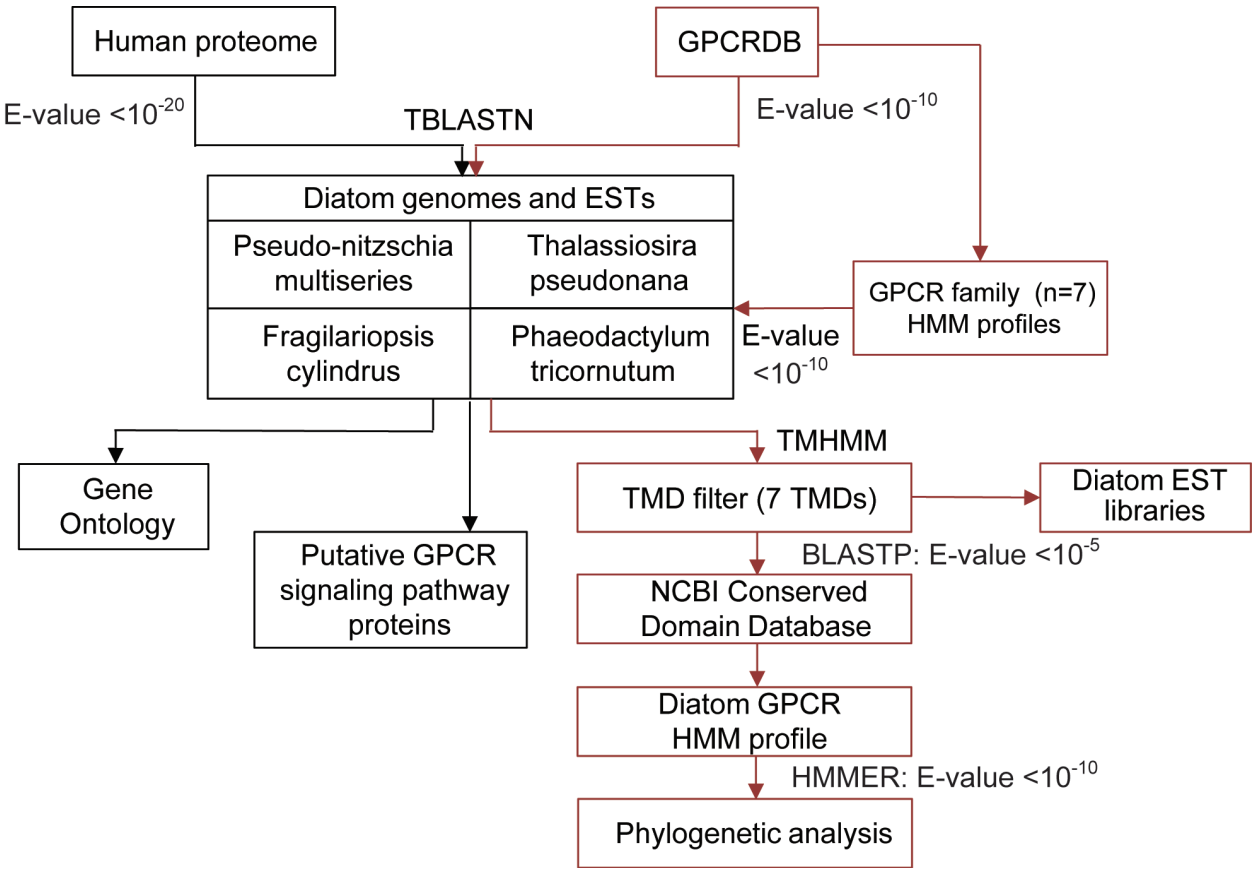

Figure 2.1. Data analysis framework for investigating the G protein-coupled receptor (GPCR) signaling pathway in diatoms. The diatom genomes and EST libraries were first BLAST searched against the human proteome to identify potential GPCR signaling pathway proteins. Putative diatom GPCRs were further characterized using transmembrane domain (TMD) region and conserved domain analyses. A seed alignment was generated using the TMD regions of the putative diatom GPCRs and converted to an HMM profile to recruit related sequences from a custom microeukaryote database and GenBank. Phylogenetic analysis was then performed. Individual GPCR family and class C GPCR subfamily alignments were also downloaded from the GPCRDB and used as seed alignments for HMM searches to further identify potential diatom GPCR sequences that may have been missed using the previous approach. BLAST, basic local alignment search tool; GPCRDB, G protein-coupled receptor database; HMM, hidden markov model; MUSCLE, multiple sequence comparison by log-expectation; TBLASTN, protein query versus translated nucleotide BLAST; TMHMM, transmembrane hidden markov model.

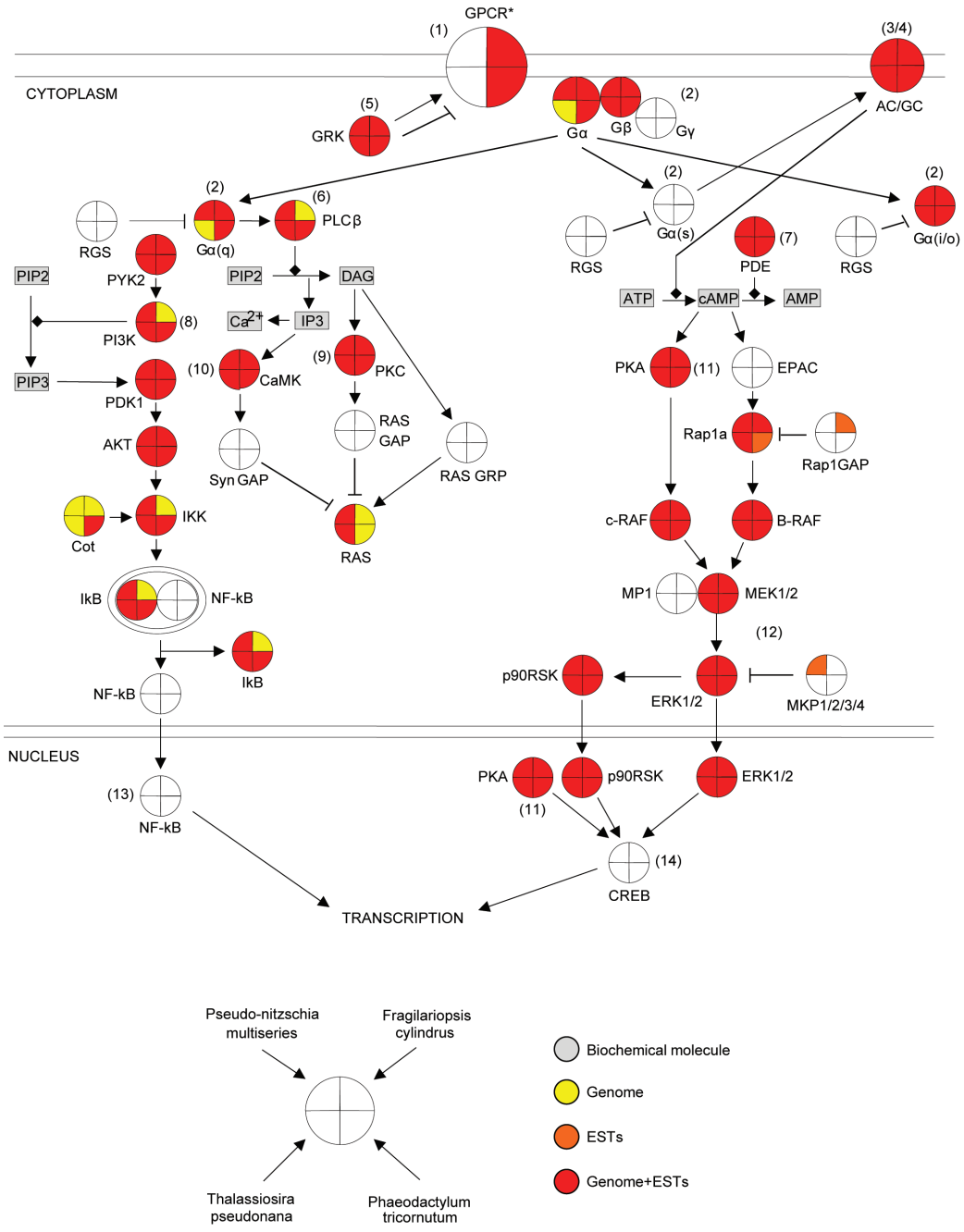

Figure 2.2. Human G protein-coupled receptor (GPCR) signaling pathway proteins that are conserved across the diatom genomes and EST libraries. Conservation is based on sequence similarity criteria consisting of an e-value $<10^{-20}$, alignment length >200 amino acids and bit score >100. Numbers beside proteins refer to descriptions in Table 2.1. AKT, protein kinase B; B-RAF, serine/threonine-protein kinase B-Raf; c-RAF, RAF proto-oncogene serine/threonine-protein kinase; COT, mitogen-activated protein kinase 8; DAG, diacylglycerol; EPAC, cAMP-regulated guanine nucleotide exchange factor; ERK1/2, mitogen-activated protein kinase; Gα(q), Gq alpha subunit protein; Gα(s), Gs alpha subunit protein; Gα(i/o), Gi alpha subunit protein; IkB, inhibitory kappa B; IKK, IkB kinase; IP3, inositol 1,4,5-triphosphate; MEK1/2, mitogen-

activated protein kinase kinase; MKP1/2/3/4, mitogen-activated protein kinase phosphatases; MP1, MEK partner 1; NF-kB, nuclear factor-kappa B; p90RSK, p90 ribosomal S6 kinase; PIP2, phosphatidylinositol 4,5-bisphosphate; PIP3, phosphatidylinositol 3,4,5-triphosphate; PYK2, protein tyrosine kinase 2; Rap1a, Ras-related protein 1; Rap1GAP, Rap1-GTPase activating protein; RGS, regulator of G protein signaling; RasGAP, Ras GTPase activating protein; RasGRP, RAS guanyl nucleotide-releasing protein; SynGAP, synaptic Ras GTPase activating protein.

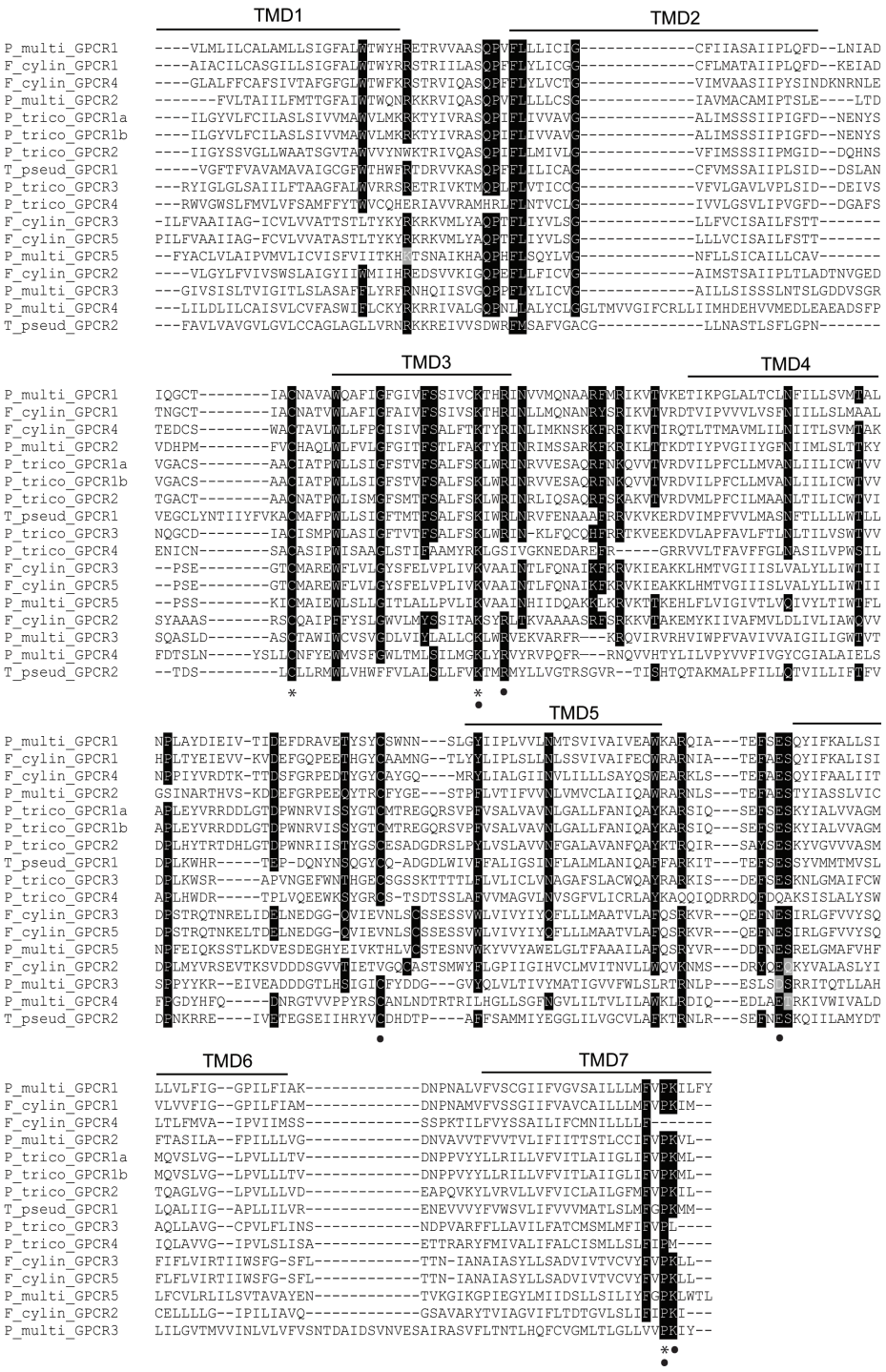

Figure 2.3. Multiple sequence alignment for the transmembrane domain (TMD) regions of the putative diatom GPCRs. Highlighted regions represent the most conserved residue positions among the diatom sequences. Asterisks denote residue positions conserved across all sequences and circles denote residue positions critical in family C GPCR activation or conserved across metazoan class C receptors.

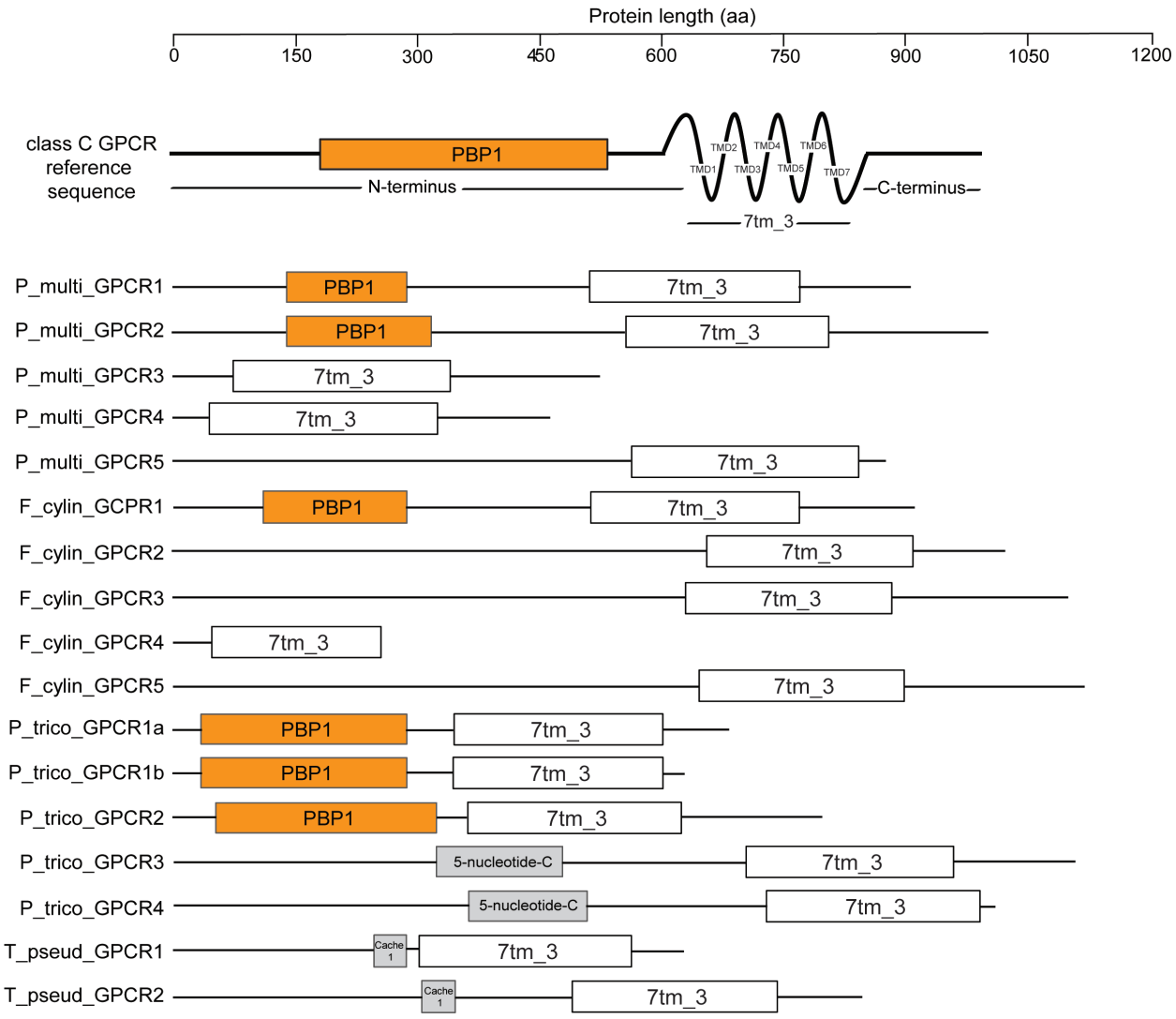

Figure 2.4. Structure and size of putative diatom GPCRs. A reference structure for a mammalian class C GPCR is provided for comparison. Conserved domains are boxed and include: PBP1, periplasmic binding protein domain which is considered to be the evolutionary origin of the class C GPCR ligand binding domain; 7tm_3, seven transmembrane domain of class C GPCRs; 5-nucleotide-C, associated with enzymatic degradation of sugars; Cache 1, extracellular protein domain involved in recognition of small molecules. TMD, transmembrane domain.

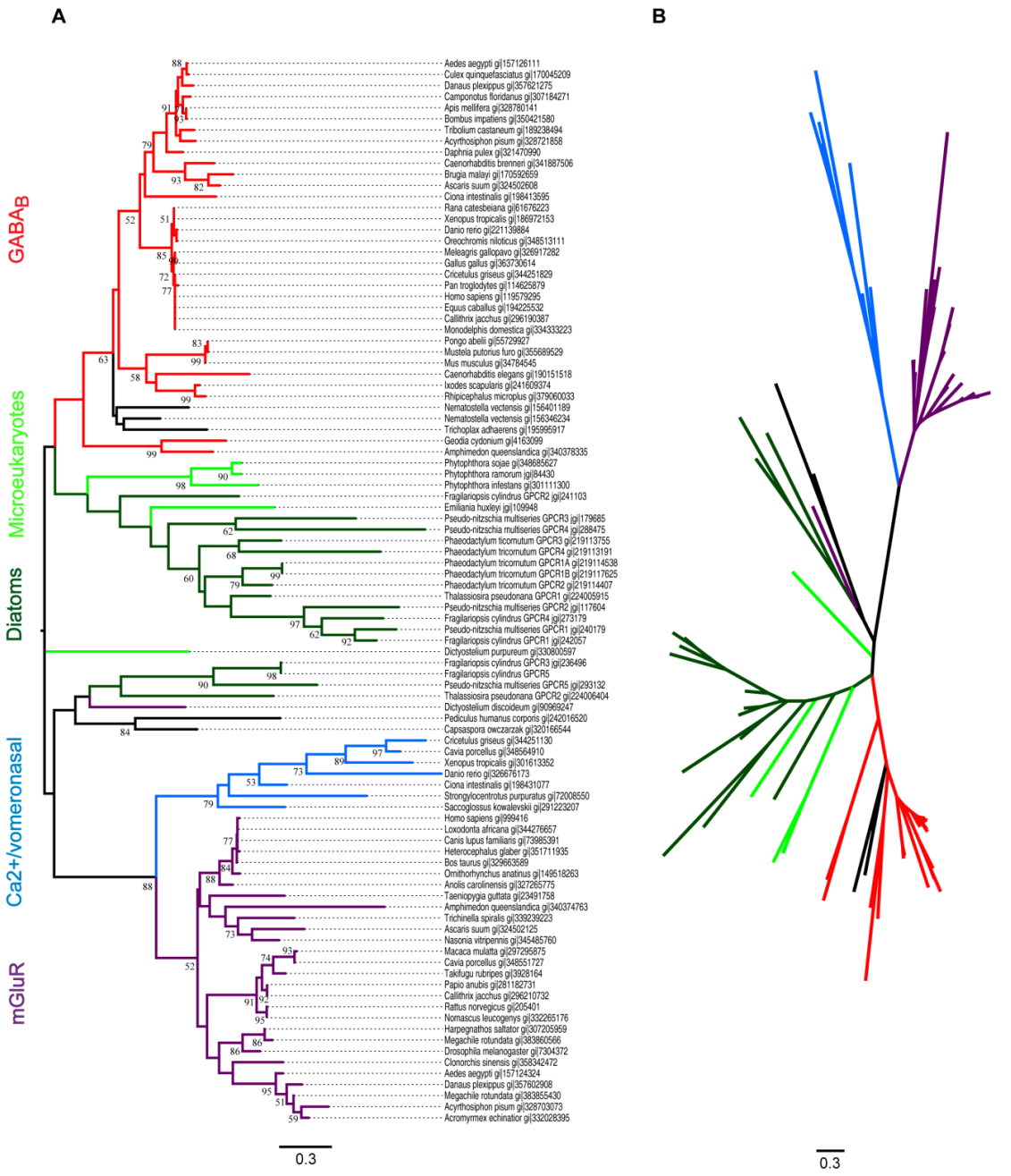

Figure 2.5. Maximum likelihood phylogenetic trees of the putative diatom GPCRs in relation to known class C GPCRs and microeukaryotic homologs. (A) Rooted and (B) unrooted trees. The unrooted tree highlights the overall genetic distance and relationships between branches. Class C GPCR subfamilies and putative diatom and microeukaryote GPCRs are color-coded: diatoms, dark green; microeukaryotes, light green; $GABA_B$, red; metabotropic glutamate, purple; calcium-sensing and vomeronasal, blue; hypothetical proteins, black. Bootstrap values ≥50 were included in the tree.

Table 2.1. Known functions of the major proteins involved in the mammalian G protein-coupled receptor (GPCR) signaling pathway. Numbered positions listed for each protein category refer to their respective locations in Figure 2.2, with the order of the numbering proceeding from left to right as one moves down the pathway. Protein abbreviations also correspond to those presented in Figure 2.2.

Protein category	**Function**
1) G protein-coupled receptor (GPCR)	Cell surface receptor; binds agonist/ligand, catalyzing exchange of GDP for GTP on G protein; dissociates and activates G protein subunits.
2) G protein (Gα,β,γ)	Heterotrimeric protein composed of α,β, and γ subunits; activated by GPCR to bind to and activate/deactivate various effectors (e.g. second messengers); amplifies receptor signal.
3) Adenylate cylcase (AC)	Transmembrane protein regulated by G protein; catalyzes formation of the second messenger cyclic adenosine monophosphate (cAMP) from ATP.
4) Guanylate cyclase (GC)	Transmembrane protein; catalyzes conversion of GTP to the second messenger cyclic guanosine monophosphate (cGMP); main receptor for nitric oxide.
5) GPCR kinase (GRK)	Regulates GPCR activity via phosphorylation; desensitizes the receptor signal.
6) Phospholipase A/C/D (PLC)	Catalyzes hydrolysis of phospholipids to generate the second messengers inositol 1,4,5-triphosphate (IP3) and diacylglycerol (DAG); amplifies signal by stimulating Ca^{2+} release and protein kinase activation.
7) Phosphodiesterase (PDE)	Degrades the phosphodiester bond in the second messengers cAMP and cGMP; terminates receptor signal.
8) Phosphoinositide-3 kinase (PI3K)	Recruited to the cell membrane following GPCR activation; binds G protein and initiates assembly of signaling complexes and priming of protein kinase cascades; hyperactivation of this pathway has been associated with cancer and diabetes.
9) Protein kinase C (PKC)	Regulates signal transduction; activated by G proteins or increases in cytosolic Ca^{2+}; phosphorylates a wide variety of proteins including small GTPases and MAPKs.
10) Ca2+/calmodulin protein kinase (CaMK)	Phosphorylates downstream CaM kinases; amplifies signaling cascade.
11) cAMP-dependent protein kinase (PKA)	Catalyzes transfer of phosphate from ATP to serine or threonine residues of effector proteins.
12) Mitogen activated protein kinases (MAPK+)	Large group of proteins forming the MAPK cascade; couple the receptor signal to transcriptional changes (e.g. activation of CREB); involved in a wide array of physiological processes (e.g. development, apoptosis, metabolism and stress response).
13) NF-kappa beta (NF-kB)	Transcription factor; translocates to nucleus and stimulates expression of genes involved in a wide variety of functions (e.g. immune and inflammatory responses, apoptosis).
14) cAMP response element binding protein (CREB)	Transcription factor; regulated by PKA phosphorylation and binding of cAMP response elements; activated by extracellular signals (e.g. neurotransmitters, hormones, growth factors).

Table 2.2. Microeukaryotes included in the custom database and their respective sources within genome and expressed sequence tag (EST) libraries.

Database type	Organism	Source
Genome	Aureococcus_anophageferrens	JGI[a]
Genome	Chlamydomonas reinhardii	JGI
Genome	Cyaniodioschyzon merolae	University of Tokyo
Genome	Dictyostelium discoideum	Dicty_predictions.stripped.protein.aa.fasta-cleaned
Genome	Ectocarpus siliculosus	JGI
Genome	Emiliania huxleyi	JGI
Genome	Fragilariopsis cylindrus	Fragilariopsis_cylindrus.chloroplast.scaffolds.fasta-cleaned.orfs
Genome	Fragilariopsis cylindrus	Fragilariopsis_cylindrus.main_genome.scaffolds.fasta-cleaned.orfs
Genome	Fragilariopsis cylindrus	Fragilariopsis_cylindrus.mitochondrion.scaffolds.fasta-cleaned.orfs
Genome	Micromonas sp. NOUM17	JGI
Genome	Micromonas pusilla CCMP1545	JGI
Genome	Naeglaria gruberi	JGI
Genome	Ostreococcus lucimarinus	JGI
Genome	Ostreococcus tarui	JGI
Genome	Phaeodactylum tricornutum	JGI
Genome	Phytopthora ramorum	JGI
Genome	Plasmodium falciparum	JGI
Genome	Plasmodium yoelii	JGI
Genome	Pseudo-nitzschia multiseries	Pseudo_nitzschia_multiseries.chloroplast.scaffolds.fasta-cleaned.orfs
Genome	Pseudo-nitzschia multiseries	Pseudo_nitzschia_multiseries.cluster_consensusEST.fasta-cleaned.orfs
Genome	Pseudo-nitzschia multiseries	Pseudo_nitzschia_multiseries.main_genome.scaffolds.fasta-idcleaned.orfs
Genome	Pseudo-nitzschia multiseries	Pseudo_nitzschia_multiseries.mitochondrion.scaffolds.fasta-cleaned.orfs
Genome	Tetrahymena_thermophila	TIGR[b]
Genome	Thalassiosira_pseudonana	JGI
Genome	Volvox carteri	JGI
EST	Thalassiosira_pseudonana	JGI
EST	Pseudo-nitzschia australis	P.australis-EST-ec121605-cleaned.tax_id.orfs
EST	Pseudo-nitzschia granii	This study
EST	Pseudo-nitzschia multiseries	Pseudo-nitzschia_multiseries.cluster_ consensus_EST.fasta-cleaned.orfs
EST	Chaetoceros	Chaetoceros_ESTs_FElim_FEreplete_McMurdo_isolate_isotigs_092710.fna-idcleaned.tax_id.orfs.fasta
EST	Geminigera_cryophila	Geminigera_cryophila_CCMP2564_ESTs.tax_id.orfs.fasta
EST	Guillardia_theta	Guillardia_theta_clusters.tax_id.orfs.fasta
EST	Phaeocystis antartica	p_antarctica_FElim_FErepleteo_isotigs_092710.fna-idcleaned.tax_id.orfs.fasta
EST	Phaeociystis globosa	p_globosa_colonies_and_single_cells_isotigs_092710.fna-idcleaned.tax_id.orfs.fasta
EST	Phaeociystis globosa	p_globosa_various_exp_treats_isotigs_092710.fna-idcleaned.tax_idd.orfs.fasta
EST	Phaeociystis globosa	p_globosa_viral_infections_isotigs_092710.fna-idcleaned.tax_id.d.orfs.fasta
EST	Haptophyte	NCBI[c]
EST	Copepod	NCBI
EST	Dinoflagellate	NCBI
RefSeq49	Plastid	NCBI
RefSeq49	Mitrochondial	NCBI
RefSeq49	Protist	NCBI
RefSeq49	Microbial	NCBI
RefSeq49	Viral	NCBI
RefSeq49	Plant	NCBI
Curated protein	GPCRDB[d]	http://www.gpcr.org/7tm/ (Vroling et al. 2011)

[a]Department of Energy Joint Genome Insitute

[b]Institute for Genomic Research

[c]National Center for Biotechnology Information

[d]G protein-Coupled Receptor Database

Table 2.3. Expressed sequence tag (EST) data for the putative diatom G protein-coupled receptors (GPCRs). ESTs for each diatom were accessed through the Joint Genome Institute (JGI) genome projects page and GenBank. N/A, not applicable.

Sequence	JGI protein ID / NCBI accession no.	Length (aa)	% EST coverage (≥99% identity)	No. of EST sequences	EST coverage region (aa) (≥99% identity)	Culture treatment	Reference
P. multiseries GPCR1	240179	906	45	1	499-906	Phosphate starvation	[1]JGI
GPCR2	117604	1003	17	1	616-819	Domoic acid-producing conditions	(Boissonneault et al. 2008)
GPCR3	179685	521	No hits	N/A	N/A	N/A	N/A
GPCR4	288475	461	No hits	N/A	N/A	N/A	N/A
GPCR5	293132	887	No hits	N/A	N/A	N/A	N/A
F. cylindrus GPCR1	242057	911	No hits	N/A	N/A	N/A	N/A
GPCR2	241103	1020	No hits	N/A	N/A	N/A	N/A
GPCR3	236496	1102	No hits	N/A	N/A	N/A	N/A
GPCR4	273179	305	100	2	1-305	Osmotic stress, pooled RNA from 5 treatments	(Krell and Gloeckner 2004)
GPCR5	253555	257	No hits	N/A	N/A	N/A	N/A
P. tricornutum GPCR1A	219114538	688	100	64	1-688	16 different treatments	(Bowler et al. 2008)
GPCR1B	219117625	626	100	64	1-619	16 different treatments	(Bowler et al. 2008)
GPCR2	219114407	799	69	8	1-552	16 different treatments	(Bowler et al. 2008)
GPCR3	219113755	1111	86	22	1-228/ 378-1111	16 different treatments	(Bowler et al. 2008)
GPCR4	219113191	1012	74	18	1-125/ 366-986	16 different treatments	(Bowler et al. 2008)
T. pseudonana GPCR1	224005915	627	No hits	N/A	N/A	[2]Upregulated during silaffin-like response	(Shrestha et al. 2012)
GPCR2	224006404	846	26	1	570-792	Iron-limited cells	[3]JGI

[1]http://genome.jgi-psf.org/Psemu1/Psemu1.home.html
[2]T. pseudonana GPCR1 had no corresponding EST data but the sequence was found to be upregulated during transcriptome profiling under silicon starvation conditions.
[3]http://genome.jgi-psf.org/Thaps3/Thaps3.home.html

Table 2.4. Best predicted protein hits in diatom and green algae/land plant genomes to human G protein-coupled receptor (GPCR) signaling pathway proteins. Data presented as: e-value | percent identity | alignment length/human protein length.

		Genome Best hit				
Protein Category		***Pseudo-nitzschia multiseries***	***Fragilariopsis cylindrus***	***Thalassiosira pseudonana***	***Phaeodactylum tricornutum***	***Viridiplantae*** [a]
G protein-coupled receptor (GPCR): Class C family		10^{-16}|24%|343/941	10^{-20}|21%|745/941	10^{-17}|26%|330/883	10^{-24}|25%|272/883	No hits[b]
G protein subunit	α	10^{-47}|47%|197/354	10^{-42}|45%|202/354	10^{-79}|43%|357/354	10^{-67}|36%|358/355	10^{-63}|39%|371/355
	β	10^{-23}|31%|262/340	10^{-53}|37%|327/340	10^{-108}|55%|336/340	10^{-69}|41%|350/340	10^{-84}|49%|312/340
	γ	Not present at $<10^{-5}$	Not present at $<10^{-5}$	Not present at $<10^{-5}$	Not present at $<10^{-5}$	Present in *Arabidopsis*
Adenylate cyclase		10^{-23}|33%|249/1119	10^{-21}|30%|200/1353	10^{-31}|34%|241/1119	10^{-27}|30%|572/1610	Present in *Arabidopsis*
Guanylate cyclase		10^{-49}|48%|212/1108	10^{-29}|36%|242/1073	10^{-44}|46%|218/1108	10^{-34}|38%|218/1073	10^{-41}|42%|228/1032
GPCR kinase		10^{-54}|36%|322/590	10^{-42}|31%|353/590	10^{-49}|36%|311/590	10^{-55}|40|298/590	10^{-54}|38|299/590
Phospholipase A/C/D		10^{-67}|32%|534/762	10^{-75}|35%|510/762	10^{-74}|37%|457/762	10^{-76}|35%|492/762	10^{-65}|32%|582/762
Phosphodiesterase		10^{-52}|38%|278/809	10^{-56}|39%|306/673	10^{-52}|38%|283/673	10^{-39}|29%|417/609	10^{-57}|42%|293/673
Phosphoinositide-3 kinase		10^{-23}|25%|403/1102	10^{-35}| 35%|281/1358	10^{-50}|27%|575/1102	10^{-57}|25%|838/1358	10^{-61}|30%|542/1102
Protein kinase C		10^{-81}|44%|329/671	10^{-75}|46%|287/671	10^{-80}|48%|294/671	10^{-84}|46%|331/671	10^{-81}|50%|329/671
Ca2+/calmodulin protein kinase		10^{-69}|40%|334/478	10^{-72}|43%|303/385	10^{-70}|49%|261/476	10^{-70}|41%|317/489	10^{-64}|43%|305/489
cAMP-dependent protein kinase (PKA)		10^{-76}|44%|293/351	10^{-72}|44%|279/351	10^{-92}|52%|297/398	10^{-74}|43%|293/351	10^{-80}|45%|331/351
Mitogen-activated protein kinase cascade	MAPK	10^{-97}|57%|237/816	10^{-98}|58%|230/816	10^{-97}|51%|352/379	10^{-101}|50%|345/816	10^{-107}|56%|332/816
	MAPKK	10^{-32}|33%|240/448	10^{-30}|34%|243/448	10^{-39}|38%|261/448	10^{-39}| 30%| 320/400	10^{-62}|37%|321/400
	MAPKKK	10^{-42}|38%|267/1313	10^{-42}|32%|310/1512	10^{-141}|38%|277/954	10^{-45}|43%|273/619	10^{-70}| 48%|281/619
	MAPKKKK	10^{-35}|35%|276/846	10^{-45}|31%|403/1320	10^{-52}|40%|266/894	10^{-55}|40%|288/894	10^{-69}|44%|315/894
NF-kappa beta		10^{-8}|25%|265/317	10^{-13}|31%|212/900	10^{-8}|31%|212/900	10^{-10}|32%|200/900	Not present at $<10^{-5}$
cAMP response binding element (CREB)		Not present at $<10^{-5}$	Not present at $<10^{-5}$	Not present at $<10^{-5}$	Not present at $<10^{-5}$	Best hit to *Zea mays* transcription factor

[a]Land plants and green algae

[b]Predicted class C family GPCR domains but most similar to iGluRs

CHAPTER 3: Metagenomic profiling of microbial composition and antibiotic resistance determinants in Puget Sound

This chapter has been published in its entirety:

Port JA, Wallace JC, Griffith WC, Faustman EM. 2012. Metagenomic profiling of microbial community composition and antibiotic resistance determinants in Puget Sound. PLoS ONE 7(10): e48000.

3.1 Abstract

Human-health relevant impacts on marine ecosystems are increasing on both spatial and temporal scales. Traditional indicators for environmental health monitoring and microbial risk assessment have relied primarily on single species analyses and have provided only limited spatial and temporal information. More high-throughput, broad-scale approaches to evaluate these impacts are therefore needed to provide a platform for informing public health. This study uses shotgun metagenomics to survey the taxonomic composition and antibiotic resistance determinant content of surface water bacterial communities in the Puget Sound estuary. Metagenomic DNA was collected at six sites in Puget Sound in addition to one wastewater treatment plant (WWTP) that discharges into the Sound and pyrosequenced. A total of ~550 Mbp (1.4 million reads) were obtained, 22 Mbp of which could be assembled into contigs. While the taxonomic and resistance determinant profiles across the open Sound samples were similar, unique signatures were identified when comparing these profiles across the open Sound, a nearshore marina and WWTP effluent. The open Sound was dominated by α-Proteobacteria (in particular *Rhodobacterales sp.*), γ-Proteobacteria and Bacteroidetes while the marina and

effluent had increased abundances of Actinobacteria, β-Proteobacteria and Firmicutes. There was a significant increase in the antibiotic resistance gene signal from the open Sound to marina to WWTP effluent, suggestive of a potential link to human impacts. Mobile genetic elements associated with environmental and pathogenic bacteria were also differentially abundant across the samples. This study is the first comparative metagenomic survey of Puget Sound and provides baseline data for further assessments of community composition and antibiotic resistance determinants in the environment using next generation sequencing technologies. In addition, these genomic signals of potential human impact can be used to guide initial public health monitoring as well as more targeted and functionally-based investigations.

3.2 Introduction

Coastal ecosystems continue to be subjected to increasing human pressures. Over 40% of the global population currently lives within 100 km of coastlines, and this percentage continues to increase (United Nations Environment Programme (UNEP) 2006). Human impacts have led to significant degradation in marine environments in the form of habitat loss, eutrophication, hypoxia, organic and inorganic pollution, invasive species, pathogen spread and ocean acidification, among others (Crain et al. 2009). Based on a spatial analysis of 17 anthropogenic drivers, it has been estimated that over 40% of the world's oceans have medium to high impacts associated with human activities (Halpern et al. 2008). These impacts subsequently pose risks to human health in the form of bacterial and viral pathogens, harmful algal blooms, contaminated seafood and decreased well-being (NRC 1999; Fleming et al. 2006; Kite-Powell et al. 2008; Laws 2008). There is thus considerable interest in investigating the distribution and magnitude of these impacts in marine environments, including nearshore, coastal and open ocean locations.

Assessing human impacts on such a global scale is a challenge, and current approaches to environmental monitoring are not well suited for large-scale spatial and temporal analyses. This is changing with advancements in the field of environmental genomics. Metagenomics, in tandem with next generation sequencing, now provides a technical means by which to monitor environmental microbial communities in a high-throughput manner. Metagenomics refers to the sequencing of DNA directly from environmental samples (i.e. metagenomes), and as such simultaneously provides access to the genetic information from mixed environmental microbial communities (Dinsdale et al. 2008). With this approach, taxonomic and functional microbial diversity can be profiled and community change subsequently monitored over space and time in response to environmental or anthropogenic impacts relevant to human health (Nogales et al. 2011). Microbial community composition has been shown to change in differentially impacted coastal environments (Nogales et al. 2007; Aguilo-Ferretjans et al. 2008; Wu et al. 2010) and other anthropogenically impacted environments (Dinsdale et al. 2008; Tringe et al. 2008; Sanapareddy et al. 2009; Hemme et al. 2010) and thus provides a potential indicator of community stressors and stress response.

Despite the potential role aquatic ecosystems may play in harboring and disseminating antibiotic resistance, little is known regarding the distribution of antibiotic resistance determinants within marine microbial communities. Antibiotic resistance continues to be a serious public health concern, as the overuse and misuse of antibiotics has led to the selection of antibiotic resistance genes in bacterial populations and has thus compromised our ability to treat bacterial infections. Coastal environments are subject to a multitude of anthropogenic impacts such as sewage, hospital waste, agricultural runoff and aquaculture that increase the prevalence of antibiotic resistance determinants in microbial communities, and thus these ecosystems have

the potential to be important sources and sinks for antibiotic resistant bacteria and genes and may contribute to the dissemination of resistance determinants across organisms (Baquero et al. 2008; Taylor et al. 2011). In fact, many resistance genes found in pathogenic bacteria have evolved or are sourced from resistance genes found in environmental microbial communities (Martinez 2009). Clinical resistance in many cases is therefore the result of horizontal transfer of resistance genes via mobile genetic elements from ecologically and taxonomically distant bacteria (Aminov and Mackie 2007), and this gene transfer has been shown to occur across a wide range of bacterial species, including pathogens to non-pathogens and vice versa (Salyers and Amabile-Cuevas 1997; Salyers et al. 2004; Allen et al. 2010). Furthermore, the marine environment in particular has been shown to have a high rate of horizontal gene transfer (McDaniel et al. 2010).

The majority of studies assessing the prevalence and distribution of antibiotic resistance genes in natural environments to this point have focused on specific known resistance genes using PCR-based techniques (Soge et al. 2009; Storteboom et al. 2010; D'Costa et al. 2011). While these techniques provide resistance profiles for specific organisms or targeted genes, they are limited for ecosystem level applications because of these gene-by-gene and single species approaches. Thus there is limited knowledge regarding the global prevalence and distribution of antibiotic resistance genes and other resistance determinants including mobile genetic elements in natural microbial communities, and this data gap has been exacerbated by our inability to access the genomic information for unculturable bacteria, which represent >99% of bacteria in the environment (Amann et al. 1995). Metagenomics now offers a culture-independent and high-throughput approach to investigate antibiotic resistance determinants in the environment at the microbial community-level, and a number of studies have already successfully employed this

approach (Riesenfeld et al. 2004; Szczepanowski et al. 2008; Allen et al. 2009; Sommer et al. 2009; Donato et al. 2010; Kristiansson 2011; Torres-Cortes et al. 2011).

This study uses a metagenomic approach in combination with next generation sequencing to evaluate potential differences in community composition and antibiotic resistance gene signals across a natural ecosystem using the case study of Puget Sound, Washington. Puget Sound is a partially mixed fjord-like estuary that is connected to the Pacific Ocean via the Strait of Juan de Fuca and that receives the majority of its freshwater input from rivers emptying into the Whidbey Basin region (Figure 3.1) (Babson et al. 2006). There are 96 publicly owned wastewater treatment plants (WWTP) in the Puget Sound Basin that serve over 3.5 million people from urban population centers such as Seattle, Tacoma, Olympia and Everett and process over 124 million gallons of sewage per day (Washington State Department of Ecology 2010). Antibiotics, heavy metals and other pollutants also enter the Sound through agricultural and urban runoff. At the same time, Puget Sound supports one of the largest shellfish industries in the country, with over 100 commercial growing areas covering 900 miles of shoreline (Washington State Department of Health 2011). The potential for these anthropogenic impacts to alter microbial community and resistance determinant composition in Puget Sound is uncertain, and baseline data is needed in order to spatially and temporally monitor these potentially health relevant human impacts.

This is the first baseline survey of the taxonomic and resistance potential of microbial communities in Puget Sound. Using high-throughput and sequence-based comparative metagenomics, we identified common taxonomic and antibiotic resistance determinant signatures for the open Puget Sound locations. Comparison of these metagenomic profiles between the open Sound, a nearshore marina and effluent from a proximal WWTP revealed unique profiles across

the different environments that follow a gradient of human impact. This investigation provides an initial framework by which to monitor the marine environment for genomic determinants relevant to public health.

3.3 Methods

3.3.1 Sample collection

Surface water samples (~5m depth) were collected aboard the R/V Thomas Thompson from October 29-30, 2010 at five stations in Puget Sound, WA (Figure 3.1). These stations have been monitored since 1998 by the University of Washington Puget Sound Regional Synthesis Synthesis Model (PRISM) program. A nearshore surface water sample was also collected at Shilshole Bay Marina in Seattle, WA on December 20, 2010 and an effluent sample from the King County West Point Treatment Plant (WWTP) in Seattle, WA was collected on January 31, 2011. The Marina is located approximately 8 miles north of central Seattle and 2 miles north of the WWTP, and includes 1,400 moorage slips (300 slips for liveaboards), public water access and a public promenade. The West Point Treatment Plant has an average daily inflow of 133 million gallons, and the wastewater is sourced from stormwater/groundwater (53%), residential (29%), commercial (17%) and industrial (1%) processes (County 2011). At each station and sampling location, 80 liters of water were pumped through a peristaltic pump system (Cole-Palmer, U.S.A.) and filtered sequentially (i.e. size fractionated) through 3-µm (147mm) polycarbonate membranes (Sterlitech, WA) and 0.2-µm (147mm) Supor membranes (Pall Corporation, U.S.A). Filters were covered with sucrose lysis buffer (50 mMTris•HCl, 40 mM EDTA, and 0.75 M sucrose) and stored at -80°C on board and then transferred to -80°C in the laboratory. For the cruise samples, physicochemical conditions (e.g. temperature, salinity,

oxygen, fluorescence) were measured using a conductivity-temperature-depth (CTD) sensor array mounted on a Niskin bottle rosette (Table 3.1).

3.3.2. DNA extraction and sequencing

DNA was extracted from the 0.2-μm filters as these filters contain the bacterial fraction of the marine community. The filters were thawed and cut into eighths. Total metagenomic DNA was extracted using the Powerwater DNA Isolation Kit (Mo Bio Laboratories, CA) following the manufacturer's protocol with modifications. This kit is specifically designed for isolating DNA from filtered environmental water samples and includes inhibitor removal technology aimed at removing humic acid and other organic matter commonly found in environmental samples that can interfere with downstream analyses. The protocol modifications included an incubation at 65°C for 10 minutes following addition of Solution P1 and a lysozyme digestion (final concentration = 10mg/ml) at 37°C for 1 hour and an RNase digestion at 37°C for 20 minutes immediately following the bead-beating step. DNA quantity and quality were estimated using the Quant-iTPicogreen dsDNA Assay Kit (Invitrogen, U.S.A.) and the Nanodrop-1000 Spectrophotometer (Nanodrop Technologies, DE). DNA concentrations ranged from 75 to 130 ng/μl per sample. Gel electrophoresis showed high molecular weight DNA fragments 3-12 kb in size. Approximately 3μg of total DNA for each sample was used for a multiplexed pyrosequencing run on the Roche/454 GS FLX Titanium platform in the Department of Microbiology at the University of Washington.

3.3.3. Bioinformatic analyses

The 454 sequence reads were initially trimmed to remove barcoded sequences and any secondary adapter sequences. Reads containing <5% of any one nucleotide were also removed. For each sample, raw reads were assembled into contigs using Newbler v. 2.5.3 (Roche Diagnostics-454 Life Sciences) with default settings except for a more stringent 95% minimum overlap identity. For taxonomic annotation, the unassembled reads were run through the Meta Genome Rapid Annotation using Subsystems Technology (MG-RAST) pipeline (Meyer et al. 2008). This pipeline includes QC steps to eliminate low quality sequences. Reads meeting any of the following three criteria were omitted: read length > 2 standard deviations from the mean sample read length, read containing > 5 ambiguous bases or reads identical to another sequence over the first 50 bp. Unassembled reads were taxonomically annotated using the lowest common ancestor (LCA) algorithm and only sequence matches with ≥50% identity over ≥50 amino acids were retained. This algorithm assigns each read to the LCA within the set of matching taxa from the BLASTX search (Huson et al. 2007). For example, if a given read had sequence similarity to 3 different families within the same order, the read is assigned at the order level. The LCA algorithm is thus predicted to have a lower rate of false positive assignments than the best hit BLAST approach but with a higher number of unspecific assignments or no hits. Genus level annotation was performed using 16S rRNA analysis. An *Escherichia coli* 16S rRNA reference sequence was searched against the 454 reads using BLASTN (default settings) and matching 454 reads were run through the Ribosomal Database Project (RDP) Classifier (Wang et al. 2007) to classify the 16S rRNA sequences. Only those classifications with at least a 50% confidence estimate were included in the analysis. Additional environmental metagenomes used for taxonomic comparison to the Puget Sound dataset included: Sargasso Sea (GS001a, GS001b and GS001c), Chesapeake Bay (GS012) and the Gulf of Mexico (GS016) from the Global Ocean

Sampling Expedition (Rusch et al. 2007), Lake Lanier (Oh et al. 2011) and farm soil (Tringe et al. 2005). All the water samples were collected in 1-5 m depth.

Read recruitment to alpha proteobacterium HTCC2255 was performed using BLASTN (E-value $<10^{-5}$). The 2.23 Mbp genomic scaffold (GI number: 211594581) was downloaded from GenBank. The recruitment plot was generated using in-house custom visualization software.

To identify putative antibiotic resistance genes, an expanded version of the Antibiotic Resistance Gene Database (ARDB) (Liu and Pop 2009) was generated that incorporates all sequences deposited in GenBank to date that are ≥80% identical to sequences already in the ARDB (Figure 3.2). The ARDB is a commonly used database consisting of over 23,000 resistance genes from nearly 1,700 species. A nonredundant version of the ARDB was first compiled and consisted of 3,185 sequences. An expanded dataset, termed the ARDB+, was generated by searching the sequences from this nonredundant list against GenBank using the 80% identity cutoff (Liu and Pop 2009). The ARDB+ increases the number of nonredundant ARDB sequences to a total of 11,500 sequences. A list of 103 antibiotic resistance gene sequences from metagenomic samples that were functionally verified to confer resistance (Allen et al. 2009; Sommer et al. 2009; Donato et al. 2010; Torres-Cortes et al. 2011) was also compiled and included in the ARDB+. MetaGeneMark (Zhu et al. 2010) (default settings) was used to predict potential protein coding sequences for the unassembled Puget Sound and WWTP reads, and these predicted peptides were then BLASTP searched (Camacho et al. 2009) against the ARDB+ (E-value $<10^{-5}$). Reads with ≥80% identity over an alignment length ≥50 amino acids to a resistance gene within the ARDB+ were annotated according to the best hit.

Plasmid sequences were annotated by aligning the unassembled reads and contigs against the plasmid sequences available in the NCBI RefSeq database (1843 sequences) using BLASTN

(E-value $<10^{-5}$). Only those sequences with a nucleotide sequence identity ≥95% over an alignment length of at least 100 bp were retained (Kristiansson 2011; Zhang et al. 2011).

Two approaches were used to identify transposable elements in the dataset. First, 167 unique genes annotated as transposable elements were downloaded from GenBank. Unassembled reads were then compared against these sequences using BLAST, and reads with a sequence identity ≥80% (Kristiansson 2011) over an alignment length of at least 150 bp were annotated as transposable elements. Second, unassembled reads were compared to 41 protein motifs annotated as transposable elements within the Pfam 26.0 database (Punta et al. 2012) using an E-value cutoff of 10^{-10}. Pfam is a database of conserved protein families and domains across species, where the domains represent structural and functional motifs of the protein. A significance threshold of $E<10^{-10}$ represents a more stringent cutoff than the manually curated Pfam gathering threshold or trusted cutoff, and thus minimizes the number of false positive assignments. The abundance of transposable elements was normalized to total sequence reads per sample for the BLAST approach and to total motif counts per sample for the Pfam approach.

3.3.4. Statistical analyses

Overrepresentation or underrepresentation of taxonomic units (domain, phylum or class) was determined using the Statistical Analysis of Metagenomic Profiles (STAMP) software package (Parks and Beiko 2010). Only taxonomic units that had an effect size > 1 between two metagenomes were included in the statistical analysis. P-values were calculated using the two-sided Fisher's Exact test, and 95% confidence intervals were calculated with the Newcombe-Wilson method. The Benjamini-Hochberg FDR method (Benjamini and Hochberg 1995) was used to correct for the false discovery rate. Hierarchical clustering of metagenomes by taxonomic

units using the Euclidian distance metric was performed with the TIGR Multiexperiment Viewer (TMeV) (http://www.tm4.org/mev/). Counts were normalized to total taxonomic counts to give a measure of relative abundance within a metagenome. Significance tests were run to determine whether there were statistical differences in the abundance counts across samples. Data was analyzed using a generalized linear model for binomially distributed data. Differences among open Sound samples were treated as a source of extra-binomial variability in the statistical model to provide estimates of variability for the open Sound, Marina and WWTP samples in addition to the simple binomial variability based on the counts (McCullagh and Nelder 1989).

3.3.5. Metagenome sequence accession

Sequence data for the Puget Sound dataset is available through the Metagenomics RAST (MG-RAST) server (http://metagenomics.anl.gov/) under the MG-RAST identification numbers 4460178.3, 4460179.3, 4460180.3, 4460182.3, 4460188.3, 4460189.3 and 4460190.3.

3.4 Results and Discussion

Shotgun metagenomics in combination with pyrosequencing were used to explore the taxonomic and antibiotic resistance determinant composition at six locations from north to south Puget Sound in addition to an effluent sample from a WWTP that deposits into the main basin of Puget Sound. The environmental data and summary statistics for the pyrosequencing run are shown in Table 3.1. Approximately 1.4 million reads were generated that passed the QC filter, comprising nearly 530 million base pairs of sequence data. There were 22 Mbp of assembled contigs greater than 500 bp and 47 contigs greater than 10,000 bp. The N50 contig size was 1.1 kb.

3.4.1. A taxonomic signature of Puget Sound

Approximately 42% of the unassembled sequence reads were assigned by phylum to protein sequences within GenBank using our sequence similarity criteria of ≥50% identity and an alignment length ≥50 aa. Bacteria accounted for 93% of all annotated sequence reads, with Archaea (4%), Eukaryota (2%) and viruses (1%) comprising the remaining proportion. All communities were dominated by the phylum Proteobacteria, which accounted for 49-67% of all sequence reads (Figure 3.3). As shown by the positive and negative bars in Figure 3.4, the open Sound had an overrepresentation of α- and γ-Proteobacteria and Bacteriodetes and a decreased proportion of β-Proteobacteria, Actinobacteria and Firmicutes when compared to the Marina, WWTP and other metagenomes (p-value $<10^{-15}$ for all comparisons). The Marina exhibited a similar pattern of taxon overrepresentation when compared to the WWTP except on a smaller magnitude. The dominance of α-Proteobacteria and γ-Proteobacteria and low abundance of β-Proteobacteria in the open Sound samples has also been characterized in other open and coastal ocean locations (von Mering et al. 2007; Zinger et al. 2011), but α- and γ-Proteobacteria appear to represent an even larger proportion of total taxa in the open Sound samples than the other marine metagenomes included in this study except for α-Proteobacteria in the Gulf of Mexico.

The overrepresentation of α-Proteobacteria in the open Sound was due to a high abundance of *Rhodobacterales sp.* (Figure 3.4 and Figure 3.5). A total of 47,425 sequence reads from the open Sound and Marina were recruited to 92% (7x coverage) of the *Rhodobacterales bacterium* HTCC2255 reference genome assembly with a minimum 95% identity per read (Figure 3.6). This genome was also previously assembled from a metagenomic sample taken from Puget Sound (Iverson et al. 2012). The high abundance of *Rhodobacterales* is in contrast to

the other marine metagenomes where *Rickettsiales sp.*, another group of α-Proteobacteria that includes the ubiquitous SAR11 clade, was the dominant organism (Figure 3.5). Recent metatranscriptomic samples collected from Monterey Bay, CA also contained a high abundance of *Rhodobacterales sp.* HTCC2255 and a large *Rhodobacterales*: *Rickettsiales* ratio (Ottesen et al. 2011), suggesting *Rhodobacterales* may be a dominant organism in surface waters on the U.S. West coast. Members of the phylum Bacteroidetes have also been shown to be widespread in marine environments, and the increased proportion of this taxon in the open Sound relative to the other metagenomes is attributable to an overrepresentation of *Flavobacteriales sp.* (Figure 3.4). The proportion of Actinobacteria significantly increased from ~2% of reads in the open Sound to 13% in the Marina and 25% in the WWTP sample (p-value $<10^{-15}$ for all comparisons). Actinobacteria was also overrepresented in the other freshwater and freshwater-impacted environments, and further analysis revealed a strong negative correlation between Actinobacteria abundance and salinity ($r = -0.952$; $p<10^{-6}$) (Figure 3.7). Actinobacteria in coastal areas has previously been suggested to be a potential signal of terrestrial or freshwater runoff (Kelly and Chistoserdov 2001; Aguilo-Ferretjans et al. 2008), and our results support this notion. α-Proteobacteria ($r = 0.904$; $p<10^{-4}$), β-Proteobacteria ($r = -0.913$; $p<10^{-5}$), γ-Proteobacteria ($r = 0.860$; $p<10^{-4}$), Bacteroidetes ($r = 0.612$; $p<0.05$) and Firmicutes ($r = -0.644$; $p<0.05$) were also correlated with salinity, while cyanobacteria abundance was correlated with temperature ($r = 0.804$; $p<0.005$). The GC-content distribution also reveals the taxonomic similarity between the open Sound samples and the differences in community composition between the open Sound, Marina and WWTP effluent (Table 3.1). The open Sound had a mean %GC of 43.62%±0.40, with a slight increase for the Marina sample (46.8%). The GC-content of the WWTP sample was significantly higher than for all other samples (56.4%).

Approximately 3.3% of sequence reads were classified at the genus level by the RDP. Rarefaction analysis at the genus level shows that not all the taxonomic richness was accounted in the marine and WWTP samples (Figure 3.8). Still, the rarefaction curves were leveling off at 60% of the metagenomes sizes, indicating repeated sampling of the same taxa and thus coverage of the most abundant taxa. Open Sound samples were characterized by a high abundance of *Candidatus Pelagibacter* (SAR11 clade) and *Thalassobacter* (family Rhodobacteraceae) from the phylum α-Proteobacteria and *Polaribacter* from Bacteroidetes. The Marina was dominated by the γ-proteobacterium *Glaciecola*, accounting for 28.3% of genera diversity. *Glaciecola* was absent in all other samples except P5 where it comprised 0.7% of genera diversity. The abundance of bacteria from the genus Acinetobacter, which has been linked to hydrocarbon degradation in marine environments and has been shown to be seasonally prevalent in a coastal marina (Aguilo-Ferretjans et al. 2008), was not significantly different between the Marina (0.1%) and open Sound (~0.03-0.07%) but was considerably elevated in the WWTP (1.07%). The WWTP effluent also contained more specific taxonomic markers of potential human impact. Firmicutes, which is one of the dominant phyla of the human gut (Turnbaugh et al. 2007), accounted for a significantly greater proportion (8.4%) of all WWTP phyla than in the other samples except farm soil, and was also associated with a number of the WWTP putative antibiotic resistance gene sequences. *Clostridiales sp.* was the most abundant group within Firmicutes, members of which are commonly found in human or animal feces (Liu et al. 2008; Cook et al. 2010). *Blautia* is a core human gut microbe (Claesson et al. 2009), and *Hespellia* is a member of the family *Lachnospiraceae* within which phylotypes have been identified as genetic markers of human fecal contamination (Newton et al. 2011). Other taxonomic surveys of WWTP environments have also found an increased abundance of Firmicutes (McLellan et al. 2010;

Albertsen et al. 2012), and one recent study used the ratio of Firmicutes and Bacteroidetes to α-Proteobacteria abundance as an indicator of fecal pollution in watersheds (Wu et al. 2010). Thus the overabundance of Firmicutes within the WWTP distinguishes this environment from the other marine and freshwater metagenomes, and may suggest the use of this phylum as a potential indicator of human impact.

Despite the spatial variation between the open Sound samples, the taxonomic composition across these surface water samples was strikingly similar and can likely be attributed to similar salinity and temperature gradients. Salinity has been shown to be one of, if not the most, important environmental parameters determining the level of similarity between isolated microbial communities (Lozupone and Knight 2007). This is evidenced by the fact that the samples cluster by salinity (Figure 3.3). The open Sound samples cluster with other more saline environments including the Sargasso Sea (open ocean) and the Gulf of Mexico (coastal ocean), while the Marina and WWTP effluent cluster within a larger clade that contains other freshwater and freshwater-impacted metagenomes (Lake Lanier, Chesapeake Bay and farm soil). The freshwater clade can be described by a high proportion of Actinobacteria, an increased abundance of β-Proteobacteria and an overall increased level of diversity. The metagenomes composing this freshwater clade can also be defined as being human impacted environments, and while it is not possible to further tease apart the potential influence freshwater input versus human impact has on the taxonomic profiles, it is important to note that the two are related in that environmental pollutants can enter marine ecosystems via freshwater runoff and thus a freshwater signal may in some cases be a potential indicator of human impact.

It is also important to note the potential impact temporal differences may have had on community composition. Seasonality has been shown to be a strong driver of marine microbial

composition (Gilbert et al. 2012). While the open Sound samples were collected within a span of two days at the end of October, the Marina and effluent samples were collected approximately 7 and 12 weeks later respectively. By nature of the fact that the freshwater effluent sample is already taxonomically distinct from the marine samples due to salinity and source, temporal variability is not likely a primary explanatory variable for describing the differences in community composition. Seasonal influences are more relevant when comparing the open Sound to the Marina. Freshwater discharges from rivers into Puget Sound are at a maximum in late autumn due to rainfall runoff and this flow continues at elevated levels through the winter (Babson et al. 2006). The lower salinity of the December Marina sample (~24 ppt) may therefore correspond to increased freshwater inputs. Subsequently the community composition of the Marina sample, which includes a higher proportion of the potentially freshwater or terrestrial-sourced Actinobacteria, may reflect these seasonal differences. The temperatures of the Marina and open Sound samples were similar despite the two month lag (Table 3.1). More temporal monitoring data is needed to better explain the role seasonality has on bacterial composition at these locations.

3.4.2. Proportion of putative antibiotic resistance gene sequences differs across metagenomes

In order to determine the abundance of putative antibiotic resistance genes in the Puget Sound and WWTP datasets, the sequence data was searched against an expanded and nonredundant ARDB (termed ARDB+) which also included 103 functional antibiotic resistance genes sequences isolated from metagenomic projects. Approximately 0.0013% (n=18) of all sequence reads had ≥80% identity and an alignment length ≥50 aa to a sequence within the ARDB+. To validate this annotation, the 18 sequences were also searched against GenBank, and

results indicated that the best protein hits within GenBank were also to antibiotic resistance genes. Of these 18 sequences, the WWTP effluent had the highest representation (n=14), followed by the Marina (n=4) and the open Sound (n=0) (Figure 3.9 and Table 3.2). For comparison, the proportion of putative resistance gene sequences in the WWTP effluent (0.012%) was lower than that for other pyrosequenced metagenomes from different environmental matrices including a river sediment sample taken downstream from an Indian WWTP that processes large quantities of drugs (1.4%) (Kristiansson 2011) and the plasmid fraction of a WWTP activated sludge sample (1.1%) (Szczepanowski et al. 2008), but slightly greater than that of an activated sludge metagenome from another municipal WWTP (0.008%) [15]. Tetracycline resistance genes had the highest representation (33%) within the 18 putative antibiotic resistance gene sequences found in our samples. The WWTP putative antibiotic resistance gene sequences had similarity to those from clinically relevant bacteria species including *Clostridium*, *Enterococcus* and *Streptococcus*, while resistance gene sequences from the Marina were similar to those from environmental bacteria including *Desulfococcus*, *Polaromonas* and *Pseudomonas*. The average percent protein coverage based on the sequence alignment of the 454 reads to ARDB+ sequences was 33.8%±18.9 (mean±SD).

There was a significant increase in the proportion of putative resistance gene sequences from the open Sound to Marina to WWTP (Figure 3.9), suggestive of a relationship between resistance gene abundance and human impact across these environments. The fact that a differential resistance gene signal could be detected across the different environments supports using a pyrosequencing approach to identify trends in resistance gene abundance across diverse environments. Challenges still remain though in using this approach for quantification of resistance genes in metagenomes. First, the low sequencing depth per sample likely limits our

ability to detect the full resistance gene repertoire of complex microbial community samples. Depth of sequencing is a significant limitation in metagenomic investigations and thus targeting specific genomic elements in whole-genome sequence data has been a considerable challenge (Gilbert and Dupont 2011). This is especially relevant for aquatic samples where dilution reduces the likelihood of identifying specific genes. That only 18 sequences in the Puget Sound dataset matched to known resistance genes potentially suggests a considerable under-sampling of the resistome when using pyrosequencing. Other marine and freshwater pyrosequenced metagenomes also have a relatively rare representation of resistance genes (Figure 3.10), suggesting the need for higher sequencing depths than achievable by 454 technology to fully characterize the resistomes of surface water microbial communities. For example, Illumina sequencing technology allows for greater sequencing depth than does pyrosequencing, but with shorter read lengths and therefore a greater reliance on sequence assembly, which still remains a challenge for mixed microbial communities. Secondly, the lack of sequence data for the unculturable prokaryotic majority in combination with evidence from functional metagenomic studies indicating that many environmental resistance genes appear to be distantly related and share low sequence similarity to cultured resistance genes may result in an underestimation of the actual number of resistance genes. This challenge is being addressed through further sequencing of bacterial reference genomes and functional metagenomic projects, however more research is needed. Lastly, sequence-based identification of resistance genes does not necessarily imply these genes are functionally expressed. The resistance potential can be estimated using sequence-based metagenomics, but functional metagenomics or transcriptional analysis is needed to detect functional resistance genes and more clearly define public health risks.

The fate of antibiotic resistance determinants during the wastewater treatment process is

relatively unknown at this time. As no influent from the WWTP was collected in this study, it is unclear what impact the wastewater treatment process had on the prevalence of antibiotic resistance genes in the effluent. Previous studies have shown mixed results regarding resistance gene levels in pre- versus post-treated sewage. While concentrations of resistance genes in wastewater may significantly decrease through the wastewater treatment process (Munir et al. 2011), other studies have found that resistance gene levels are higher in the effluent and that therefore the treatment process may be selecting for resistant bacteria, genes or mobile genetic elements (Ferreira da Silva et al. 2006; Zhang et al. 2009; Uyaguari et al. 2011). Either way, the effluent had a significantly higher abundance of antibiotic resistance genes than Puget Sound, although further effluent sampling is needed to support this individual finding.

3.4.3. Differential abundance of mobile genetic elements across metagenomes

Mobile genetic elements including plasmids and transposable elements are important vectors for the transfer of antibiotic resistance genes (Boerlin and Reid-Smith 2008; Partridge et al. 2009). Although in-depth profiling of the plasmid fraction of metagenomes is currently not possible for whole genome sequencing projects due to low depth of sequence coverage, insights into the plasmid composition of the Puget Sound and WWTP metagenomic datasets could still be made from the ~0.04% of sequences that matched known plasmids. A total of 124 sequence reads (0.01%) and 3 contigs from the open Sound and Marina metagenomes and 265 reads (0.22%) and 4 contigs from the WWTP matched plasmids within the NCBI RefSeq database using a similarity criteria of ≥95% sequence identity over 100 bp (Figure 3.11A and Table 3.4). There were 13 plasmids in common between the open Sound, Marina and WWTP metagenomes. The open Sound and Marina plasmid sequences were dominated by α- and γ-Proteobacteria, while the WWTP plasmids were associated with α-, β- and γ-Proteobacteria, Firmicutes and

Actinobacteria. At the species level, nearly half of the plasmid matches to the open Sound and Marina reads were sourced from *Silicibacter sp. Silicibacter* is a member of the marine Roseobacter clade (α-Proteobacteria) and has been shown to form symbioses with phytoplankton (Moran et al. 2004). The remainder of the open Sound and Marina plasmid sequences had taxonomic assignments to other environmental bacteria in addition to a number of human pathogens including *Klebsiella pneumonia*, *Salmonella enterica*, *Vibrio sp.*, *Pseudomonas aeruginosa* and *Serratia marcescens*. Many of these plasmids contain genes associated with virulence factors, heavy metal resistance, beta-lactamase, tetracycline and aminoglycoside resistance and antibiotic resistance determinants including transposases and integrons. For example, plasmid TC68 from *Vibrio sp.* is a known chloramphenicol resistance gene that has been previously isolated from a fish farm microbial community (Furushita et al. 2011). The WWTP plasmid sequences were characterized by both pathogenic and non-pathogenic host-specific bacteria. Potential human pathogens included *Bacteroides fragilis, Campylobacter*, *Enterococcus*, *Klebsiella pneumonia*, *Listeria monocytogenes* and *Salmonella*. Plasmids associated with pathogenic bacteria that are also known to carry antibiotic resistance genes included plasmid pKA1 from *Vibrio cholerae* (TEM beta-lactamase resistance gene) and plasmid R27 from *Salmonella typhi* (tetracycline tetACD resistance genes). The presence of pathogens with plasmids carrying resistance genes is thus suggestive of a potential exposure risk and transfer potential. Also present were bacterial indicators of fecal contamination, including *Bifidobacterium* and *Escherichia coli*. Bacteria associated with food products and used in food production were also prevalent, including *Lactococcus lactis*, *Lactobacillus rhamnosus* and *Streptococcus thermophiles*. The two largest contigs from the WWTP that matched plasmid

sequences had 100% identity to plasmids from uncultured bacteria previously identified in activated sludge from other WWTPs (Schluter et al. 2007; Suenaga et al. 2009) (Table 3.3).

Both approaches to identify transposable elements resulted in an increasing abundance of transposable elements from open Sound to Marina to WWTP, with a highly elevated proportion in the WWTP (Figure 3.11B and 3.10C). The Pfam approach did detect a higher proportion of transposable elements in the Marina and WWTP than the BLAST approach, as well as a significant difference in transposable element abundance between the Marina and open Sound. No significant differences in the abundance of transposable elements were seen within the open Sound samples using either approach. The most abundant transposases in the effluent included Tn3, IS4, TnpA (gene product of IS608), the mutator transposases and IS3 (specifically IS911), in addition to a large percentage of unclassified transposases (Figure 3.11D). While the most abundant transposable elements in the Puget Sound samples were the same as that for the effluent, they were for the most part present in much lower numbers. Transposable elements from the effluent were more likely to be found on plasmids than the Puget Sound samples. Approximately 20.7% of transposable elements from the effluent were annotated as being sourced from plasmids, compared to 0.039%±0.008 for the open Sound and 13.56% for the Marina. Plasmids that were unique to and represented at least 3% of the total plasmids annotated as carrying transposable elements in the WWTP sample included pA81 (biodegradation/heavy metal resistance) (Jencova et al. 2008), pMOL28 (heavy metal resistance) (Monchy et al. 2007) and pSKYE1 (aromatic compound degradation) (Suenaga et al. 2009). Plasmid pAA1 (xenobiotic compound metabolism) (Sajjaphan et al. 2004) was also abundant in both the WWTP and Marina samples. Plasmids pDSHI01 from *Dinoroseobacter* and pMAQU02 from *Marinobacter* were commonly found within the Puget Sound samples.

3.5 Conclusions

This comparative metagenomic survey of Puget Sound has provided baseline data describing the community composition and antibiotic resistance determinant abundance across this marine environment in addition to an effluent sample from a proximal WWTP. Our results support the use of whole-genome pyrosequencing for comparing community composition across differentially impacted environments and for profiling antibiotic resistance determinants in highly impacted environments. To more completely capture the resistomes of natural environments with low human impact, sequencing technologies allowing for greater depth of sequencing may be needed. This study has highlighted the similarity of these metagenomic components in the open estuarine environment and the differences that emerge when analyzing more nearshore and human impacted ecosystems. Taken together, these results warrant further investigation into the potential for WWTP effluent to disseminate resistance determinants into the marine environment. As only one Marina and one WWTP sample were included in this study, further sampling of these environments is needed in order to support the observed trends. Furthermore, this study only analyzed surface water communities, and community composition has been show to change significantly with depth (DeLong et al. 2006). Increasing the spatial and temporal extents of metagenomic sampling and analysis will thus be important for longitudinal monitoring and further assessment of human impacts in marine environments.

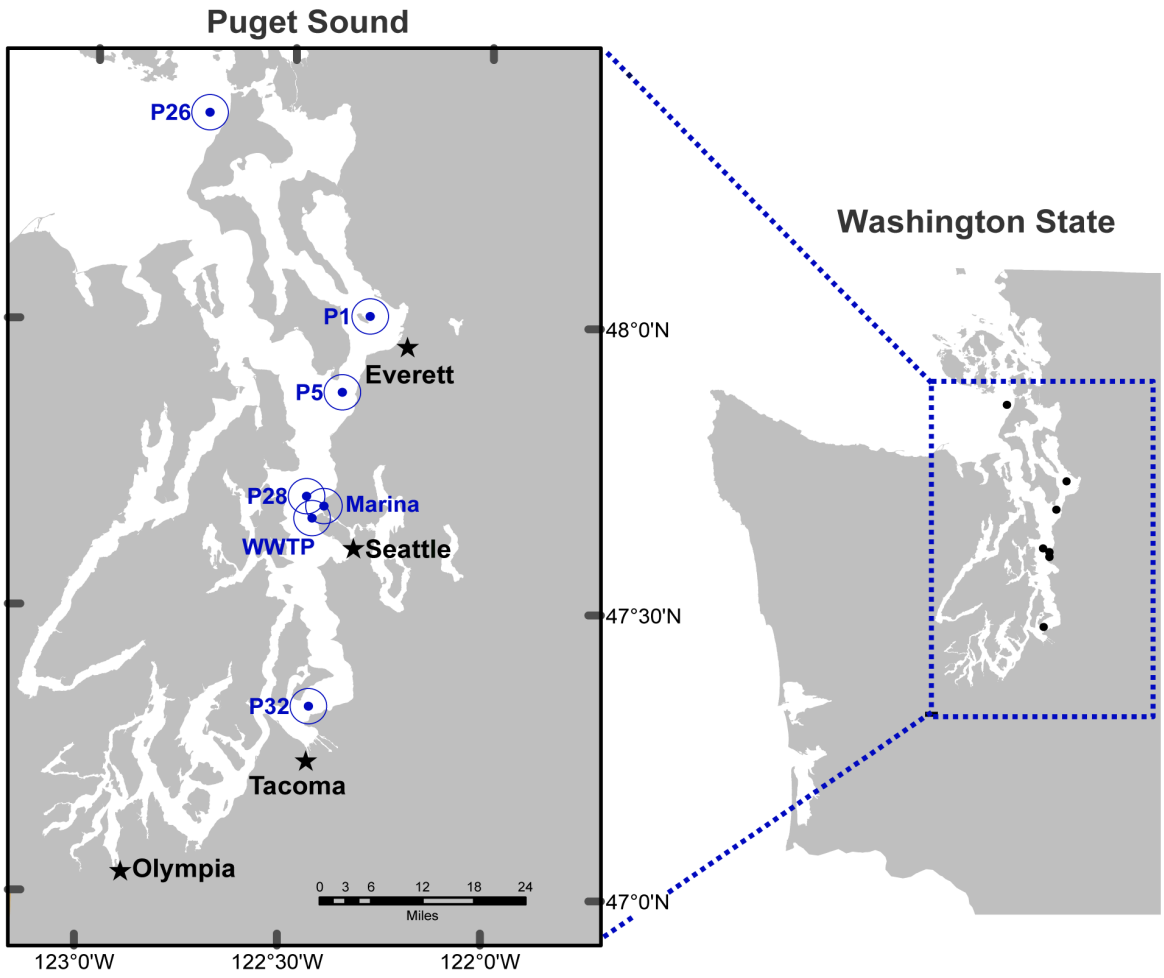

Figure 3.1. Locations of sampling sites in Puget Sound. Refer to Table 3.1 for the geographic coordinates of the sampling stations. WWTP, Wastewater treatment plant.

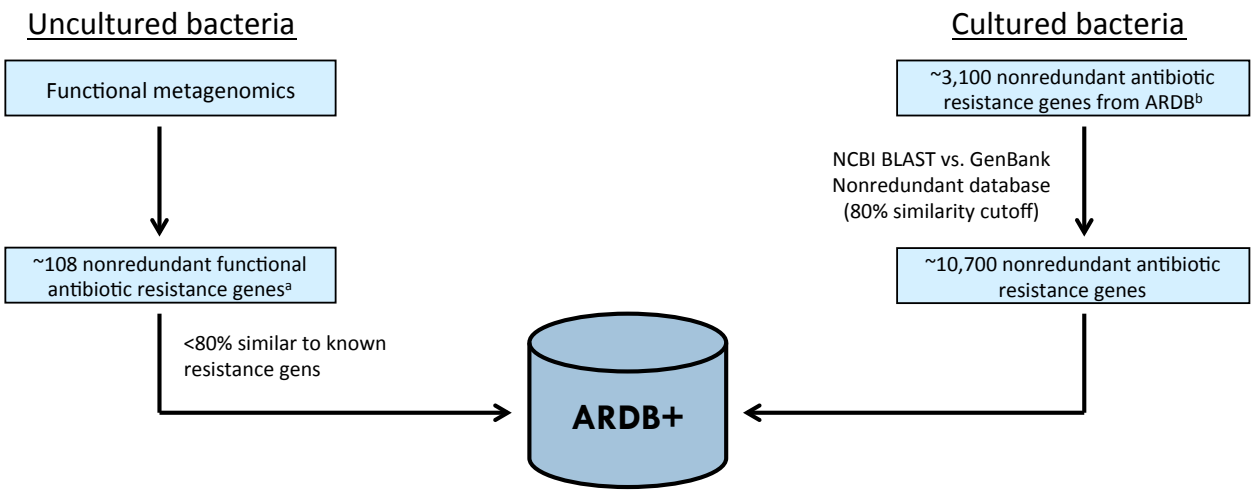

Figure 3.2. Contruction of the expanded Antibiotic Resistance Genes Database (ARDB+). Both antibiotic resistance genes sequences similar to known resistance genes within the ARDB and resistance genes identified through functional metagenomic projects are included in the database. [a]Sommer et al. 2009 (feces, saliva), Torres-Cortes et al. 2011 (soil), Donato et al. 2010 (soil), Lang et al. 2010 (soil), Allen et al. 2009 (soil), Riesenfeld et al. 2004 (soil), Mori et al., 2008 (sludge), Allen et al. 2009 (moth gut);[b]Liu and Pop 2009.

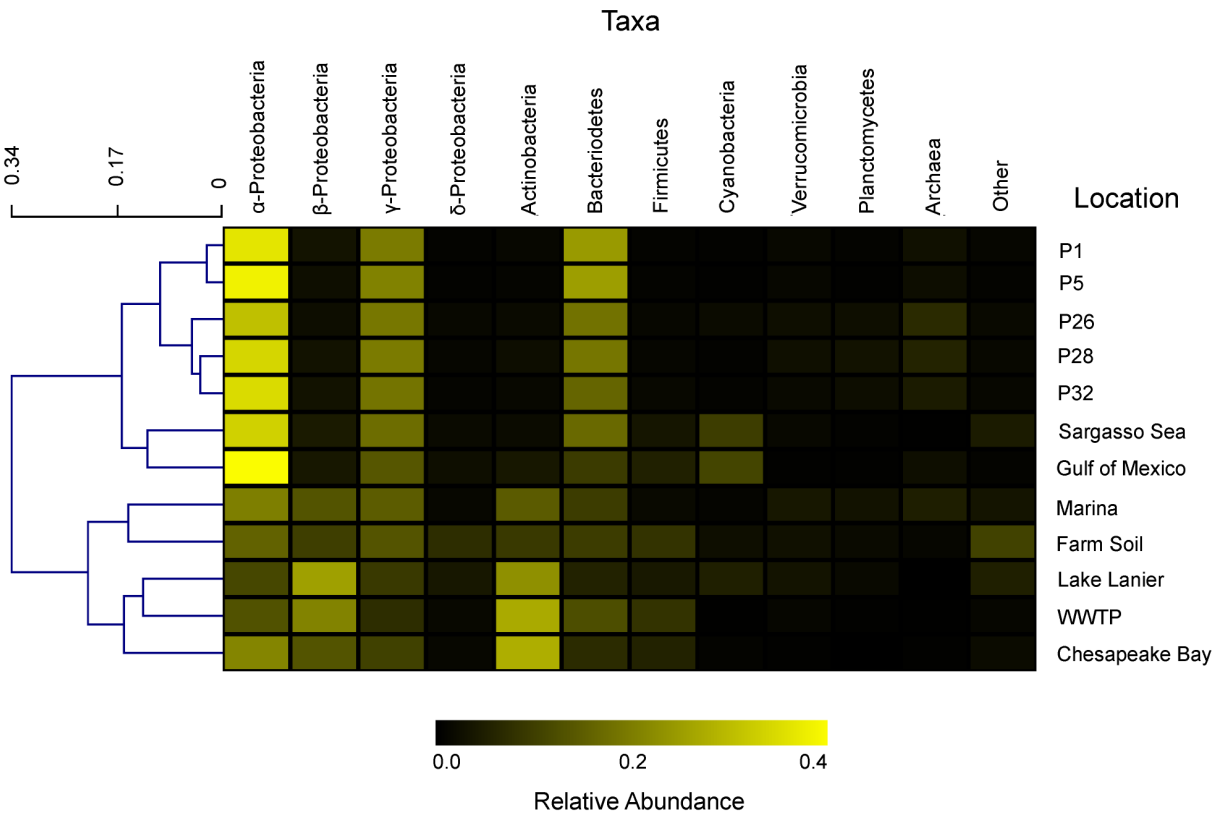

Figure 3.3. Relative abundance of major taxonomic groups in the Puget Sound samples and other selected metagenomes. Taxonomic groupings were based on BLASTX comparison to the NCBI taxonomy using MG-Rast [39] and ≥50% identity and an alignment length ≥50 amino acids. The 'other' category includes bacteria taxa present in <1% of sequences in all samples, eukaryotes and viruses. Shading is proportional to the relative abundance of each taxon within a metagenome. The cladogram was displayed using hierarchical clustering and the Euclidian distance metric. See Materials and Methods for references describing the additional metagenomes used for comparison. The open Sound locations cluster with other more saline environments including open and coastal ocean samples while the Marina and WWTP effluent samples cluster within a larger clade containing other freshwater and freshwater-impacted metagenomes.

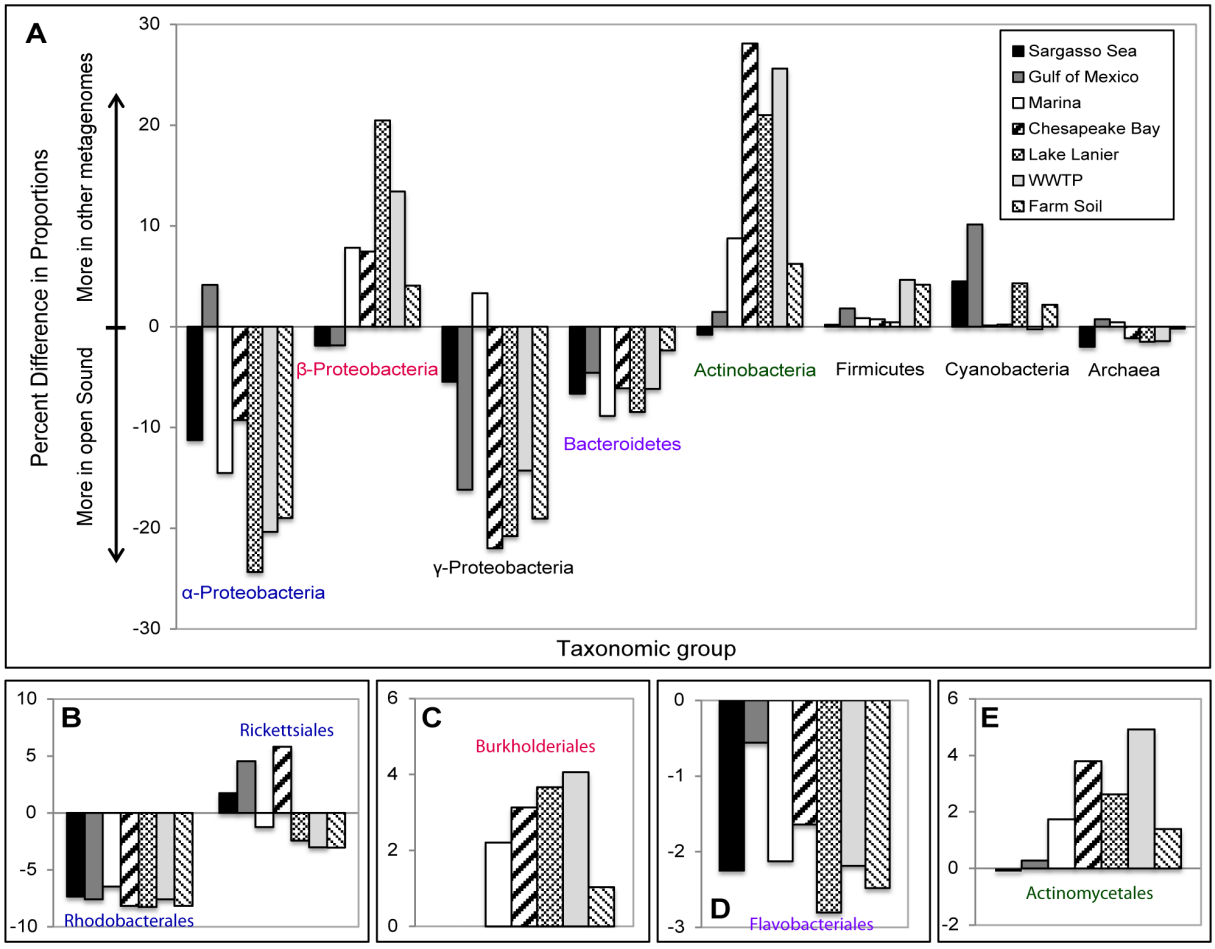

Figure 3.4. Over- and under-representation of predominant taxa in selected biomes relative to the open Puget Sound metagenome. (A) Overall taxa, (B) α-Proteobacteria, (C) β-Proteobacteria, (D) Bacteroidetes and (E) Actinobacteria. The difference in proportions refers to the percent difference in the relative abundances of a given taxa between two locations. Only taxa with a difference of proportions >1% are shown. Negative values indicate higher relative abundance in the Puget Sound ($p < 10^{-15}$ for all comparisons). See Materials and Methods for references describing the additional metagenomes used for comparison. A unique taxonomic signature was identified for the Puget Sound consisting of an over-representation of α-Proteobacteria, γ-Proteobacteria and Bacteriodetes and an under-representation of β-Proteobacteria and Actinobacteria.

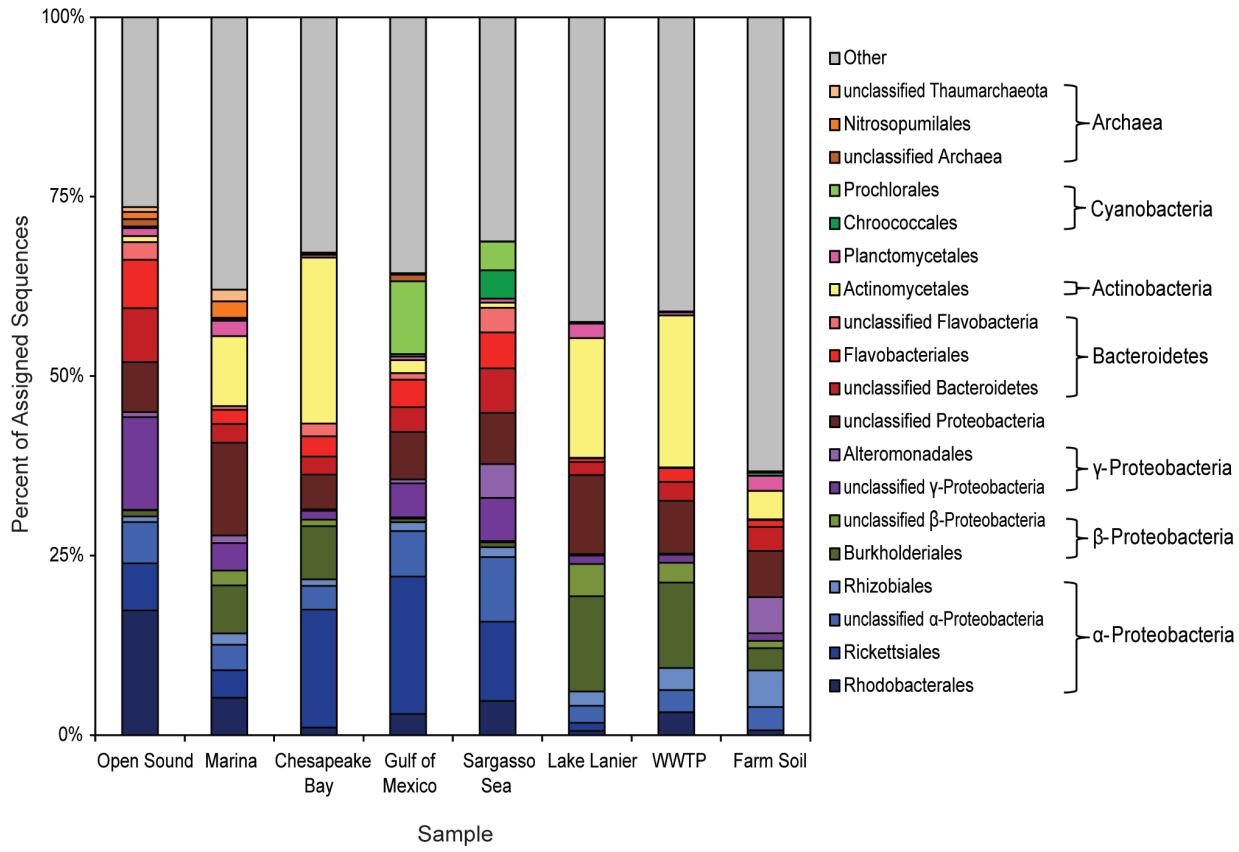

Figure 3.5. Relative abundance (order level) of major taxonomic groups in the Puget Sound samples and other selected metagenomes. Sequences were assigned to the NCBI taxonomy using MG-Rast (Meyer et al. 2008) and the lowest common ancestor algorithm (≥50% identity and alignment length ≥50 amino acids). Taxa representing >1% of assignable sequences in one or more samples are shown, while taxa present in <1% of sequences in all samples are grouped in the 'other' category.

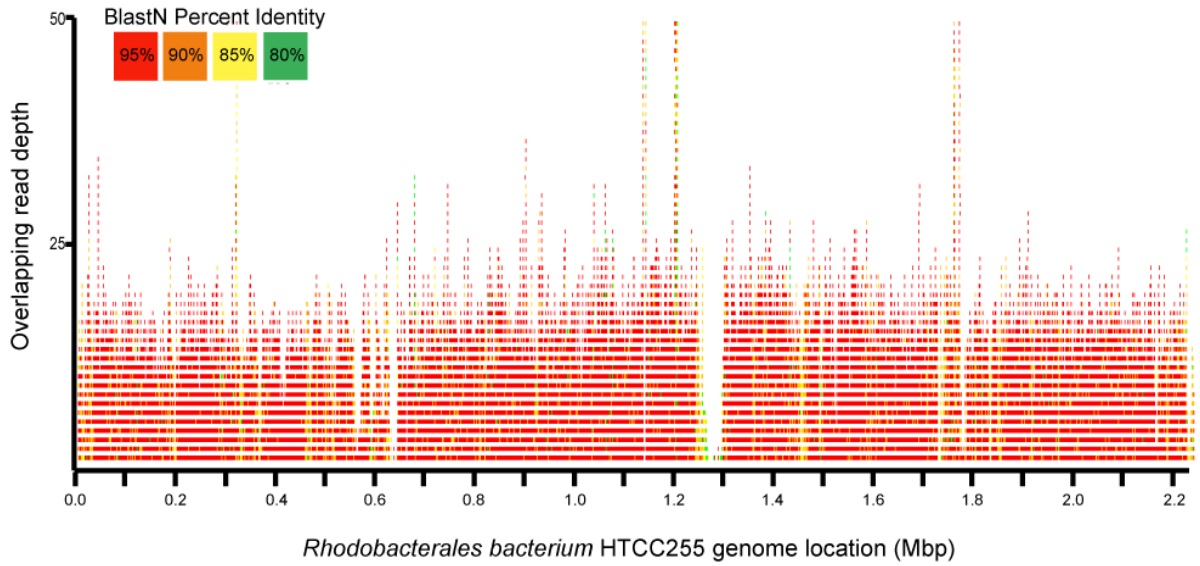

Figure 3.6. Recruitment plot of the unassembled Puget Sound reads to the alpha-Proteobacterium HTCC2255 genome assembly. Open Sound and Marina metagenomic sequence reads were recruited to the 2.23 MB assembly (GI number: 211594581) using BLASTN and an E-value cutoff of 10^{-5}. Over 91% of the assembly was matched at greater than 95% identity, equaling approximately 7X coverage.

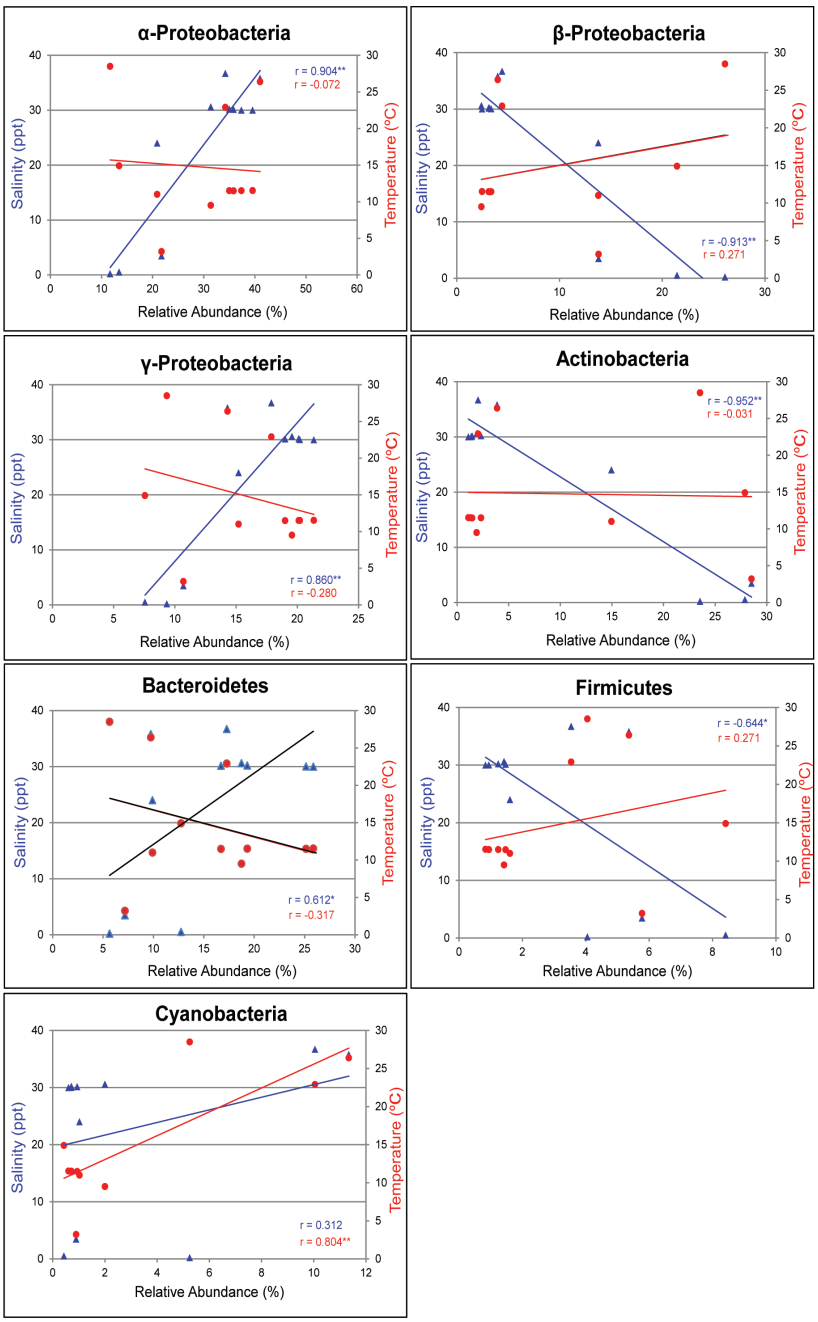

Figure 3.7. Relationship between the relative abundance of predominant taxa and salinity and temperature gradients. Linear regression lines and Pearson's coefficients (*p<0.05 and **p<0.005) are shown. Data points include the open Sound, Marina, WWTP, Chesapeake Bay, Sargasso Sea, Gulf of Mexico and Lake Lanier metagenomic samples.

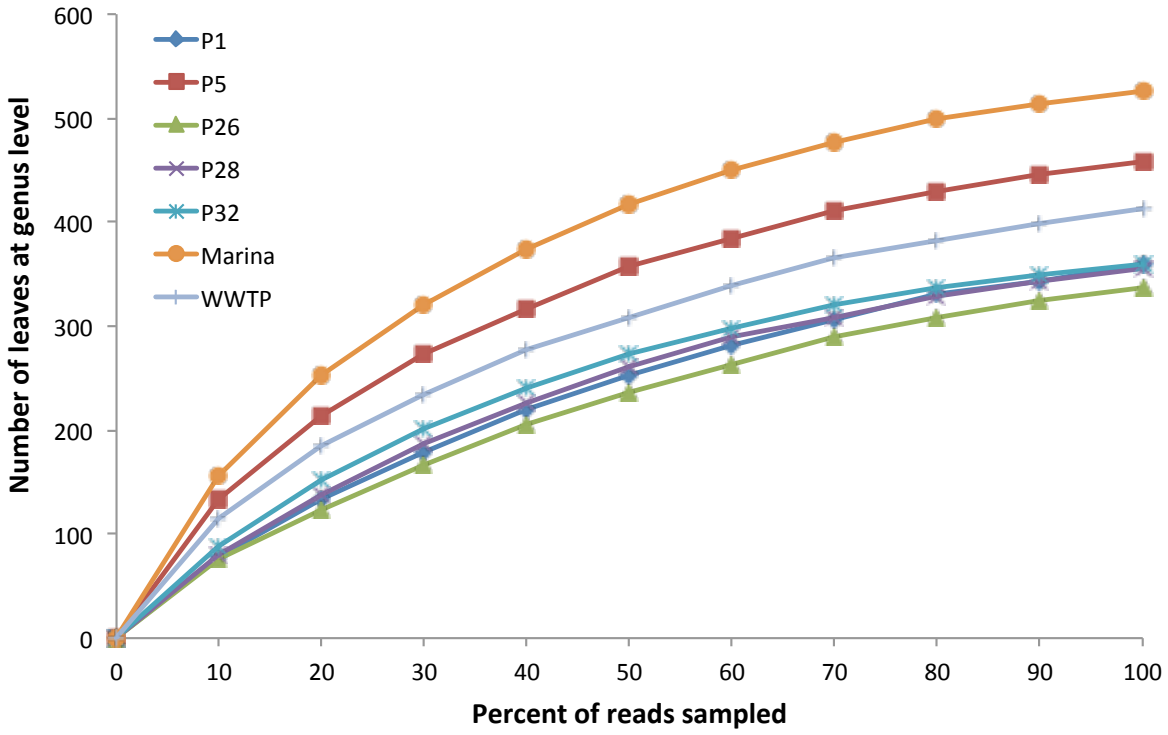

Figure 3.8. Rarefaction curves for each metagenome performed at the genus level. MEGAN (Huson et al. 2007) was used for the rarefaction analysis. This program uses the LCA algorithm to bin reads to taxa based on their BLAST hits. A rooted tree is generated where each node represents a taxon, and the leaves of the tree are then used as operational taxonomic units (OTUs) in the rarefaction analysis.

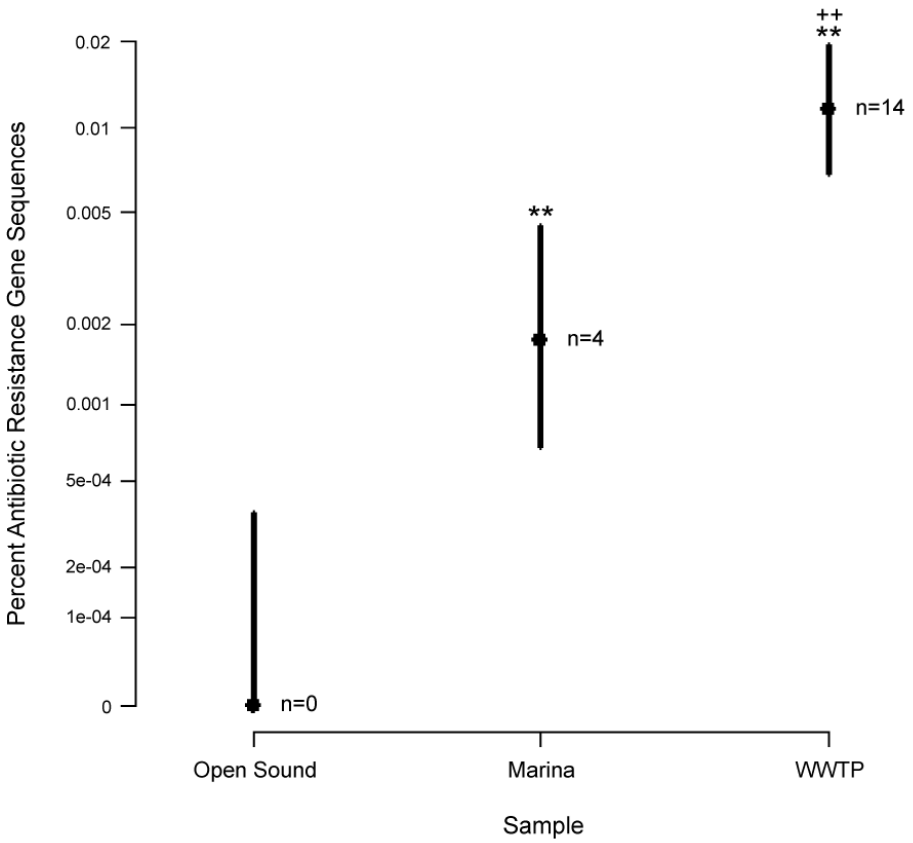

Figure 3.9. Prevalence of antibiotic resistance gene sequences in the Puget Sound metagenomes. Sequences were classified as antibiotic resistance genes if sharing ≥80% identity and an alignment length ≥ 50aa to a sequence within an expanded Antibiotic Resistance Genes Database (ARDB) (Liu and Pop 2009). Open Sound samples were pooled together for the analysis. Bars represent 95% confidence intervals for binomial proportions. For open Sound vs. Marina and WWTP, *$p<0.05$ and **$p<0.005$. For Marina vs. WWTP, $^{+}p<0.05$ and $^{++}p<0.005$. 'n=' refers to the number of antibiotic resistance gene sequences. The y-axis represents a modified log scale. While the resistance gene signals were low for the samples, the signals were significantly different across the sample types, suggesting antibiotic resistance gene abundance may reflect differences in potential human health impacts.

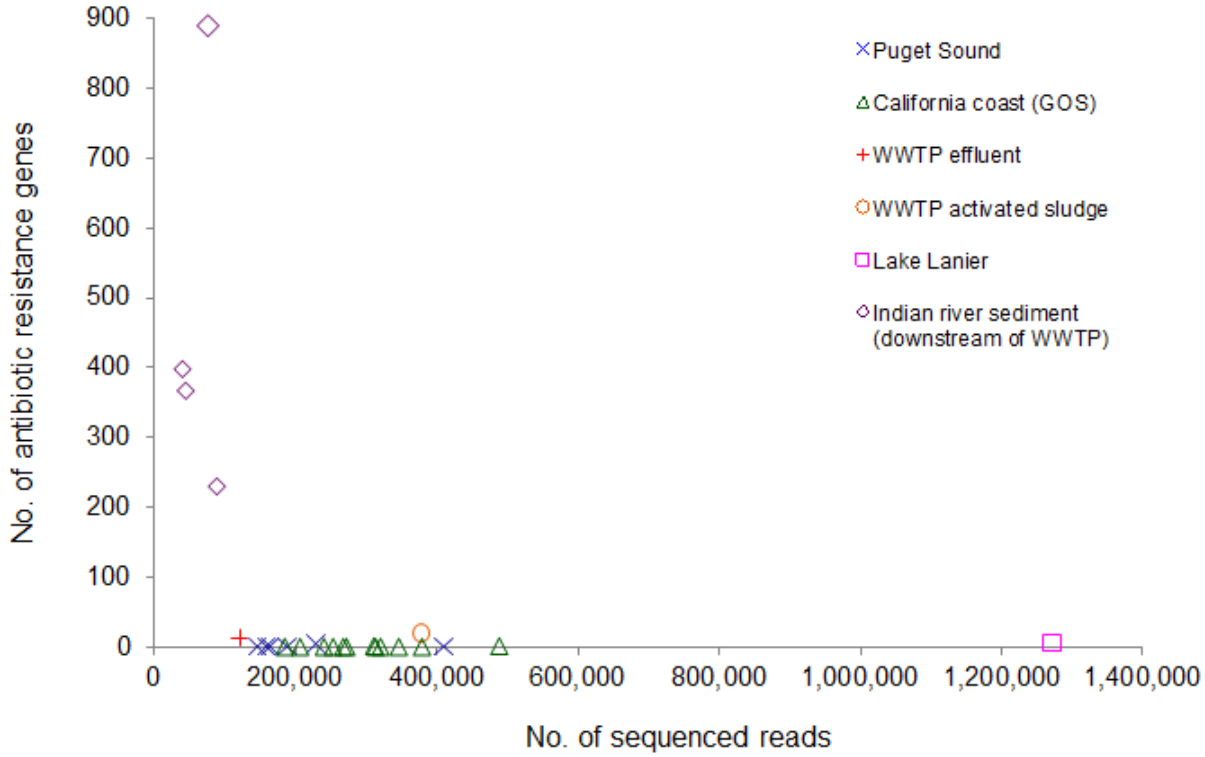

Figure 3.10. Abundance of antibiotic resistance gene sequences in environmental metagenomes with varying sequencing depth. These metagenomes represent a diverse mix of environments, including river sediment samples taken downstream from a wastewater treatment plant (WWTP) processing high volumes of antibiotics (Kristiansson 2011), coastal surface water samples taken as part of the Global Ocean Sampling Expedition (Zeigler Allen et al. 2012), the activated sludge fraction of a WWTP (Sanapareddy et al. 2009) and an urban freshwater lake (Oh et al. 2011). Reads that aligned to a sequence within the ARDB+ with ≥80% sequence identity over at least 50 amino acids were classified as putative resistance genes.

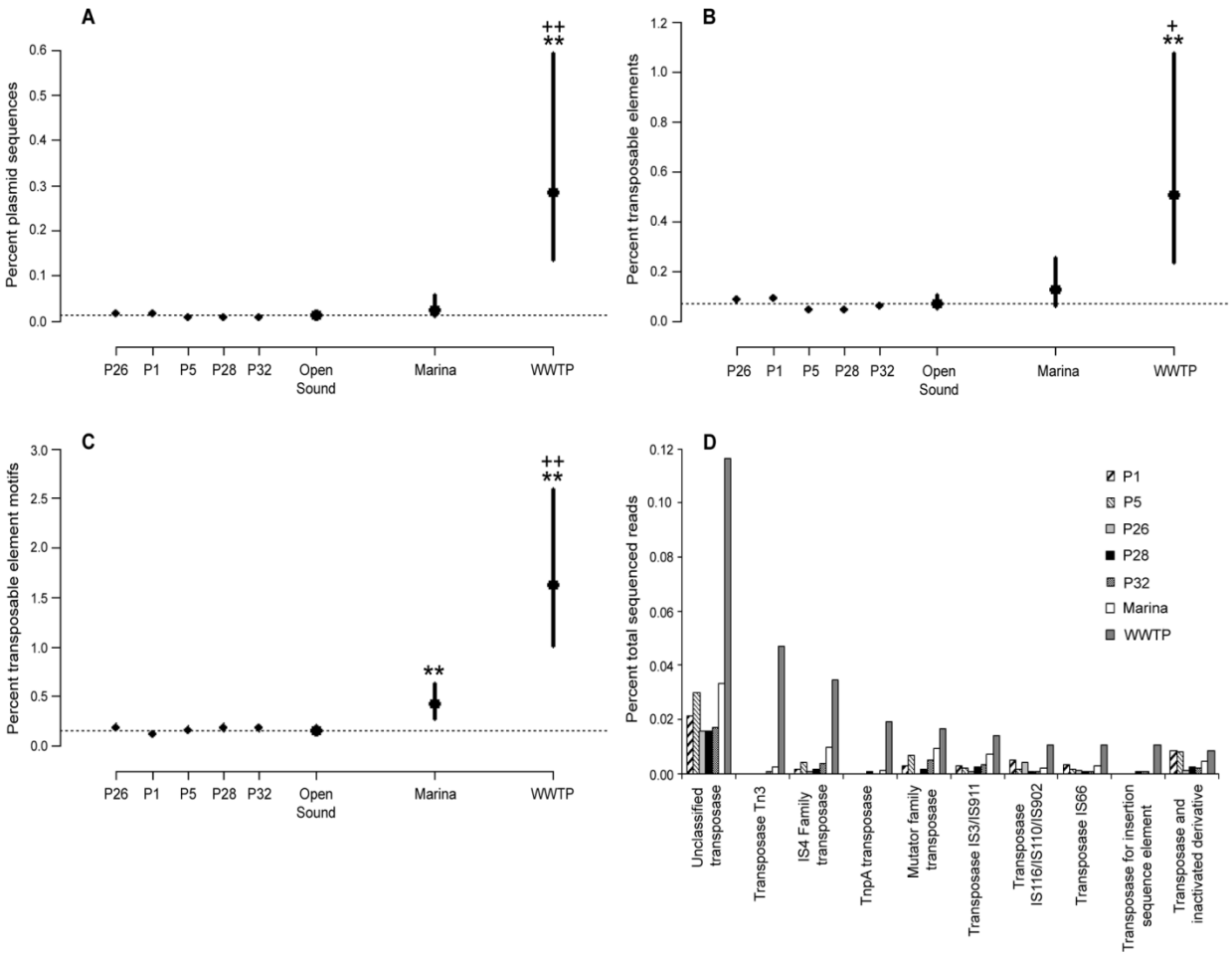

Figure 3.11. Prevalence of mobile genetic elements in the Puget Sound metagenomes. (A) Plasmid sequences, (B) transposable elements using a BLAST search against GenBank, (C) transposable elements using a hidden markov model (HMM) search against the Pfam database and (D) top ten most abundant transposable elements from the WWTP effluent by percent abundance. Bars represent 95% confidence intervals for binomial proportions. P-values are relative to the open Sound and are corrected for multiple testing. For open Sound vs. Marina and WWTP, *$p<0.05$ and **$p<0.005$. For Marina vs. WWTP, $^{+}p<0.05$ and $^{++}p<0.005$. Bars represent 95% confidence intervals for binomial proportions. The dashed line represents the mean percent for the open Sound. Similar profiles for mobile genetic elements were seen for the open Sound samples while significant differences emerged when comparing to the Marina and effluent.

Table 3.1. Sampling locations and summary statistics for the Puget Sound metagenomes.

Sample	Location	Date	Depth (m)	Temp (°C)	Salinity (ppt)	Chlorophyll (mg/m^{-3})	No. of reads	Mean read length (bp)	% GC	No. of contigs (> 1kb)
P1	48°01'0"N; 122°18'2"W	10/29/10	5	11.5	30.02	0.2	187,776	367	43.8	81
P5	47°53'0"N; 122°22'0"W	10/29/10	5	11.53	30	0.51	409,672	372	43.2	1,277
P26	48°22'5"N; 122°43'0"W	10/29/10	5	9.5	30.6	0.3	146,173	367	43.3	6
P28	47°42'2"N; 122°27'2"W	10/29/10	5	11.51	32.2	0.15	164,069	369	44.2	22
P32	47°20'0"N; 122°26'5"W	10/30/10	5	11.48	30.15	0.58	159,314	367	43.6	12
Marina	47°41'2"N; 122°24'25"W	12/20/10	0.5	~11.00[b]	~24.00[b]	No data	228,712	379	46.8	86
WWTP	47°39'42"N; 122°26'0"W	1/31/11	NA[a]	14.9	<0.5	No data	120,857	381	56.4	77

[a]NA, not applicable
[b]Based on observational data

Table 3.2. Puget Sound 454 sequence reads with ≥80% similarity and an alignment length ≥50 amino acids to sequences within the expanded Antibiotic Resistance Genes Database (ARDB+).

454 Read ID	Sample	Best hit to ARDB+	Percent identity	Alignment length (aa)	Query coverage	Antibiotic resistance gene	Antibiotic resistance class	Species	Best hit Genbank ID	Genbank protein
GYS1QVW02GFSBG	WWTP	122934824	96	141	57	BL3_imp	Beta-lactam	*Serratia marcescens*	122934824	Metallo-beta-lactamase
GYS1QVW02JR7KA	WWTP	58200478	100	120	56	tetA(39)	Tetracycline	*Acinetobacter sp. LUH5605*	58200478	TetR(39)
GYS1QVW01EZL9G	Marina	241992557	100	53	55	qacG2	Quaternary ammonium compounds	*Uncultured bacterium*	241992557	QacG
GYS1QVW01C3WDM	WWTP	314993276	95	137	55	ermB	Macrolide	*Enterococcus faecium TX0133B*	256964193	rRNA methylase
GYS1QVW01EPBXP	WWTP	314993276	100	132	53	ermB	Macrolide	*Enterococcus faecium TX0133B*	146285381	Ribosomal methylase
GYS1QVW01A42OG	WWTP	289812760	89	119	51	BL2d_ oxa10	Beta-lactam	*Providencia rettgeri*	56791710	OXA-10 beta lactamase
GYS1QVW02IDABT	WWTP	224496169	99	144	42	MefA	Macrolide	*Streptococcus anginosus*	169809207	Macrolide-efflux protein
GYS1QVW02G6714	WWTP	58200479	94	155	39	tet39	Tetracycline	*Acinetobacter sp. LUH5605*	58200479	TetA(39)
GYS1QVW02GVJSG	Marina	121604263	82	108	39	BacA	Bacitracin	*Polaromonas naphthalenivorans CJ2*	332528682	UDP pyrophospate phosphatase
GYS1QVW02HMRHR	Marina	84616919	80	151	38	orf11	Beta-lactam	*Desulfococcus multivorans*	84616919	Putative ABC-type permease
GYS1QVW02JRJX1	WWTP	283798930	99	132	33	tet40	Tetracycline	*Clostridium sp. M62/1*	283798930	MDR-type permease, tet resistance protein
GYS1QVW02F32RH	WWTP	296285266	99	161	25	tetW	Tetracycline	*Corynebacterium resistens*	323141252	Elongation factor G
GYS1QVW01AG7BR	WWTP	56479010	85	74	19	orf11	Beta-lactam	*Aromatoleum aromaticum EbN1*	56479010	ABC transporter permease
GYS1QVW01DHHKX	Marina	146306240	83	141	14	MexW	Multidrug resistant efflux protein	*Pseudomonas mendocina ymp*	77460714	Multidrug efflux protein
GYS1QVW02HM984	WWTP	254967143	97	66	10	tetW	Tetracycline	*Uncultured organism*	204789616	Tetracycline resistance protein
GYS1QVW02FI0S3	WWTP	254967143	100	50	8	tetW	Tetracycline	*Uncultured organism*	204789616	Tetracycline resistance protein
GYS1QVW02JXAIU	WWTP	260597281	82	50	8	MacB	Macrolide	*Cronobacter turicensis z3032*	346224427	Hypothetical protein
GYS1QVW02JY5OX	WWTP	260596113	87	62	6	AcrB	Macrolide	*Cronobacter turicensis z3032*	320540972	putative multidrug efflux system protein

Table 3.3. Plasmid sequences in the NCBI RefSeq database that match contigs from the Puget Sound and wastewater treatment plant (WWTP) effluent metagenomic datasets.

Contig ID	**Location**	**Plasmid name**	**GI number**	**Bacterial host**	**Identity (%)**	**Hit length (bp)**
00012	WWTP	pGNB1	149350899	*Uncultured bacterium*	100	2,130
00060	WWTP	pSKYE1	255292227	*Uncultured bacterium*	100	1,127
00059	WWTP	pAOVO01	120608524	*Acidovorax sp.*	100	1,093
01935	P5	TM1040	99034845	*Silicibacter sp.*	97	537
00155	WWTP	pK214	2467210	*Lactococcus lactis*	97	749
00139	P32	TM1040	99034845	*Silicibacter sp.*	96	562
00332	P1	TM1040	99034845	*Silicibacter sp.*	96	556

Table 3.4. Plasmid sequences in the NCBI RefSeq database that match sequence reads from the Puget Sound and wastewater treatment plant (WWTP) effluent datasets.

GI number	Plasmid name	Bacterial host	Identity (%) ≥	Hit length (bp) ≥	No. of reads	Location
99034845	unnamed	*Silicibacter sp.*	95	109	60	P1, P5, P26, P28, P32, Marina
56410263	pMOL28	*Ralstonia metallidurans*	97	102	38	P28, WWTP
310639221	pYP1	*Ketogulonicigenium vulgare*	95	108	29	P1, P5, P26, P28, P32, Marina, WWTP
32263857	pOL18	*Paracoccus sp.*	95	106	19	P1, P5, P26, P32, Marina
299068436	CMR15_mp	*Ralstonia solanacearum*	95	116	13	P5, Marina, WWTP
255292227	pSKYE1	*Uncultured bacterium*	96	261	13	WWTP
51470556	pFBAOT6	*Aeromonas punctata*	97	238	11	P32, WWTP
221236817	unnamed	*Escherichia sp.*	95	105	9	P1, P5, P26, P28, P32, Marina, WWTP
145557411	pRSPA01	*Rhodobacter sphaeroides*	95	128	9	P1, P5, Marina, WWTP
16605595	pRVS1	*Vibrio salmonicida*	98	241	9	Marina
33867057	pBD2	*Rhodococcus erythropolis*	98	300	8	WWTP
326407899	pCV56A	*Lactococcus lactis*	98	259	7	WWTP
146322225	pGdh442	*Lactococcus lactis*	96	266	7	WWTP
66862639	Rms149	*Pseudomonas aeruginosa*	99	178	7	Marina, WWTP
2467210	pK214	*Lactococcus lactis*	96	128	6	WWTP
18077099	pWW0	*Pseudomonas putida*	100	119	6	WWTP
113473678	pCAR3	*Sphingomonas sp.*	96	251	6	WWTP
149350899	pGNB1	*Uncultured bacterium*	97	225	6	WWTP
120608524	pAOVO01	*Acidovorax sp.*	98	264	5	WWTP
76151975	pA17sv1	*Enterococcus faecium*	97	126	5	WWTP
238873439	unnamed	*Eubacterium eligens*	98	217	5	WWTP
326407968	pCV56C	*Lactococcus lactis*	98	226	5	WWTP
24394861	pKLH201	*Acinetobacter calcoaceticus*	99	239	4	WWTP
288959735	pAB510a	*Azospirillum sp.*	95	170	4	P1, P5, P32, Marina
306415518	pMC1	*Delftia acidovorans*	96	193	4	WWTP

239976803	unnamed	*Enterococcus faecium*	99	320	4	WWTP
179348463	pMPOP01	*Methylobacterium populi*	99	230	4	WWTP
321170327	pRAHAQ01	*Rahnella sp.*	99	251	4	P5, Marina
194709275	pCVM19633_110	*Salmonella enterica*	100	158	4	Marina
292677436	pCHQ1	*Sphingobium japonicum*	95	322	4	WWTP
31746361	pB10	*Uncultured bacterium*	99	187	4	WWTP
58416215	pA81	*Achromobacter xylosoxidans*	99	148	3	WWTP
283484478	pMMD	*Acinetobacter baumannii*	97	315	3	P32, WWTP
83833713	pAV2	*Acinetobacter venetianus*	98	329	3	P28, WWTP
307111956	pBUN24	*Bacteroides uniformis*	98	191	3	WWTP
30387438	PNAC2	*Bifidobacterium longum*	99	362	3	WWTP
57019077	pCC178	*Campylobacter coli*	97	226	3	WWTP
17059596	pTET3	*Corynebacterium glutamicum*	98	179	3	WWTP
34500478	pUO1	*Delftia acidovorans*	98	165	3	WWTP
121490884	pVEF1	*Enterococcus faecium*	98	159	3	WWTP
38016624	pLVPK	*Klebsiella pneumoniae*	95	340	3	Marina
150958389	pKPN4	*Klebsiella pneumoniae*	95	185	3	WWTP
179366399	pLR581	*Lactobacillus reuteri*	98	325	3	WWTP
183178785	pMM23	*Mycobacterium marinum*	99	453	3	WWTP
17548221	pGMI1000MP	*Ralstonia solanacearum*	95	227	3	Marina, WWTP
160431608	pMAK2	*Salmonella enterica*	95	273	3	WWTP
295797733	pTINT01	*Thiomonas intermedia*	95	257	3	P1, WWTP
205320843	pHHV35	*Uncultured bacterium*	96	167	3	WWTP
311109684	pA81	*Achromobacter xylosoxidans*	99	314	2	WWTP
58200477	ptet5605	*Acinetobacter sp.*	100	432	2	WWTP
30172175	pKLH205	*Acinetobacter sp.*	99	178	2	WWTP
10957030	pRAY	*Acinetobacter sp.*	95	115	2	WWTP

294351974	pBM400	*Bacillus megaterium*	95	199	2	WWTP
194359988	pBFP35	*Bacteroides fragilis*	99	419	2	WWTP
146411246	pBBta01	*Bradyrhizobium sp.*	96	159	2	Marina
134132180	pBVIE02	*Burkholderia vietnamiensis*	100	250	2	WWTP
57019266	pCC178	*Campylobacter coli*	99	385	2	WWTP
190571920	pCNB	*Comamonas testosteroni*	99	102	2	WWTP
295059951	pECL_A	*Enterobacter cloacae*	99	403	2	Marina, WWTP
12956985	pRE25	*Enterococcus faecalis*	99	261	2	WWTP
309385929	pLG1	*Enterococcus faecium*	97	437	2	WWTP
241248270	unnamed	*Enterococcus faecium*	95	394	2	WWTP
239977009	unnamed	*Enterococcus faecium*	98	242	2	WWTP
193783420	pJIBE401	*Klebsiella pneumoniae*	98	161	2	Marina
171854418	pKL0018	*Lactococcus garvieae*	99	119	2	WWTP
326407939	pCV56B	*Lactococcus lactis*	96	342	2	WWTP
190571754	pNP40	*Lactococcus lactis*	99	284	2	WWTP
116108915	Plasmid 3	*Lactococcus lactis*	97	245	2	WWTP
44078	lactose	*Lactococcus lactis*	96	179	2	WWTP
32455473	pAH82	*Lactococcus lactis*	98	478	2	WWTP
47018986	pLM80	*Listeria monocytogenes*	98	235	2	WWTP
32469878	pDTG1	*Pseudomonas putida*	95	384	2	WWTP
190572013	pCT14	*Pseudomonas sp.*	98	180	2	Marina, WWTP
56068618	pMOL30	*Ralstonia metallidurans*	96	309	2	WWTP
77454567	pREL1	*Rhodococcus erythropolis*	99	446	2	WWTP
148550551	pSWIT01	*Sphingomonas wittichii*	99	376	2	WWTP
317109770	PB5	*Uncultured bacterium*	98	106	2	WWTP
57236769	pKA1	*Vibrio cholerae*	96	482	2	Marina
183211582	pACICU1	*Acinetobacter baumannii*	100	145	1	Marina
30409103	pKLH207	*Actinobacter sp.*	99	286	1	WWTP

142855988	Plasmid 4	*Aeromonas salmonicida*	100	433	1	WWTP
317119630	pALIDE01	*Alicycliphilus denitrificans*	99	368	1	WWTP
288960867	pAB510b	*Azospirillum sp.*	96	212	1	P1
288962595	pAB510e	*Azospirillum sp.*	95	129	1	Marina
222822840	unnamed	*Bacteroides fragilis*	100	418	1	WWTP
222822802	unnamed	*Bacteroides sp.*	100	431	1	WWTP
222822750	unnamed	*Bacteroides sp.*	96	307	1	WWTP
853781	pIP417	*Bacteroides sp.*	100	413	1	WWTP
73665544	pBIF10	*Bifidobacterium bifidum*	100	289	1	WWTP
121505102	pTet	*Campylobacter jejuni*	100	358	1	WWTP
57118012	pCG8245	*Campylobacter jejuni*	99	328	1	WWTP
257048712	pAph03	*Candidatus Accumulibacter*	96	451	1	WWTP
50727963	pTSA	*Comamonas testosteroni*	98	359	1	WWTP
296172990	pJA144188	*Corynebacterium resistens*	96	153	1	WWTP
89513168	pLEW279a	*Corynebacterium sp.*	98	169	1	WWTP
32479367	pTP10	*Corynebacterium striatum*	99	512	1	WWTP
260600006	pCTU3	*Cronobacter turicensis*	97	393	1	Marina
226319394	Plasmid 1	*Deinococcus deserti*	95	116	1	Marina
113706807	pDGEO01	*Deinococcus geothermalis*	95	154	1	P26
226807567	pEC-IMP	*Enterobacter cloacae*	97	426	1	WWTP
270208272	pBEE99	*Enterococcus faecalis*	99	221	1	WWTP
241252865	unnamed	*Enterococcus faecalis*	100	205	1	WWTP
239825540	unnamed	*Enterococcus faecalis*	99	186	1	WWTP
190350257	pVEF3	*Enterococcus faecium*	100	459	1	WWTP
172051323	pIP1206	*Escherichia coli*	99	424	1	WWTP
99867038	pAPEC-O1-R	*Escherichia coli*	100	366	1	WWTP
89513164	pLEW517	*Escherichia coli*	100	176	1	WWTP
89033265	NR1	*Escherichia coli*	100	275	1	WWTP

12024948	R751	*Escherichia coli*	99	361	1	WWTP
195537732	pLTK13	*Lactobacillus plantarum*	99	369	1	P5
32455506	pMD5057	*Lactobacillus plantarum*	99	200	1	Marina
257152781	N/A	*Lactobacillus rhamnosus*	99	206	1	WWTP
125631981	pEps352	*Lactococcus lactis*	98	150	1	WWTP
76574874	pSK11L	*Lactococcus lactis*	99	427	1	WWTP
32455464	pBL1	*Lactococcus lactis*	100	273	1	WWTP
9507248	pND861	*Lactococcus lactis*	98	524	1	WWTP
6739582	pCI2000	*Lactococcus lactis*	95	462	1	WWTP
53755675	pLPP	*Legionella pneumophila*	97	241	1	WWTP
295831620	LkipL4726	*Leuconostoc kimchii*	100	431	1	WWTP
222121345	pMCCL2	*Macrococcus caseolyticus*	99	393	1	WWTP
82619190	unnamed	*Mesorhizobium sp.*	95	212	1	P1
110346917	Plasmid 1	*Mesorhizobium sp.*	100	317	1	WWTP
170658659	pMRAD01	*Methylobacterium radiotolerans*	95	254	1	P1
216774	unnamed	*Moraxella sp.*	99	378	1	WWTP
315265130	pMSPYR101	*Mycobacterium sp.*	97	104	1	WWTP
325983714	pNAL21201	*Nitrosomonas sp.*	99	489	1	WWTP
325980815	pNAL21202	*Nitrosomonas sp.*	97	139	1	WWTP
145322134	pNL1	*Novosphingobium aromaticivorans*	97	262	1	WWTP
294869143	pAMI7	*Paracoccus aminophilus*	99	449	1	WWTP
154818276	pMTH1	*Paracoccus methylutens*	100	379	1	WWTP
118504932	pPRO2	*Pelobacter propionicus*	100	407	1	WWTP
121583017	pPNAP02	*Polaromonas naphthalenivorans*	95	423	1	WWTP
120595973	pPNAP01	*Polaromonas naphthalenivorans*	98	102	1	WWTP
194359396	pOZ176	*Pseudomonas aeruginosa*	99	385	1	P28
156104616	pMATVIM-7	*Pseudomonas aeruginosa*	100	362	1	WWTP
599573	pSA1700	*Pseudomonas aeruginosa*	95	487	1	WWTP

49188490	pRA2	*Pseudomonas alcaligenes*	95	253	1	WWTP
296100168	pDK1	*Pseudomonas putida*	100	404	1	Marina
90576544	NAH7	*Pseudomonas putida*	95	230	1	WWTP
42632299	pND6-1	*Pseudomonas sp.*	100	437	1	WWTP
32455785	pADP-1	*Pseudomonas sp.*	100	501	1	WWTP
299073288	RCFBPv3_mp	*Ralstonia solanacearum*	95	201	1	WWTP
77019866	pREC1	*Rhodococcus erythropolis*	99	348	1	WWTP
145226750	pDK2	*Rhodococcus sp.*	97	483	1	WWTP
456362	pBT233	*S.pyogenes*	100	396	1	WWTP
18873674	pUR400	*S.typhimurium*	100	243	1	WWTP
10957190	R27	*Salmonella typhi*	100	246	1	WWTP
38259307	R478	*Serratia marcescens*	99	301	1	P32
117676079	Plasmid 1	*Shewanella sp.*	99	342	1	WWTP
18462515	pCP301	*Shigella flexneri*	100	105	1	WWTP
99035147	unnamed	*Silicibacter sp.*	96	466	1	P1
292677706	pUT2	*Sphingobium japonicum*	100	242	1	WWTP
110346757	pYAN-1	*Sphingobium yanoikuyae*	95	287	1	WWTP
291167465	pISP3	*Sphingomonas sp.*	100	350	1	WWTP
148550845	pSWIT02	*Sphingomonas wittichii*	95	421	1	WWTP
284005967	pSLIN06	*Spirosoma linguale*	96	141	1	WWTP
284005577	pSLIN01	*Spirosoma linguale*	97	277	1	WWTP
85057109	pMTSm1	*Stenotrophomonas maltophilia*	100	478	1	WWTP
251819044	pBM407	*Streptococcus suis*	95	313	1	WWTP
10956194	pER35	*Streptococcus thermophilus*	98	438	1	WWTP
54969619	pRSB101	*Uncultured bacterium*	100	314	1	Marina
317109853	PB11	*Uncultured bacterium*	95	200	1	WWTP
290791046	pAKD4	*Uncultured bacterium*	100	384	1	WWTP
205320734	pHHV216	*Uncultured bacterium*	100	388	1	WWTP

108859310	pTRACA	*Uncultured bacterium*	98	439	1	WWTP
84094946	pTP6	*Uncultured bacterium*	100	427	1	WWTP
28875495	pAK107	*Uncultured bacterium*	98	281	1	WWTP
19070006	pB4	*Uncultured bacterium*	95	457	1	WWTP
119416936	TC68	*Vibrio sp.*	100	453	1	Marina
154162790	pXAUT01	*Xanthobacter autotrophicus*	100	491	1	WWTP

CHAPTER 4: Incorporating metagenomics into Oceans and Human Health decision-making: Considerations for antibiotic resistance surveillance

4.1 Abstract

High-throughput genomic technologies now offer new approaches for environmental health monitoring. One potential application is metagenomic surveillance of antibiotic resistance determinants (ARDs) in the environment. While natural environments serve as reservoirs for antibiotic resistance genes that can be transferred to pathogenic and human commensal bacteria, environmental monitoring of these determinants has been infrequent and incomplete, focusing on a single or limited number of organisms or genes of particular concern. Furthermore, most surveillance efforts have not been integrated into public health decision-making. We utilized a metagenomic epidemiology-based approach to develop an ARD index that quantifies antibiotic resistance potential. This potential is defined by the abundance of antibiotic resistance genes, metal resistance genes, mobile genetic elements and pathogens. Our second objective was to analyze this index for common modalities across environmental samples. Thirdly, we explored how metagenomic data such as this index could be conceptually framed within a public health surveillance context. This study analyzed 25 samples from shotgun metagenomic projects that were sequenced using 454 pyrosequencing and consisted of microbial community DNA collected from natural and human impacted environments. A sample's ARD index was calculated using four different sequence similarity stringency classes. Principal component analysis was used to identify index patterns across samples. Significant differences were observed in the overall index and index sub-category levels when comparing highly to less impacted marine and freshwater environments. The index levels were greatly influenced by the

selection of the stringency classification. There were unique index sub-category fingerprints that distinguished the different metagenomes, including differences within the marine samples alone. Broad-scale screening of ARD potential using the ARD index reveals utility for framing environmental health monitoring. Developing screening metrics for ARD surveillance such as the number of metagenomic bases ml^{-1} will allow for comparative analyses of differently impacted environments that ultimately can form a basis for public health decision-making. Furthermore, the current approach holds promise as an initial screening tool for establishing baseline ARD levels in the environment that can be used to frame future decision-making regarding management of sources and exposure routes of ARDs.

4.2. Introduction

Advances in genomic technologies now offer novel approaches for environmental public health monitoring. High-throughput sequencing of whole microbial communities provides global snaphots of community and functional composition, as opposed to more conventional analyses that are species and gene specific. Because these new techniques rely on culture-independent approaches, they are able to access the genomic information of the >99% of bacteria that are not culturable (Amann et al. 1995). These technologies are also less labor intensive, reduce laboratory time and can generate massive volumes of genomic data in less than a day (Metzker 2010). Metagenomics, or the direct extraction, sequencing and analysis of DNA from a community of microorganisms (Handelsman 2004), is one high-throughput approach that in tandem with next generation sequencing has potential utility for environmental public health surveillance. While the public health applications of metagenomics remain to be fully elucidated, this approach has been used to track fecal contamination in watersheds via community

composition profiling (Wu et al. 2010), detect pathogens in wastewater (Ye and Zhang 2011) and identify indicators of sewage contamination (McLellan et al. 2010). A key step for scientific translation will be the development of frameworks for incorporating environmental metagenomic data into management or public health decision-making. While these techniques are promising, they pose a series of challenges for public health decision-makers. Determining the significance of a given genomic signal in the context of risk, defining the levels of genomic response that are needed to drive a decision, or identifying the cost-benefit balance of using these methods versus more traditional approaches, require investigation. In this paper we develop a public health framework that begins to address these questions and represents a first step towards developing a decision-monitoring tool.

One potential application of high-throughput metagenomics for public health surveillance involves characterization of antibiotic resistance determinants (ARDs) in the environment. ARDs refer here to the genomic factors related to the presence and dissemination of antibiotic resistance genes (ARGs). Antibiotic resistance occurs when bacteria evolve under selective pressure to confer resistance to antibiotics used to treat their infection. Antibiotic resistance is a global phenomenon and is a growing source of morbidity and mortality (Bush et al. 2011). While the majority of antibiotic resistance investigations have been focused on pathogenic bacteria in clinical settings, antibiotic resistance has been shown to be widespread in environmental bacteria (Wright 2010), and furthermore many resistance genes found in pathogenic bacteria have evolved or are sourced from resistance genes found in environmental microbial communities (Martinez 2009). The antibiotic resistomes (Wright 2007) of natural environments including soil, marine, freshwater and wastewater ecosystems have revealed an abundance of resistance genes and mobile genetic elements (MGEs). MGEs serve as vectors for the transfer of ARGs and other

virulence factors across bacteria and ecosystems. In many cases, these genes have been shown to be functionally resistant to selected antibiotics (Riesenfeld et al. 2004; Allen et al. 2009; Sommer et al. 2009; Donato et al. 2010; Torres-Cortes et al. 2011; Forsberg et al. 2012; Schmieder and Edwards 2012). The presence of resistance genes in the environment may be due to selective pressures favoring these genes, including antibiotic overuse and misuse in clinical treatment, agricultural and aquaculture applications and metal pollution. ARDs are ultimately disseminated into watersheds and coastal systems via sewage from wastewater treatment plants (WWTPs) and cruise ships, animal waste and urban/agricultural runoff, and thus form environmental reservoirs of ARDs (Davies and Davies 2010). Humans can be exposed to these reservoirs through consumption of contaminated food and drinking water, recreational activities such as swimming or direct contact with organisms carrying antibiotic resistant bacteria. Because of these widespread and generally uncharacterized reservoirs, there is a need for their identification, characterization and control (Bush et al. 2011).

Monitoring for antibiotic resistance in the environment has been infrequent and incomplete (Allen et al. 2010), and, unlike pathogen detection, has not been formalized into environmental public health surveillance and decision-making frameworks. There are currently no regulatory frameworks for assessing the public health risk of environmental ARDs. Global surveillance efforts have predominantly focused on the prevalence of antibiotic usage and antibiotic resistance isolates in clinical settings. For example, the European Antimicrobial Resistance Surveillance Network (EARS-Net) is a network of national surveillance systems that systematically collect data from a total of 900 public health laboratories serving over 1,400 hospitals in Europe. EARS-Net and other regional and global surveillance programs implemented by the World Health Organization monitor resistance trends in common

community and healthcare-associated bacterial pathogens (Grundmann et al. 2011). In the United States, the National Antimicrobial Resistance Monitoring System: Enteric Bacteria is a collaboration between the Center for Disease Control and Prevention (CDC), Food and Drug Administration (FDA) and Department of Agriculture that tracks antibiotic resistant isolates submitted by state and local public health laboratories.

The objectives of this study were three-fold. First, a metagenomic epidemiology-based approach was used to develop an index that quantifies the resistance potential of an environment. This multi-layered approach considers the entire microbiotic context for environmental antibiotic resistance by characterizing simultaneously the different levels of microbiome complexity that drive antibiotic resistance, including ARGs, genetic vectors for these genes and the species in which these genes occur (Baquero 2012). Secondly, the index was analyzed for common modalities across a diverse set of natural and human impacted marine and freshwater samples. The third objective was to integrate the index into a public health surveillance framework in order to provide an example by which high-throughput metagenomic data could potentially be utilized within a regulatory or management context.

4.3 Methods

The 25 metagenomes included in this analysis represent microbial community DNA collected from natural and human impacted marine and freshwater environments. These metagenomes were shotgun sequenced using 454 pyrosequencing (Table 4.1). The Puget Sound dataset was previously generated by our group and consists of open estuarine samples (P1, P5, P26, P28, P32) and an urban marina (Port et al. 2012). The other metagenomes were obtained from public databases including the National Center for Biotechnology Information (NCBI)

Sequence Read Archive (SRA) and MG-Rast (Meyer et al. 2008) and consist of: 12 California coastal surface water samples collected as part of the Global Ocean Sampling Expedition (Zeigler Allen et al. 2012), an urban freshwater lake (Oh et al. 2011), 4 Indian river sediment samples collected downstream from a wastewater treatment plant (WWTP) processing high volumes of antibiotics (Kristiansson 2011), effluent from a WWTP that discharges into Puget Sound (Port et al. 2012) and activated sludge from a domestic WWTP (Sanapareddy et al. 2009). The unassembled DNA sequence reads for each metagenome were run through a bioinformatic pipeline to quantify the ARD index (Figure 4.1). This pipeline consisted of 1) gene and taxonomic annotation through the MG-Rast server, 2) 16s rRNA annotation through the Ribosomal Database Project (RDP) Classifier (Wang et al. 2007) and 3) gene prediction, peptide prediction and subsequent protein searches against an expanded Antibiotic Resistance Genes Database (ARDB) (Liu and Pop 2009) and the Pfam 26.0 database (Punta et al. 2012).

4.3.1. Antibiotic resistance determinant index

Metagenomic data relevant to environmental surveillance of ARDs was classified into three categories: Gene transfer potential, antibiotic resistance gene potential and pathogenicity potential (Figure 4.2). A fourth category, source tracking, relates to identifying potential anthropogenic sources of resistance determinants through a community composition profiling, but is not included in the current index due to the need for further evaluation. This is also the case for virulence factors and human commensal bacteria that are also found in the environment and thus may be more likely to pass ARDs to humans. These categories comprise an ecological context for resistance potential; that is they provide both the prevalence of the resistance genes themselves in addition to the potential mechanisms by, and species in which, these genes may be

passed. As such, these categories represent factors that have potential relevance to human exposure risk. The three categories were quantified via their respective genomic sub-categories (Figure 4.2) using different sequence similarity stringency levels in order to generate a distribution of values for each sub-category. The stringency levels and associated sequence similarity criteria for each index sub-category are presented in Table 4.2. The high stringency level represents the most conservative approach for sub-category sequence annotation, followed by a moving threshold that gradually reduced the stringency and included medium-high, medium-low and low classifications. Based on the public health decision in question, different applications of the index are explored that consider these stringency criteria and the consequent balance of false positive to false negative sequence assignments. The following is a description of the rationale for and methods used to generate each sub-category.

Category 1: Gene transfer potential

Plasmids. Plasmids are mobile genetic elements that are important vectors for the dissemination of ARGs across bacteria (Boerlin and Reid-Smith 2008; Partridge et al. 2009). Individual sequence reads were searched against the NCBI RefSeq plasmid database using BLASTN (Altschul et al. 1990). Sequences with a nucleotide sequence identity ≥95% over an alignment length ≥100, 200, 300 and 400 bp were retained and grouped by alignment length (Kristiansson 2011; Zhang et al. 2011). The plasmid count was normalized to the total number of sequence reads per metagenome to obtain the proportion of sequences assigned to plasmids.

Transposable elements. Transposable elements (TEs) are also mobile genetic elements that allow for transfer of ARGs (Boerlin and Reid-Smith 2008; Partridge et al. 2009). TEs, unlike plasmids, are mobile segments of genomic DNA that can move to different locations on the

chromosome, onto plasmids, or can also transfer by conjugation to another bacterial cell and integrate into the recipient's genome (Salyers et al. 1995). To identify TEs, 167 unique genes annotated as transposable elements were downloaded from GenBank and searched against the metagenomic data using a threshold of 80% sequence identity over 150 bp. This criteria was then expanded as described for plasmids. The TE count was normalized to the total number of sequenced reads per metagenome. TEs were also identified through identification of TE-related protein motifs using the Pfam database. Pfam is a database of conserved protein families and domains across species, where the domains represent structural and functional motifs of the protein (Punta et al. 2012). 41 Pfams annotated as TEs were searched against the metagenomic reads from each dataset and those sequences with an E-value $\leq 10^{-10}$ were retained for further analysis. The number of TEs identified using the Pfam approach was normalized to the total Pfam count (i.e. for all possible assigned motifs from the Pfam database) for each metagenome.

Phages. Bacteriophages, or phages, are viruses that infect bacteria and are widespread in marine and freshwater ecosystems (Breitbart 2012). Phages are classified as mobile genetic elements, and also have the ability to transfer ARGs (Colomer-Lluch et al. 2011). Sequences were taxonomically annotated through the MG-Rast server using BLASTP and reads matching to the domain Viruses with ≥50% identity over an alignment length ≥50, 75, 100 and 150 amino acids were retained. The total phage count for each metagenome was normalized to the total number of sequences assigned at the domain level.

Category 2: Antibiotic resistance gene potential

Antibiotic resistance genes. ARGs were identified using the same approach as we have previously described (Port et al. 2012). We first compiled a nonredundant version of the ARDB

consisting of 3,185 sequences. The ARDB is a commonly used database consisting of over 23,000 resistance genes from nearly 1,700 species. These sequences were then searched against GenBank using an 80% identity cutoff (Liu and Pop 2009) to update the ARDB, increasing the number of nonredundant ARDB sequences to a total of 11,500 sequences. A list of 103 ARG sequences from metagenomic samples that were functionally verified to confer resistance (Allen et al. 2009; Sommer et al. 2009; Donato et al. 2010; Torres-Cortes et al. 2011) was also compiled and included in the expanded dataset. Predicted proteins from the metagenomic datasets were generated using MetaGeneMark (Zhu et al. 2010) and then BLASTP searched against the expanded ARDB (E-value $<10^{-5}$). Sequences with ≥80% identity and an alignment length ≥50, 75, 100 and 150 amino acids to an ARG were annotated according to the best hit. ARGs were normalized to the total number of sequence reads per metagenome.

Metal resistance genes. Substantial evidence indicates that metal contamination exerts selective pressure on the spread of ARGs (Baker-Austin et al. 2006). Genes encoding resistance to metals and antibiotics are frequently located together on the same mobile genetic element, resulting in co-selection of metal and antibiotic resistance. Sequences with similarity to metal resistance genes were identified by searching the SEED database subsystem 'Resistance to antibiotics and toxic compounds' (Overbeek et al. 2005). This subsystem contains genes and gene clusters encoding resistance to arsenic, mecury, cadmium and zinc. Metagenomic reads matching a metal resistance gene with ≥50% identity over an alignment length ≥50, 75, 100 and 150 amino acids were retained. Metal resistance gene sequences were normalized to the total subsystem count (i.e. for all assigned subsystems from the SEED).

Category 3: Pathogenicity potential

Pathogens. The greatest public health threat associated with antibiotic resistance is that conferred by bacterial pathogens. Transfer of ARGs to pathogens and subsequent human infection can result in infections that are difficult and in some cases not treatable with usual antibiotic therapies, leading to substantial morbidity and mortality. Two approaches were used to identify pathogens. First, 16S rRNA analysis was used to identify sequences with similarity to pathogenic bacteria. An *Escherichia coli* 16S rRNA reference sequence was searched against the metagenomic reads using BLASTN (default settings) and matching reads were then run through the Ribosomal Database Project (RDP) Classifier (Wang et al. 2007). Classifications with 50%, 80% and 95% confidence estimates were included in the analysis. The resulting sequences were then searched against the Microbial Rosetta Stone Database (Ecker et al. 2005) to identify bacterial pathogens that impact public health. Second, sequences were taxonomically annotated with the lowest common ancestor algorithm (LCA) using the MG-Rast server and reads matching to the species level with ≥95% identity and an alignment length ≥100 amino acids were retained and run against the Microbial Rosetta Stone Database.

4.3.2. Principal Component Analysis

The abundance counts for the index sub-categories were normalized to the total number of sequences in the index for a given metagenome. Principal component analysis was performed on the normalized data using the JMP v.10.0 statistical package (SAS Institute, Inc.). Eigen vectors and loading values were extracted for the first two principal components.

4.4. Results

4.4.1 Antibiotic resistance potential

The public health framework for ARD surveillance was assembled using a qualitative risk matrix approach. An ARD index was developed that consisted of three categories related to the molecular etiology of antibiotic resistance: gene transfer, antibiotic resistance gene and pathogenicity potential. To first compare the antibiotic resistance potential across the metagenomic samples, the index scores were calculated for each metagenome using four sequence similarity stringency levels (Table 4.2). By using different stringency levels, one can see how the index scores change with bioinformatic stringency criteria and consequently how these differences can impact public health monitoring and decision-making. Figure 4.3 shows that the high stringency classification generated the lowest percentage of index-positive sequences (mean =0.025%) for all samples except the Indian river sediment. As the stringency thresholds are reduced, this percentage increases to 0.033% (medium-high), 0.28% (medium-low) and 0.55% (low). This may be explained in part by the alignment length criteria exceeding the majority of reads from a given metagenome (e.g. activated sludge). The low stringency classification obtains the highest number of index-positive sequences but also potentially the most false positives. This is an important consideration for public health monitoring, where management decisions rely on robust data. Individual index sub-categories were also differentially sensitive to increases in alignment length and hence stringency (Figure 4.4).

As would be predicted, the most highly impacted environments had the highest cumulative ARD index scores at all stringency levels (Figure 4.3). The Indian river sediment samples taken downstream from a WWTP processing high volumes of antibiotics and to a lesser extent the WWTP effluent sample had significantly higher proportions of index-positive sequences. The dramatically elevated abundance of index-positive sequences in the Indian river sediment samples is due to an over-abundance of antibiotic resistance genes (ARGs), plasmids

and transposable elements (TEs) when compared to the other samples (Figure 4.4). While the activated sludge sample resembled the marine samples overall, there was an increased level of ARGs, plasmids and TEs compared to the marine and freshwater samples. The Puget Sound samples on average had a slightly increased cumulative score when compared to the California coastal samples. In particular, Puget Sound had significantly higher levels of metal resistance genes (MRGs), and to a lesser extent TEs, than the other marine samples. Pathogens were rarely detected in any of the samples, but from a public health perspective, any positive pathogen match is informative. Four pathogen sequences were detected in the effluent sample using the highest stringency confidence estimate criteria, while an additional pathogen sequence was identified when the stringency criteria was reduced to medium-low. These pathogens included *Clostridium* and *Legionella* species. *Enterococcus*, which is commonly used as an indicator species in marine waters, was not identified in the 16s rDNA sequences for any sample.

Multivariate analysis of all samples revealed ARGs and plasmids to be the most strongly correlated index sub-categories ($r=0.83$-0.98, $p<0.0001$) at all bioinformatic stringency levels (Table 4.3). TEs and ARGs were negatively correlated at the medium-high and high stringencies, while TEs and phages were negatively correlated at the low and medium-low stringencies. TEs and plasmids exhibited positive correlations at the lower stringencies and negative correlations at the higher stringencies. When considering the marine samples only, phages had strong negative correlations to both MRGs and TEs at the lower stringencies, ARG and TEs had a strong positive correlation at the low stringency, and as was the case for the analysis including all samples, TEs and plasmids switched from a significantly positive to negative correlation as the stringency was increased.

4.4.2 Antibiotic resistance determinant index patterns

We used principal component analysis (PCA) to identify modalities (or "fingerprints") for the index sub-categories for each metagenome within the set of samples. The medium-low stringency classification was applied for this analysis in order to maximize the number of sequences while reducing the false positive assignments. In the analysis of the full set of samples, PC1 was characterized by the presence of ARGs, plasmids and TEs and the relative absence of phages, while PC2 reflected high loadings for MRGs and pathogens and the relative absence of phages and ARGs (Figure 4.5A). There was a clear division between the marine, WWTP and Indian River sediment samples along PC1. PC2 distinguished the Puget Sound, California and WWTP effluent sample sets from one another. These first two principal components explained 71% of the variance in the underlying data. These results suggest that the Indian River sediment samples were characterized by the presence of ARGs and plasmids, and by the relative absence of MRGs and to a lesser extent phages. The WWTP effluent was characterized by the presence of MRGs, pathogens and to lesser extent TEs, and the relative absence of phages. The effluent and activated sludge were significantly different from one another along PC2, mainly due to a higher proportion of MRGs and pathogens in the effluent. Despite the inclusion of the varied sample types in the PCA, the marine locations could still be distinguished from one another. Puget Sound had PC1 and PC2 loadings consistent with the presence of MRGs and phages and the relative absence of ARGs and plasmids, while the California coastal samples were characterized by phages and the relative absence of MRGs, pathogens and to a lesser extent TEs. The freshwater lake samples had similar loading profiles to those from Puget Sound, but with increased PC2 loadings, signifying a higher proportion of MRGs and pathogens.

To further distinguish the modalities of the marine samples, PCA was repeated on the Puget Sound and California coastal samples only. In this analysis PC1 reflected high proportions of TEs, plasmids, ARGs and the relative absence of phages, while PC2 reflected the presence of phages and plasmids and a modest loading of ARGs as well as the relative absence of MRGs. As with the previous analysis, Puget Sound was divided from the California coastal samples along both PC1 and PC2 (Figure 4.5B). In the marine only analysis the primary Puget Sound cluster was slightly positive in PC1 and negative in PC2, thus it had a higher proportion of TEs, plasmids, ARGs and MRGs and a relative absence of phages. California coastal samples generally exhibited the opposite profile (negative in PC1 and positive in PC2), with a greater proportion of phages but a relative lack of plasmids, TEs, ARGs and MRGs. These samples also had a lower proportion of TEs and plasmids. Outliers included sample P26 (from Puget Sound), which grouped tightly with the California cluster, as well as samples P5 (from Puget Sound) and GS260 (from California coast), which were separated from their respective clusters. The fact that samples P26 and P5 are from locations that lie closer to open ocean than any others within the Puget Sound dataset may explain this pattern. P26 is located within the Strait of Juan de Fuca which connects Puget Sound to the Pacific Ocean, while P5 is located in an area of Puget Sound that experiences heavy mixing of oceanic and Sound waters.

4.5 Discussion

This study investigated the potential for high-throughput sequence data from environmental microbial communities to be informative for public health concerns such as antibiotic resistance. A metagenomic index of ARDs was developed and shown to differ across both diverse environmental samples and also within a group of marine samples. Significantly

impacted environments, including WWTP effluent, and river sediment collected downstream from a WWTP processing high volumes of pharmaceuticals, had the highest cumulative index scores. These samples were distinguished by multiple factors including higher potential for gene transfer, pathogenicity and the presence of antibiotic resistance genes. While the presence of antibiotic resistance genes may not pose a direct human health risk, characterizing environmental reservoirs of resistance genes is important for predicting potential transfer of these genes to bacteria that could ultimately colonize or infect humans. Less impacted environments, including marine samples and a freshwater lake, had indices likely reflecting reduced public health concern while exhibiting a distinct fingerprint characterized by either phages or metal resistance genes depending on location. Increased abundance of metal resistance genes may be indicative of metal pollution from anthropogenic sources, and at the molecular level metal resistance genes are commonly associated with the presence of antibiotic resistance genes on plasmids. The coastal samples had a greater proportion of phages than the Puget Sound samples, while the Sound samples were characterized by an increased abundance of metal resistance genes. Phages are ubiquitous in the marine environment (Breitbart 2012), thus any link to actual dissemination of antibiotic resistance genes will require more targeted investigations. The ability of phages to act as vectors of antibiotic resistance gene transfer in marine ecosystems though warrants their preliminary inclusion in the ARD index in order to establish abundance trends across differentially impacted environments. The variation represented by the index accounts for common index sub-category modalities or patterns across sample type that may related to anthropogenic inputs. This is important in terms of identifying potential sources for antibiotic resistance potential at a given location and for establishing management criteria to minimize this potential.

This study also shows that the choice of sequence similarity criteria for annotating metagenomic data has a significant impact on the number of index-positive sequences. There was a significant decrease in the number of index-positive sequences for each sample and index sub-category with each increasing stringency class. This trend may be related to sequence read length in that index-positive sequences at the lower stringencies may be too short to reach the alignment length criteria of the higher stringencies (e.g. activated sludge sample) or that the lower stringencies assign a significant number of false positives. As the high sequence similarity stringency classification for a number of the index sub-categories matched or exceeded the criteria used in other studies investigating antibiotic resistance gene and gene transfer in environmental water samples (Kristiansson 2011; Zhang et al. 2011), further bioinformatic and laboratory analyses are needed to determine the most robust criteria for annotating ARDs in metagenomic data. This is especially important if these data are to be applied within a public health surveillance context.

4.5.1. Applications to public health surveillance and management

Water quality management decisions have not specifically considered ARDs or antibiotics, likely because of a lack of data and the uncertainty regarding risk. Given the global magnitude of antibiotic resistance and the emergence of multi-drug resistance bacterial strains, information pertaining to the status and trends in ARDs in the environment is needed. Public health management decisions pertaining to water quality that may benefit from information regarding ARD potential include actions aimed at reducing the sources and exposure routes of ARDs and framing of adaptive monitoring protocols (Figure 4.6). Source control of ARDs entering coastal environments primarily involves waste management of antibiotics and the

regulation of antibiotic use in agriculture, aquaculture, hospitals and households. Exposure control of ARDs may involve beach or shellfish bed closures or advisories or aquaculture siting. Due to the uncertainty between exposure to ARDs and actual human health risk, current applications of the ARD index as a potential initial environmental screening tool for informing public health decisions are best suited to ARD source control. Here we provide an example demonstrating the potential application of a high-throughput metagenomic approach (e.g. ARD index) for water quality monitoring in recreational marine waters.

Current regulatory standards for water quality in recreational waters are based on culture-based methods that are highly sensitive and specific for targeted organisms of interest (Figure 4.7). For example, beach and shellfishery closures in Washington State occur when fecal coliform levels exceed a geometric mean of 14 colony forming units (CFUs) per 100 ml marine water or enterococci levels exceed a geometric mean of 70 CFUs per 100 ml marine water (Washington Administrative Code 173-201A-210). While this targeted approach is appropriate for public health regulatory decisions in environmental and clinical contexts, early risk management may benefit from more broad-based, population level screening. High-throughput metagenomic screening of environmental samples would allow for initial profiling of a portfolio of pathogenic and non-pathogenic microbial communities and genes simultaneously, including the vast majority of bacteria that are unculturable. Test sensitivity and specificity are defined differently when applied to environmental screening (Figure 4.7). Screening using more traditional methods such as culture entails high sensitivity but low specificity. The fact that these techniques are for the most part organism or gene specific leads to a low number of false positives due to their high accuracy in correctly identifying target organisms, but at the same time increases false negatives due to the limited number of organisms or genes that are

detectable. The false negative rate is also potentially expanded by the fact that the vast majority of bacteria are unculturable. On the other hand, high-throughput metagenomic screening may yield higher false positives than culture-based methods due to sequence misannotation or identification of sequences from cells that are not viable. Furthermore, the massive amounts of data generated by next generation sequencing allow for profiling of tens to hundreds of thousands of organisms and genes in a sample, thereby increasing the likelihood of detecting their presence and reducing the false negative rate. For more specific metagenomic screening though (e.g. ARD index), the false negative rate may increase due to limitations in sequencing depth. One can begin to understand the trade-offs in sensitivity and specificity by using an optimization function directly related to the public health question. For example, by applying a moving sequence similarity stringency threshold for the ARD index, the sensitivity of the index to answer specific risk questions that may require more or less conservative criteria can be analyzed. A certain percentage of false positives may be accepted when using the ARD index in order to gain a broader understanding of the antibiotic resistance potential of an environmental sample and furthermore to detect the emergence of ARDs in the environment. This screening tool could then trigger additional monitoring or risk assessment. Cost and time are currently important considerations for high-throughput metagenomic surveying, but genome sequencing is becoming more cost-effective and rapid and there is potential for automated in situ systems in the future.

To begin to frame metagenomic screening data within a risk and decision context for ARD source or eventually exposure control, a metric that relates sequence abundance to both sequencing depth and water volume can be generated. The following equations calculates the

abundance of a public health genomic marker of interest (X) in a metagenomic sample using screening data:

1) Fraction of total bases sequenced annotated as X = No. of bases sequenced for X / Total no. of bases sequenced

Equation 1 determines the fraction of bases sequenced that were annotated as belonging to X by normalizing to sequencing depth. The number of bases sequenced for X can be determined by summing the bases for all X annotated sequences.

(2) No. of bases in 1 ml of seawater = 10^6 bacterial genomes/ml x effective genome size

Equation 2 estimates the number of bases in 1 ml of seawater by multiplying the number of bacterial genomes in 1 ml of seawater (10^6 genomes) (Gilbert and Dupont 2011) by the effective genome size for the sample. The effective genome size is an estimate of the average genome size for a sample containing mixed communities of organisms (Raes et al. 2007).

(3) Predicted fraction of bases in 1 ml of seawater annotated as X = Equation 1 x Equation 2

Equation 3 predicts the number of X-related bases one would expect to find in 1 ml of seawater by multiplying the results from equations 1 and 2.

For example, the following calculations estimate the antibiotic resistance gene (ARG) signal in the WWTP effluent sample from the Puget Sound dataset:

(1) Fraction of bases sequenced annotated as ARG-related in WWTP effluent sample =

4,629 ARG-related bases / 46,046,577 total bases sequenced = 9.9×10^{-5}

(2) No. of bases in 1 ml of seawater = 10^6 genomes/ml x 3×10^6 bases/genome =

3×10^{12} bases/ml

(3) Predicted fraction of bases in 1 ml of seawater annotated as ARG-related =

$9.9 \times 10^{-5} \times 3 \times 10^{12}$ bases/ml = 3×10^8 bases/ml

If related to the volume of water commonly used for regulatory purposes (100 ml), the predicted fraction of ARG-related bases would equal 3×10^{10} bases per 100 ml. These equations can be used to further calculate the abundance of the remaining ARD index sub-categories in the effluent sample, and the cumulative abundance of ARD index positive sequences for the effluent sample.

This metric is informative for environmental health monitoring as it relates the abundance of a genomic marker of interest to a sampled volume of seawater, freshwater or wastewater. As such it provides an initial approach for comparative analyses across samples that can distinguish differently impacted environments and ultimately form a basis for public health decision-making. A detection rate can also be estimated using equation 1, and for the WWTP effluent example above, one ARG-related base is detected approximately every 10,000 bases sequenced. Extrapolating from this ratio, to detect one ARG (~1,000 bases), a sequencing depth of 4 Mb is needed. This detection rate is likely to be much smaller for marine samples. Quantitative microbial risk assessments (QMRA) have used gene abundance counts (i.e. genome copies liter^{-1}

detected via qPCR) for pathogenic markers in fecally-contaminated recreational waters to determine pathogen dose, and this information is then used to estimate the human health risk due to ingestion of these waters (Staley et al. 2012). Conceptually this approach has relevance to microbial risk assessment using high-throughput metagenomic approaches, and the metric described above (i.e. metagenomic bases ml^{-1}) begins to lay out an approach that could eventually be informative for estimating dose across microbial communities.

4.5.2. Future data needs and applications

While metagenomics in tandem with next generation sequencing provides a valuable tool for identifying trends in data and making comparisons across samples, quantification of the full complement of specific genes or functions in the data will require a more optimal balance between sequencing depth and read length. For example, ARGs or pathogens, which appear to be rare in surface water microbial communities (Port et al. 2012), are likely underestimated considering the limit of sequencing depth for current 454 pyrosequencing (~500-600 Mbp). In terms of public health surveillance and environmental screening, this may lead to an increased false negative rate. For highly impacted environments, the detection rate using this 454 technology was much higher. For example, the detection rates of ARGs for the downstream Indian river sediment sample taken nearest to the WWTP discharge site and the Puget Sound WWTP effluent sample were 1 ARG sequence/85 sequence reads and 1 ARG sequence/9300 sequence reads respectively. Given the current rate of development for sequencing technologies, it is likely that a more complete quantification of selected metagenomic markers balance will be possible in the near future.

In addition, the ARD index is a high-throughput measure of ARD potential, and as such cannot be directly related to human health risk. For environmentally-sourced ARGs to pose a health risk, they must be transferable via mobile genetic elements, be transferred to either pathogenic or commensal bacteria that then infect or colonize humans and confer resistance to antibiotics of clinical importance. Qualitative risk assessment can be used to predict the potential for these conditions to be met but the actual health risk cannot be quantified without more detailed molecular data and fate and transport modeling. A qualitative assessment could identify specific ARGs and predict the likelihood for transfer by classifying transfer potential as low, medium or high based on known molecular data for that gene. Proximity to human exposure routes, such as the distance to recreational beaches or food supplies such as shellfish beds or other mariculture operations could then be used as proxies for potential receptor interactions and thus human exposure risk.

Trends in the ARD index or other metagenomic markers of interest will also benefit from increased temporal and spatial sampling. The number of next generation sequenced shotgun metagenomes is currently limited likely due to high cost and the required bioinformatic infrastructure. This approach is quickly becoming more cost-effective though, and there are a number of publicly available annotation and analysis pipelines specifically designed for metagenomic data. Larger sample sizes representing a diverse mix of environments, capturing temporal trends and different next generation sequencing technologies such as the Illumina platform would allow for a more comprehensive profiling of the ARD index.

Given these future data needs, current applications of the ARD index should focus on public health management decisions associated with source control and not exact quantitation of human health risks. Longitudinal monitoring and screening of the abundance of ARD index sub-

categories associated with WWTP and cruise ship effluent and discharge sites, freshwater inputs such as river mouths and coastal aquaculture operations could provide baseline environmental levels for anthropogenically-sourced ARDs. This data could be incorporated into a value of information (VOI) decision framework to evaluate the potential for high-throughput metagenomics to inform a decision to, for example, reduce ARD dissemination into the environment by improving WWTP technologies or reducing the use of activated sludge sourced from WWTPs as fertilizer for agricultural crops. VOI is used to evaluate how the acquisition of additional monitoring information (i.e. metagenomic data) is able to reduce uncertainty that matters to, or thwarts, decision management (Raiffa 1968). The availability of increased monitoring data pertaining to ARD levels in wastewater or sludge would benefit these decisions that currently do not account for the potential risk associated with antibiotic resistance release into the environment. In order to actually quantify the value of the information provided by ARD index a decision analytic framework that allows for the identification of the full range of data gaps and uncertainties would need to be implemented.

4.6. Conclusions

This study had three objectives, to develop a metagenomic ARD index that quantifies the antibiotic resistance signal within marine and freshwater environments, analyze this index for common modalities across environmental samples and thirdly conceptually frame the index within a public health surveillance context. Significant differences were seen in the overall index and index sub-category levels when comparing highly to less impacted marine and freshwater environments. Furthermore, there were unique index sub-category fingerprints across the different metagenomes. We developed a metric using metagenomic data to show how broad-

scale screening of the ARD potential of an environment has the potential to inform environmental health monitoring. The ultimate application of this metric for surveillance will require increased sequencing depth and sampling in order to more thoroughly characterize antibiotic resistance potential in water environments. Furthermore, to define index threshold levels of concern and furthermore link these levels to actual exposure and risk management will require a better understanding of the prevalence and mobility of ARGs in the marine environment. Nevertheless, this approach holds promise as screening tool for establishing baseline ARD levels in the environment that can be used for future decision-making regarding controlling and monitoring sources and exposure routes of ARDs.

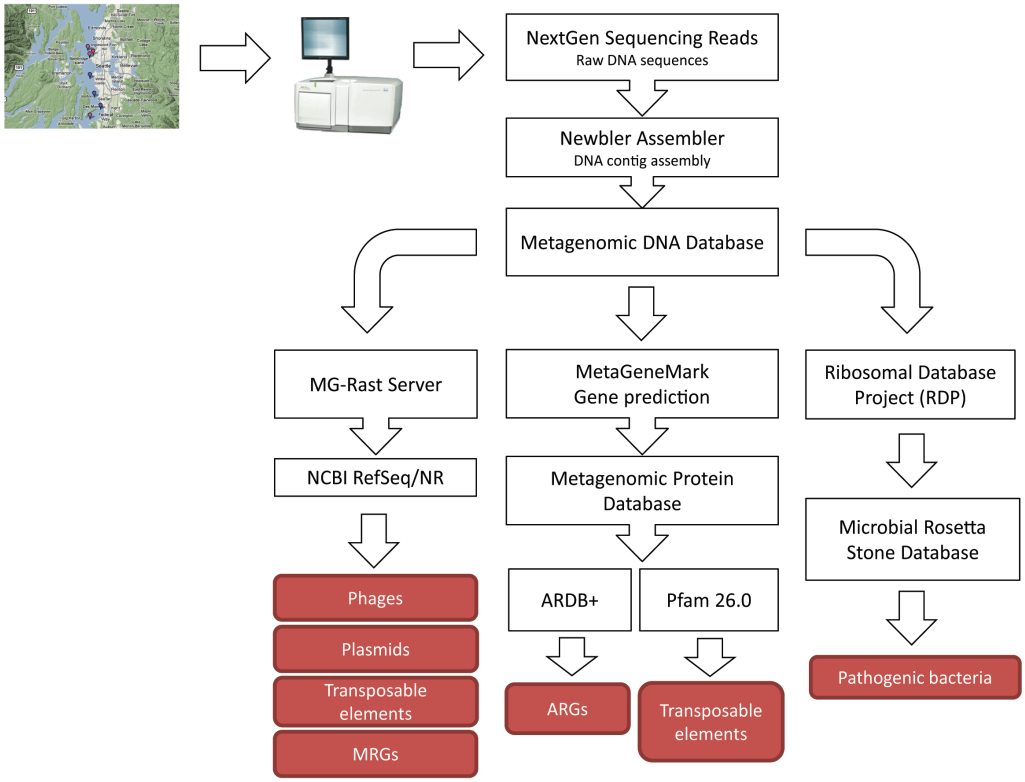

Figure 4.1. Bioinformatic framework for quantifying the metagenomic index of antibiotic resistance determinants (ARDs). Index sub-categories are shown in red. Abbrevations: ARDB+, expanded Antibiotic Resistance Genes Database; ARG, antibiotic resistance gene; MRG, metal resistance gene; NCBI, National Center for Biotechnology Information; NR, nonredundant database; RefSeq, Reference Sequence database.

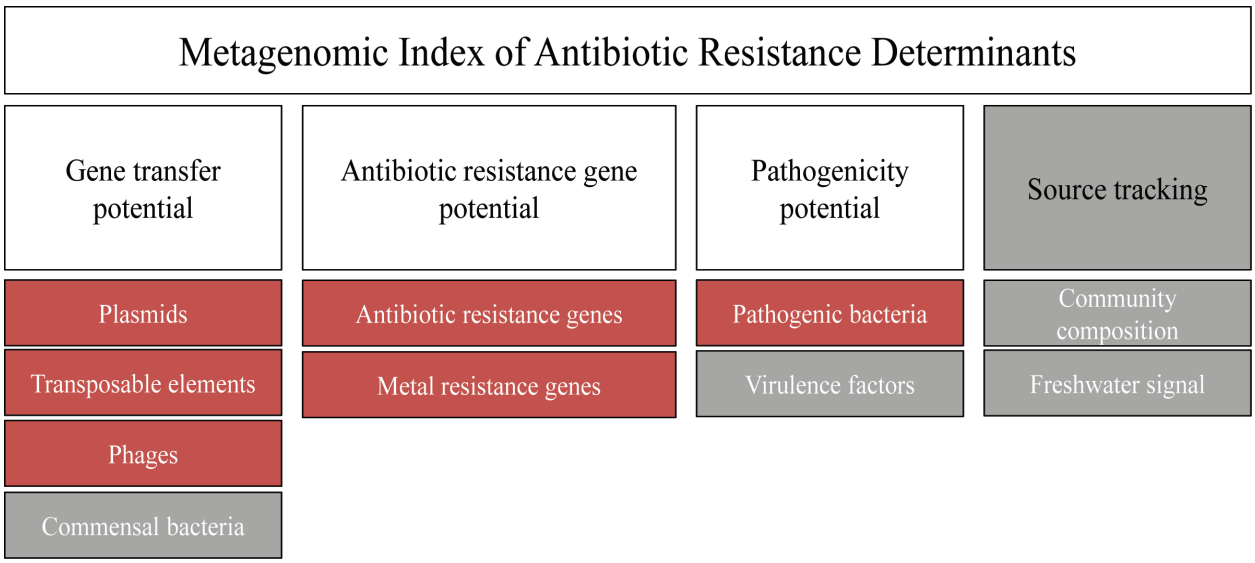

Figure 4.2. Metagenomic index of antibiotic resistance determinants for public health surveillance. The white boxes denote the primary index categories, while the red boxes represent the quantifiable sub-categories. The gray boxes are categories and sub-categories that have not yet been incorporated into the index but have the potential to be upon further investigation.

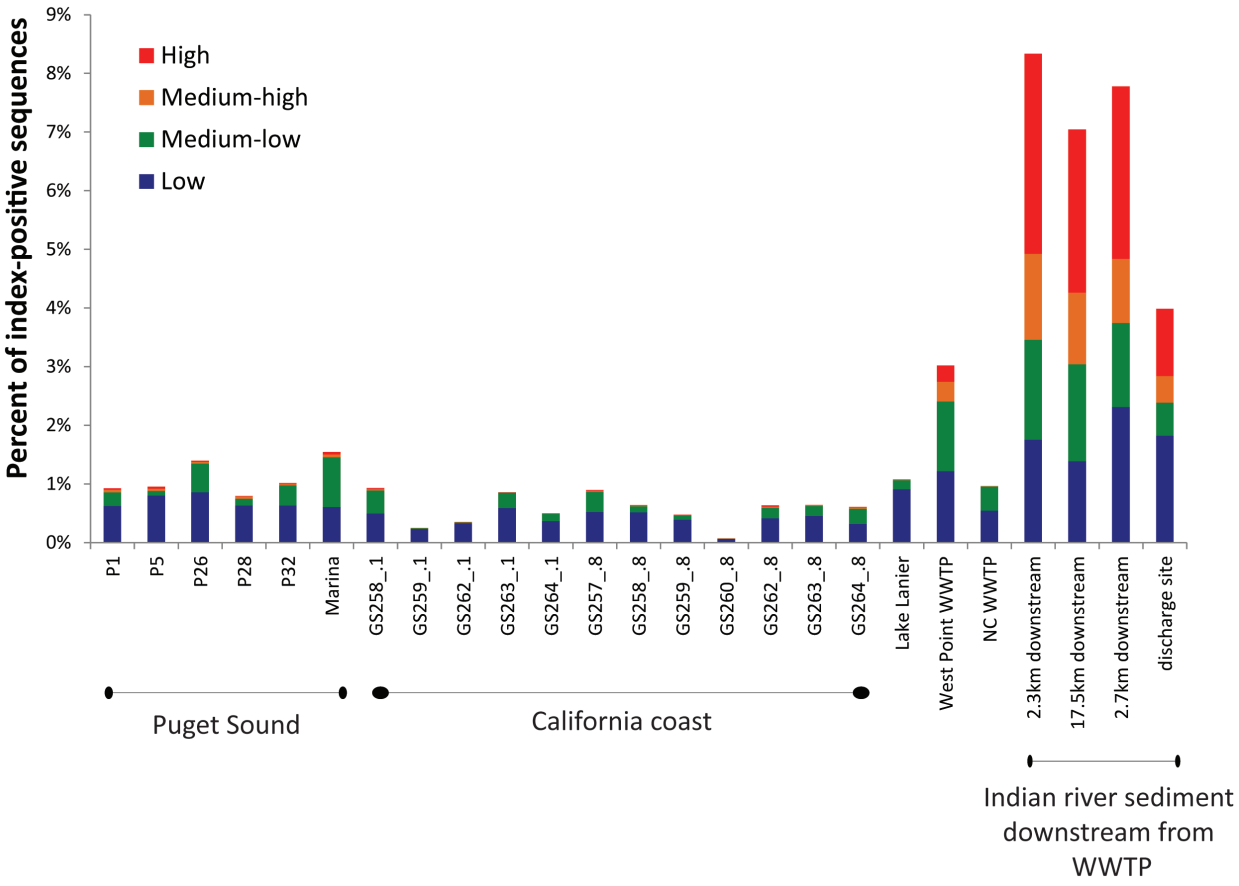

Figure 4.3. Proportion of total sequence reads per metagenome that match sequences within the antibiotic resistance determinant (ARD) index. The proportions are shown for four different sequence similarity criteria, including high, medium-high, medium and low sequence stringencies.

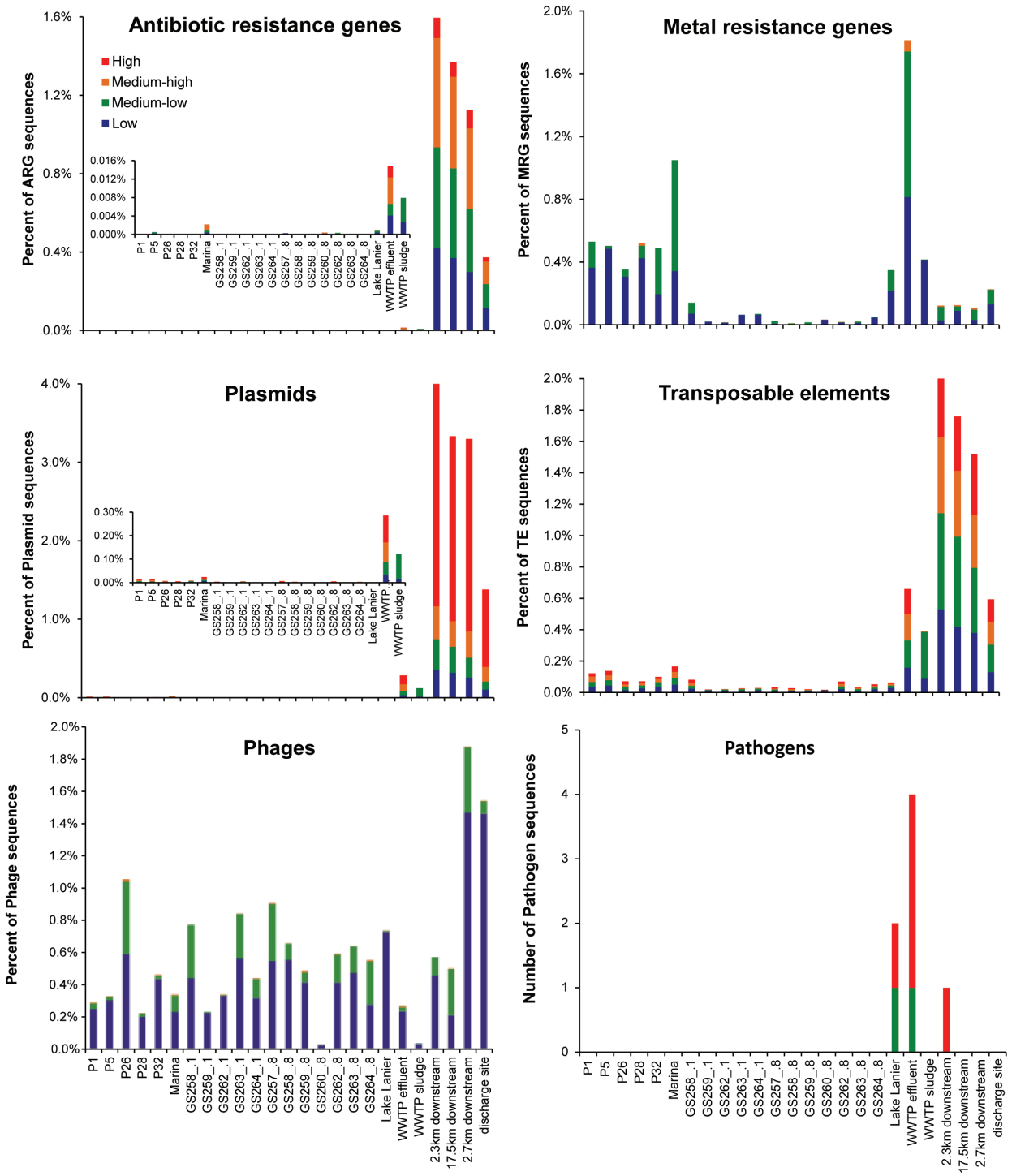

Figure 4.4. Proportion of total sequenced reads per metagenome assigned to each index sub-category at the different bioinformatic stringency levels. For the pathogen sub-category, the number of pathogen-annotated sequences is shown instead as any positive signal for this category may be of direct public health concern.

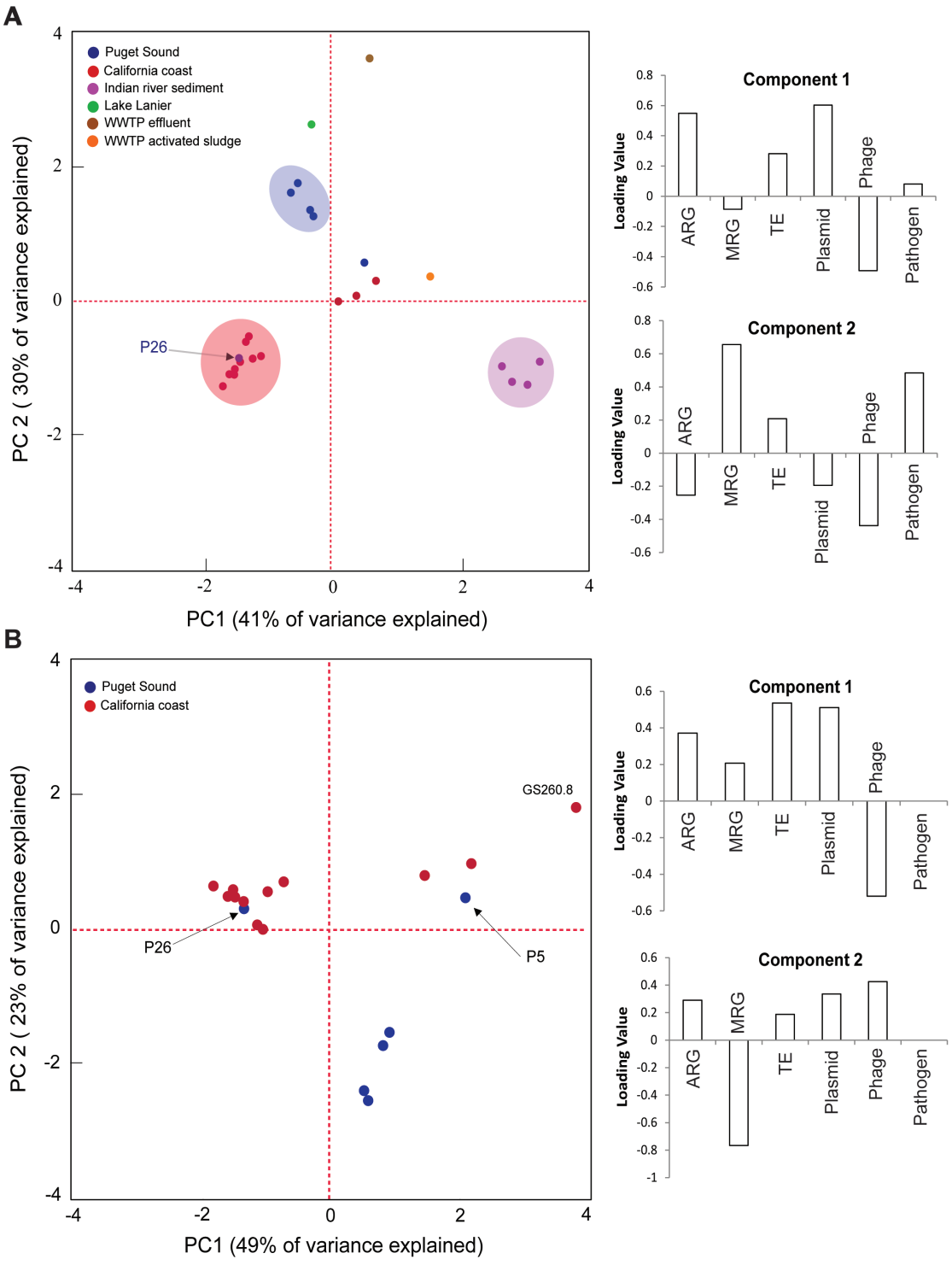

Figure 4.5. Principal component analysis and corresponding loading values of index footprints for (A) all metagenomic samples and (B) marine samples. The medium-low sequence similarity stringency level was used for this analysis. Abbreviations: PC1, principal component 1; PC2, principal component 2.

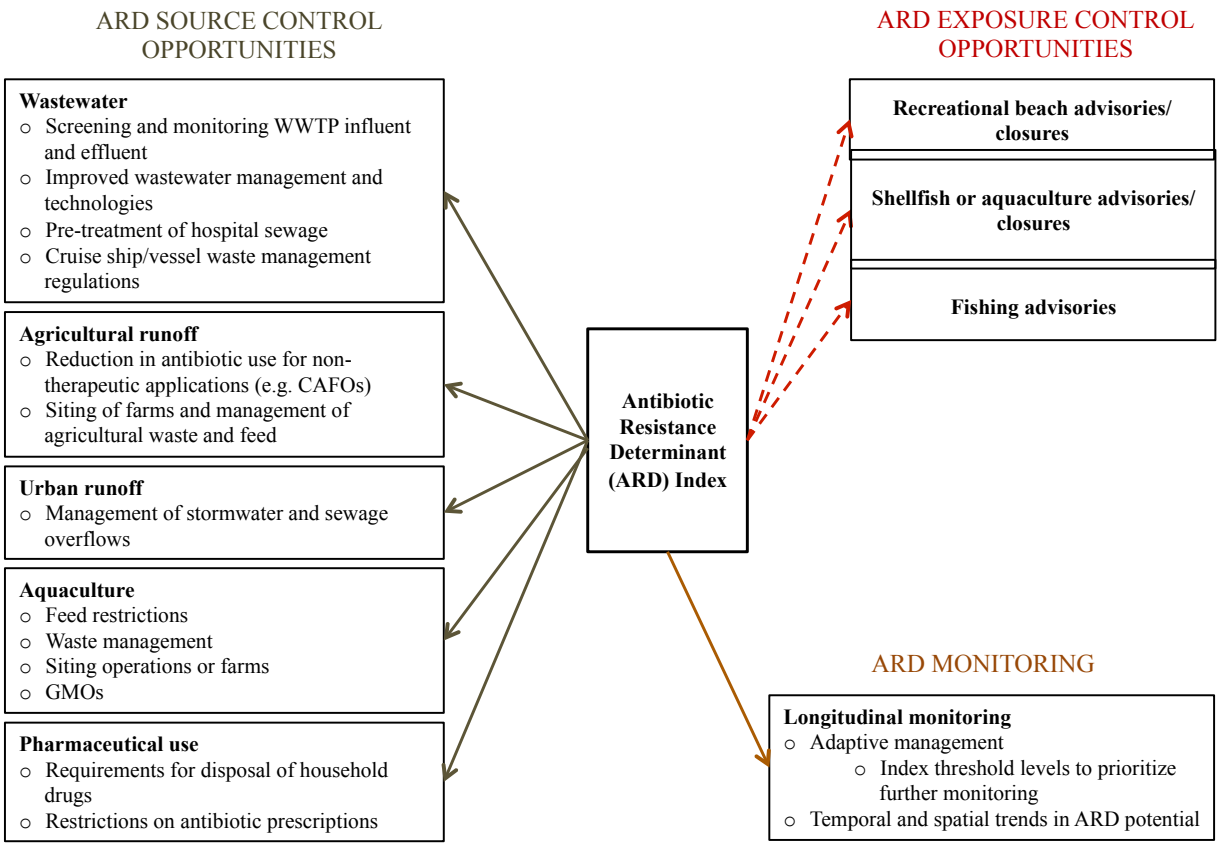

Figure 4.6. Public health management decisions and opportunities pertaining to the control and monitoring of potential sources of and exposures routes for environmental antibiotic resistance using the antibiotic determinant (ARD) index. Dashed lines relating to exposure control opportunities reflect the fact that the index cannot be directly related to exposure risk but can potentially be informative for decisions pertaining to human health. Abbreviations: CAFO, concentrated animal feeding operation; GMO, genetically modified organism; WWTP, wastewater treatment plant.

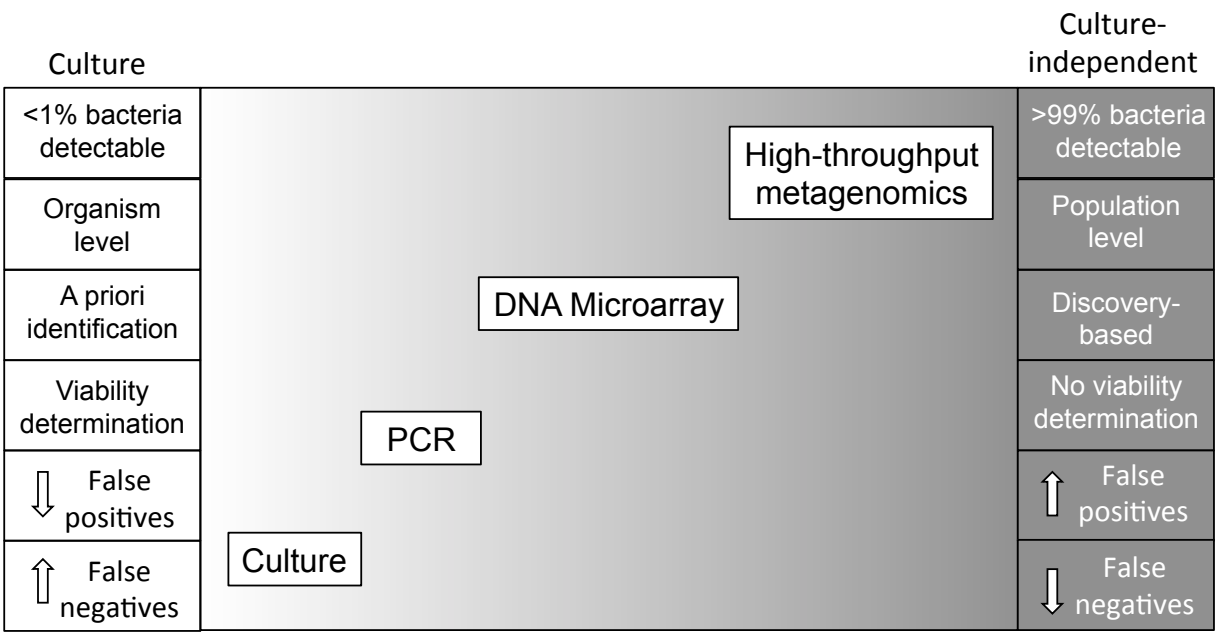

Figure 4.7. Methods for broad-scale screening of public health signals in environmental samples. The gradient represents a spectrum from more culture-based, organism specific methods to culture-independent, population level approaches. Culture-based methods are highly sensitive for targeted organisms of interest but only access a small proportion of organisms and genes present while culture-independent approaches sacrifice sensitivity for increased specificity. Environmental screening of genomic information using high-throughput metagenomics can be used to trigger further monitoring or action relevant to public health.

Table 4.1. Metagenomic samples used in this study.

Project	No. of samples	Environment	Size fraction (μm)	Depth (m)	Mbp	Mean read length (bp)	Reference
Puget Sound	6	Marine	0.2-3	5	504	370	Port et al. 2012
California coast	12	Marine	0.1-0.8/ 0.8-3	2	1940	551	Allen et al. 2012
Lake Lanier	1	Freshwater	0.22-1.6	5	502	395	Oh et al. 2011
Indian river sediment	4	Freshwater sediment downstream from WWTP	No size fractionation	Unknown	91	365	Kristiansson et al. 2011
Effluent	1	WWTP	0.2-3	N/A	46	381	Port et al. 2012
Activated sludge	1	WWTP	No size fractionation	N/A	95	250	Sanapareddy et al. 2009

Abbreviations: N/A, not applicable; WWTP, wastewater treatment plant.

Table 4.2. Sequence similarity criteria for the varying stringency classification levels used to quantify the index sub-categories.

	Sequence similarity stringency classification			
Index	**High**	**Medium-high**	**Medium-low**	**Low**
Gene transfer potential				
Plasmids	95% ID; ≥ 400 bp	95% ID; ≥ 300 bp	95% ID; ≥ 200 bp	95% ID; ≥ 100 bp
Transposable elements	80% ID; ≥ 120 aa	80% ID; ≥ 90 aa	80% ID; ≥ 60 aa	80% ID; ≥ 30 aa
Phages	50% ID; ≥ 150 aa	50% ID; ≥ 100 aa	50% ID; ≥ 75 aa	50% ID; ≥ 50 aa
Antibiotic resistance gene potential				
Antibiotic resistance genes	80% ID; ≥ 150 aa	80% ID; ≥ 100 aa	80% ID; ≥ 75 aa	80% ID; ≥ 50 aa
Metal resistance genes	50% ID; ≥ 150 aa	50% ID; ≥ 100 aa	50% ID; ≥ 75 aa	50% ID; ≥ 50 aa
Pathogenicity potential				
Pathogens	RDP 95% CE	RDP 95% CE	RDP 80% CE	RDP 50% CE

Abbreviations: CE, confidence estimate; ID, identity; RDP, Ribosomal Database Project.

Table 4.3. Pearson's correlation coefficients (r) for the index sub-categories. (A) Multivariate analysis including all samples, (B) Multivariate analysis including only marine samples. The values for each sequence similarity stringency (low|medium-low|medium-high|high) are shown for each comparison.

A.

	All samples					
All	ARG	MRG	Plasmid	TE	Phage	Pathogen
ARG						
MRG	-0.36|-0.31| -0.16|ND					
Plasmid	0.98***|0.94***| 0.90***|0.83***	-0.31|-0.34| -0.07|ND				
TE	0.48|-0.01| -0.78***|-0.72***	0.26|-0.07|- 0.17|ND	0.58*|0.23|- 0.87***|-0.77***			
Phage	-0.52*|-0.45| -0.36|-0.19	-0.57*|-0.47| -0.07|ND	-0.59*|-0.56*| -0.39|-0.17	-0.82***|-0.55*| -0.03|0.11		
Pathogen	0|-0.02|-0.06|0.02	0.30|0.38| 0.08|ND	0.07|0.02|-0.04| -0.09	0.24|0.02|0.12| 0.13	-0.31|-0.30| -0.21|-0.11	

B.

	Marine samples					
Marine	ARG	MRG	Plasmid	TE	Phage	Pathogen
ARG						
MRG	0.34|0.01| 0.22|ND					
Plasmid	0.68*|0.52|0.27| ND	0.61*|0.09|0.20| ND				
TE	0.82***|0.40| -0.25|ND	0.61*|0.02|0.52| ND	0.81***|0.81***| -0.49|-0.97***			
Phage	-0.46|-0.30| -0.11|ND	-0.98***|-0.75**| -0.19|ND	-0.69*|-0.51| -0.29|-0.09	-0.72**|-0.66*| -0.52|-0.16		
Pathogen	ND|ND|ND|ND	ND|ND|ND|ND	ND|ND|ND|ND	ND|ND|ND|ND	ND|ND|ND|ND	

Abbreviations: ARGs, antibiotic resistance genes; MRGs, metal resistance genes; ND, non-detect; TE, transposable elements. *p<0.01. **p<0.001. ***p<0.0001.

CHAPTER 5: Significance of Research Findings and Implications for future OHH GxE investigations

For this work, we sought to test the overall hypothesis that environmental genomic information can provide sensitive and functional markers of human impacts on marine ecosystems which can then be used to improve our understanding of how the composition of micro-organism communities relates to public health. We utilized a gene-environment (GxE) approach whereby the complex interplay between genetic and environmental factors in marine ecosystems can provide insight into potential OHH concerns. In Chapter 1, this GxE framework was presented within the context of Oceans and Human Health (OHH) investigations. The following chapters then applied framework to specific OHH concerns. In Chapter 2, we investigated the repertoire of GPCR signaling pathway proteins in sequenced diatoms in order to provide insight into potential signaling mechanisms related to environmental perception and response in these organisms. The presence and expression of GPCRs and GPCR signaling pathway proteins in these diatoms warrants further investigation into their potential OHH relevance regarding diatom stress response or bloom formation. In Chapter 3, we used metagenomic pyrosequencing to show differences in microbial composition and antibiotic resistance determinant signals across differentially impacted environments, including marine, nearshore and wastewater ecosystems. The environment serves as a reservoir for resistance genes that can be disseminated to pathogenic bacteria in clinical or environmental settings, and this study provides baseline environmental data for the prevalence and distribution of antibiotic resistance genes and their genomic vectors. In Chapter 4, we developed a metagenomic index for

screening antibiotic resistance determinants in the environment and conceptually framed this index within a public health surveillance and decision-making context.

This work is of scientific value since few studies have examined GxE interactions within an OHH context. Only recently have advances in sequencing technology dramatically increased our ability to obtain and analyze large volumes of genomic and metagenomic data from marine micro-organism communities. This information is now paving the way for novel and integrated approaches for fields such as environmental health monitoring. The approaches and results presented here lay out a foundation for conceptually framing marine GxE interactions in the context of public health. Making these OHH links requires cross-discipline investigations that integrate databases, tools and concepts from oceanography, human biology, environmental microbiology, environmental health and risk assessment. Furthermore, the results presented here rely to a great degree on bioinformatic annotation of sequence data, and thus aside from experimental confirmation, there is a need for standardization of annotation pipelines and criteria. This is especially important for the identification of public health relevant signals such as antibiotic resistance genes, pathogens or signaling proteins that may related to environmental stress response. Chapter 4 begins to lay out how results may change given different sequence similarity criteria, and why initial public health screening applications may favor an optimal balance of test sensitivity and specificity.

GxE investigations relevant to OHH will continue to gain traction as sequencing technologies continue to advance and become less expensive and as more environments are explored. An emphasis though must be placed on increasing the spatial and temporal extents of genomic and metagenomic sampling. As the majority of samples are currently collected from coastal and open ocean locations, there is limited application for GxE investigations relating to

anthropogenic impacts in nearshore areas. More longitudinal genomic data is needed for human-impacted areas, including recreational beaches, shellfish harvesting areas, aquaculture pens and wastewater outfall and freshwater input points. The collection of time-series data across locations in tandem with sequencing technology that allows for greater depth of sequencing into community structure and function will ultimately allow for greater utility of GxE for public health monitoring applications.

As depth and read length increase for next generation sequencing technologies, the utility of the ARD index may move beyond an initial environmental screening tool. These advances will better enable the index to be used for human health risk assessment via more robust predictions of antibiotic resistance gene transfer and potential across bacteria and organisms relevant to OHH. For example, future high-throughput metagenomic datasets may allow one to more fully answer questions such as "Of the plasmids identified, which are associated with pathogens and are also known to carry antibiotic resistance genes?" or "Which antibiotic resistance genes are associated with plasmids or transposable elements within the dataset?". Furthermore, increased read length and sequencing depth will enable deeper profiling of the resistomes of marine microbial communities and may therefore shed insight into potential associations between resistance gene abundance and community composition.

REFERENCES

Aguilo-Ferretjans MM, Bosch R, Martin-Cardona C, Lalucat J and Nogales B. 2008. Phylogenetic analysis of the composition of bacterial communities in human-exploited coastal environments from Mallorca Island (Spain). *Syst Appl Microbiol* 31(3): 231-240.

Albertsen M, Hansen LB, Saunders AM, Nielsen PH and Nielsen KL. 2012. A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. *ISME J* 6(6): 1094-1106.

Allen AE, Dupont CL, Obornik M, Horak A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu HH, Fernie AR and Bowler C. 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* 473(7346): 203-207.

Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J and Handelsman J. 2010. Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* 8(4): 251-259.

Allen HK, Moe LA, Rodbumrer J, Gaarder A and Handelsman J. 2009. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J* 3(2): 243-251.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215(3): 403-410.

Amann RI, Ludwig W and Schleifer KH. 1995. Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol Rev* 59(1): 143-169.

Aminov RI and Mackie RI. 2007. Evolution and ecology of antibiotic resistance genes. *FEMS Microbiol Lett* 271(2): 147-161.

Annesley SJ and Fisher PR. 2009. Dictyostelium discoideum-a model for many reasons. *Mol Cell Biochem* 329(1-2): 73-91.

Armbrust EV. 2009. The life of diatoms in the world's oceans. *Nature* 459(7244): 185-192.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WW, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP and Rokhsar DS. 2004. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science* 306(5693): 79-86.

Babson AL, Kawase A and MacCready P. 2006. Seasonal and interannual variability in the circulation of Puget Sound, Washington: A box model study. *Atmos Ocean* 44(1): 29-45.

Baker-Austin C, Wright MS, Stepanauskas R and McArthur JV. 2006. Co-selection of antibiotic and metal resistance. *Trends Microbiol* 14(4): 176-182.

Baquero F. 2012. Metagenomic epidemiology: a public health need for the control of antimicrobial resistance. *Clin Microbiol Infect* 18 Suppl 4: 67-73.

Baquero F, Martinez JL and Canton R. 2008. Antibiotics and antibiotic resistance in water environments. *Curr Opin Biotechnol* 19(3): 260-265.

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Series B Met* 57(1): 289-300.

Berridge MJ, Bootman MD and Roderick HL. 2003. Calcium signalling: Dynamics, homeostasis and remodelling. *Nat Rev Mol Cell Bio* 4(7): 517-529.

Biao X and Yu KJ. 2007. Shrimp farming in China: Operating characteristics, environmental impact and perspectives. *Ocean Coast Manage* 50(7): 538-550.

Binet V, Duthey B, Lecaillon J, Vol C, Quoyer J, Labesse G, Pin JP and Prezeau L. 2007. Common structural requirements for heptahelical domain function in class A and class C G protein-coupled receptors. *J Biol Chem* 282(16): 12154-12163.

Bockaert J and Pin JP. 1999. Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J* 18(7): 1723-1729.

Boerlin P and Reid-Smith RJ. 2008. Antimicrobial resistance: its emergence and transmission. *Anim Health Res Rev* 9(2): 115-126.

Boissonneault KR, Bates SS, Milton S, Pelletier J and Housman DE. 2008. Gene discovery and expression profiling in the toxin (domoic acid)-producing marine diatom Pseudo-nitzschia multiseries (Bacillariophyceae) using cDNA microarrays. *Unpublished*.

Bormann J. 2000. The 'ABC' of GABA receptors. *Trends Pharmacol Sci* 21(1): 16-19.

Bouche N and Fromm H. 2004. GABA in plants: just a metabolite? *Trends Plant Sci* 9(3): 110-115.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, Rayko E, Salamov A, Vandepoele K, Beszteri B, Gruber A, Heijde M, Katinka M, Mock T, Valentin K, Verret F, Berges JA, Brownlee C, Cadoret JP, Chiovitti A, Choi CJ, Coesel S, De Martino A, Detter JC, Durkin C, Falciatore A, Fournet J, Haruta M, Huysman MJ, Jenkins BD, Jiroutova K, Jorgensen RE, Joubert Y, Kaplan A, Kroger N, Kroth PG, La Roche J, Lindquist E, Lommer M, Martin-Jezequel V, Lopez PJ, Lucas S, Mangogna M, McGinnis K, Medlin LK, Montsant A, Oudot-Le Secq MP, Napoli C, Obornik M, Parker MS, Petit JL, Porcel BM, Poulsen N, Robison M, Rychlewski L, Rynearson TA, Schmutz J, Shapiro H, Siaut M, Stanley M, Sussman MR, Taylor AR, Vardi A, von Dassow P, Vyverman W, Willis A, Wyrwicz LS, Rokhsar DS, Weissenbach J, Armbrust EV, Green BR, Van de Peer Y and Grigoriev IV. 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* 456(7219): 239-244.

Bowler C, Allen AE and Vardi A. 2006. An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr Opin Plant Biol* 9(3): 264-273.

Breitbart M. 2012. Marine viruses: truth or dare. *Ann Rev Mar Sci* 4: 425-448.

Bush K, Courvalin P, Dantas G, Davies J, Eisenstein B, Huovinen P, Jacoby GA, Kishony R, Kreiswirth BN, Kutter E, Lerner SA, Levy S, Lewis K, Lomovskaya O, Miller JH, Mobashery S, Piddock LJ, Projan S, Thomas CM, Tomasz A, Tulkens PM, Walsh TR, Watson JD, Witkowski J, Witte W, Wright G, Yeh P and Zgurskaya HI. 2011. Tackling antibiotic resistance. *Nat Rev Microbiol* 9(12): 894-896.

Cabello FC. 2006. Heavy use of prophylactic antibiotics in aquaculture: a growing problem for human and animal health and for the environment. *Environ Microbiol* 8(7): 1137-1144.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

Cheng Z, Tu C, Rodriguez L, Chen TH, Dvorak MM, Margeta M, Gassmann M, Bettler B, Shoback D and Chang W. 2007. Type B gamma-aminobutyric acid receptors modulate the function of the extracellular Ca2+-sensing receptor and cell differentiation in murine growth plate chondrocytes. *Endocrinology* 148(10): 4984-4992.

Claesson MJ, O'Sullivan O, Wang Q, Nikkila J, Marchesi JR, Smidt H, de Vos WM, Ross RP and O'Toole PW. 2009. Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS One* 4(8): e6669.

Clapham DE and Neer EJ. 1997. G protein beta gamma subunits. *Annu Rev Pharmacol Toxicol* 37: 167-203.

Colomer-Lluch M, Jofre J and Muniesa M. 2011. Antibiotic resistance genes in the bacteriophage DNA fraction of environmental samples. *PLoS One* 6(3): e17549.

Cook KL, Rothrock MJ, Jr., Lovanh N, Sorrell JK and Loughrin JH. 2010. Spatial and temporal changes in the microbial community in an anaerobic swine waste treatment lagoon. *Anaerobe* 16(2): 74-82.

King County. 2011. Wastewater treatment process: How is wastewater treated at King County's West Point Treatment Plant? Available: http://www.kingcounty.gov/environment/wtd/About/System/West/Process.aspx. Accessed 22 February 2012.

Crain CM, Halpern BS, Beck MW and Kappel CV. 2009. Understanding and managing human threats to the coastal marine environment. *Ann N Y Acad Sci* 1162: 39-62.

D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, Froese D, Zazula G, Calmels F, Debruyne R, Golding GB, Poinar HN and Wright GD. 2011. Antibiotic resistance is ancient. *Nature* 477(7365): 457-461.

Davies J and Davies D. 2010. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* 74(3): 417-433.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW and Karl DM. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311(5760): 496-503.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA and Rohwer F. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452(7187): 629-632.

Donato JJ, Moe LA, Converse BJ, Smart KD, Berklein FC, McManus PS and Handelsman J. 2010. Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins. *Appl Environ Microbiol* 76(13): 4396-4401.

Dupre DJ, Robitaille M, Rebois RV and Hebert TE. 2009. The role of Gbγ subunits in the organization, assembly, and function of GPCR signaling complexes. *Annu Rev Pharmacol Toxicol* 49: 31-56.

Ecker DJ, Sampath R, Willett P, Wyatt JR, Samant V, Massire C, Hall TA, Hari K, McNeil JA, Buchen-Osmond C and Budowle B. 2005. The Microbial Rosetta Stone Database: a compilation of global and emerging infectious microorganisms and bioterrorist threat agents. *BMC Microbiol* 5: 19.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5): 1792-1797.

El Zein L, Omran H and Bouvagnet P. 2003. Lateralization defects and ciliary dyskinesia: lessons from algae. *Trends Genet* 19(3): 162-167.

Falciatore A, d'Alcala MR, Croot P and Bowler C. 2000. Perception of environmental signal by a marine diatom. *Science* 288(5475): 2363-2366.

Ferreira da Silva M, Tiago I, Verissimo A, Boaventura RA, Nunes OC and Manaia CM. 2006. Antibiotic resistance of enterococci and related bacteria in an urban wastewater treatment plant. *FEMS Microbiol Ecol* 55(2): 322-329.

Fleming LE, Broad K, Clement A, Dewailly E, Elmir S, Knap A, Pomponi SA, Smith S, Solo Gabriele H and Walsh P. 2006. Oceans and human health: Emerging public health risks in the marine environment. *Mar Pollut Bull* 53(10-12): 545-560.

Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO and Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* 337(6098): 1107-1111.

Fountain SJ. 2010. Neurotransmitter receptor homologues of Dictyostelium discoideum. *J Mol Neurosci* 41(2): 263-266.

Fredriksson R, Lagerstrom MC, Lundin LG and Schioth HB. 2003. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology* 63(6): 1256-1272.

Fredriksson R and Schioth HB. 2005. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* 67(5): 1414-1425.

Furushita M, Akagi H, Kaneoka A, Awamura K, Maeda T, Ohta M and Shiba T. 2011. Structural variation of Tn10 that carries tetB found in fish farm bacteria. *Microbes Environ* 26(1): 84-87.

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P and Gerstein MB. 2009. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA* 106(5): 1374-1379.

Gilbert JA and Dupont CL. 2011. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci* 3: 347-371.

Gilbert JA, Steele JA, Caporaso JG, Steinbrueck L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, Somerfield P, Fuhrman JA and Field D. 2012. Defining seasonal marine microbial community dynamics. *ISME J* 6(2): 298-308.

Grundmann H, Klugman KP, Walsh T, Ramon-Pardo P, Sigauque B, Khan W, Laxminarayan R, Heddini A and Stelling J. 2011. A framework for global surveillance of antibiotic resistance. *Drug Resist Updat* 14(2): 79-87.

Halpern BS, Kappel CV, Selkoe KA, Micheli F, Ebert CM, Kontgis C, Crain CM, Martone RG, Shearer C and Teck SJ. 2009. Mapping cumulative human impacts to California Current marine ecosystems. *Conserv Lett* 2(3): 138-148.

Halpern BS, Walbridge S, Selkoe KA, Kappel CV, Micheli F, D'Agrosa C, Bruno JF, Casey KS, Ebert C, Fox HE, Fujita R, Heinemann D, Lenihan HS, Madin EM, Perry MT, Selig ER, Spalding M, Steneck R and Watson R. 2008. A global map of human impact on marine ecosystems. *Science* 319(5865): 948-952.

Handelsman J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* 68(4): 669-685.

Harada H, Nakajima K, Sakaue K and Matsuda Y. 2006. CO2 sensing at ocean surface mediated by cAMP in a marine diatom. *Plant Physiol* 142(3): 1318-1328.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A,

Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T and White R. 2004. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32(Database issue): D258-261.

Heisler J, Glibert PM, Burkholder JM, Anderson DM, Cochlan W, Dennison WC, Dortch Q, Gobler CJ, Heil CA, Humphries E, Lewitus A, Magnien R, Marshall HG, Sellner K, Stockwell DA, Stoecker DK and Suddleson M. 2008. Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* 8(1): 3-13.

Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S, Barry K, Tringe SG, Watson DB, He Z, Hazen TC, Tiedje JM, Rubin EM and Zhou J. 2010. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J* 4(5): 660-672.

Herman E and Challiss RA. 2001. Structural, signalling and regulatory properties of the group 1 metabotropic glutamate receptors: prototypic family C G-protein coupled receptors. *Biochem J* 359: 465-484.

Ho MK, Su Y, Yeung WW and Wong YH. 2009. Regulation of transcription factors by heterotrimeric G proteins. *Curr Mol Pharmacol* 2(1): 19-31.

Hunter DJ. 2005. Gene-environment interactions in human diseases. *Nat Rev Genet* 6(4): 287-298.

Huson DH, Auch AF, Qi J and Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res* 17(3): 377-386.

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL and Armbrust EV. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335(6068): 587-590.

Jacoby GA. 2005. Mechanisms of resistance to quinolones. *Clin Infect Dis* 41 Suppl 2: S120-126.

Jencova V, Strnad H, Chodora Z, Ulbrich P, Vlcek C, Hickey WJ and Paces V. 2008. Nucleotide sequence, organization and characterization of the (halo)aromatic acid catabolic plasmid pA81 from Achromobacter xylosoxidans A8. *Res Microbiol* 159(2): 118-127.

Kelly KM and Chistoserdov AY. 2001. Phylogenetic analysis of the succession of bacterial communities in the Great South Bay (Long Island). *FEMS Microbiol Ecol* 35(1): 85-95.

Kennedy J, Flemer B, Jackson SA, Lejon DP, Morrissey JP, O'Gara F and Dobson AD. 2010. Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. *Mar Drugs* 8(3): 608-628.

Kinnersley AM and Turano FJ. 2000. Gamma aminobutyric acid (GABA) and plant responses to stress. *Crit Rev Plant Sci* 19(6): 479-509.

Kite-Powell HL, Fleming LE, Backer LC, Faustman EM, Hoagland P, Tsuchiya A, Younglove LR, Wilcox BA and Gast RJ. 2008. Linking the oceans to public health: current efforts and future directions. *Environ Health* 7 Suppl 2: S6.

Krell A and Gloeckner G. 2004. Analysis of an osmotic stress induced cDNA library of the psychrophilic diatom Fragilariopsis cylindrus. *Unpublished*.

Kristiansson E, Fick J, Janzon A, Grabic R, Rutgersson C, Weijdegard B, Soderstrom H and Larsson J. 2011. Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS One* 6(2): e17038.

Krogh A, Larsson B, von Heijne G and Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305(3): 567-580.
Kuang D, Yao Y, Maclean D, Wang M, Hampson DR and Chang BS. 2006. Ancestral reconstruction of the ligand-binding pocket of Family C G protein-coupled receptors. *Proc Natl Acad Sci USA* 103(38): 14050-14055.
Lagerstrom MC and Schioth HB. 2008. Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* 7(4): 339-357.
Landry Y and Gies JP. 2008. Drugs and their molecular targets: an updated overview. *Fund Clin Pharmacol* 22(1): 1-18.
Laws EA, Fleming, L.E. and Stegeman, J.J. . 2008. Centers for Oceans and Human Health: contributions to an emerging discipline. *Environ Health* 7(Suppl 2): S1.
Levy SB and Marshall B. 2004. Antibacterial resistance worldwide: causes, challenges and responses. *Nat Med* 10(12 Suppl): S122-129.
Liu B and Pop M. 2009. ARDB-Antibiotic Resistance Genes Database. *Nucleic Acids Res* 37(Database issue): D443-447.
Liu C, Finegold SM, Song Y and Lawson PA. 2008. Reclassification of Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus and Ruminococcus schinkii as Blautia coccoides gen. nov., comb. nov., Blautia hansenii comb. nov., Blautia hydrogenotrophica comb. nov., Blautia luti comb. nov., Blautia producta comb. nov., Blautia schinkii comb. nov. and description of Blautia wexlerae sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol* 58(Pt 8): 1896-1902.
Loomis WF and Anjard C. 2006. GABA induces terminal differentiation of Dictyostelium through a GABA(B) receptor. *Development* 133(11): 2253-2261.
Lozupone CA and Knight R. 2007. Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* 104(27): 11436-11440.
Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWweese-Scott C, Geer LY, Gwadz M, He SQ, Hurwitz DI, Jackson JD, Ke ZX, Lanczycki CJ, Liebert CA, Liu CL, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang DC and Bryant SH. 2005. CDD: a conserved domain database for protein classification. *Nucleic Acids Res* 33: D192-D196.
Marinissen MJ and Gutkind JS. 2001. G-protein-coupled receptors and signaling networks: emerging paradigms. *Trends Pharmacol Sci* 22(7): 368-376.
Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P and Kyrpides NC. 2008. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 36(Database issue): D534-538.
Martinez JL. 2009. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proc Biol Sci* 276(1667): 2521-2530.
McCullagh P and Nelder JA. 1989. Generalized linear models. Boca Raton: Chapman & Hall/CRC. 532 p.
McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB and Paul JH. 2010. High frequency of horizontal gene transfer in the oceans. *Science* 330(6000): 50.

McLellan SL, Huse SM, Mueller-Spitz SR, Andreishcheva EN and Sogin ML. 2010. Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. *Environ Microbiol* 12(2): 378-392.

Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11(1): 31-46.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J and Edwards RA. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.

Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M, Kallas T, Huttlin EL, Cerrina F, Sussman MR and Armbrust EV. 2008. Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. *Proc Natl Acad Sci USA* 105(5): 1579-1584.

Monchy S, Benotmane MA, Janssen P, Vallaeys T, Taghavi S, van der Lelie D and Mergeay M. 2007. Plasmids pMOL28 and pMOL30 of Cupriavidus metallidurans are specialized in the maximal viable response to heavy metals. *J Bacteriol* 189(20): 7417-7425.

Montsant A, Allen AE, Coesel S, De Martino A, Falciatore A, Mangogna M, Siaut M, Heijde M, Jabbari K, Maheswari U, Rayko E, Vardi A, Apt KE, Berges JA, Chiovitti A, Davis AK, Thamatrakoln K, Hadi MZ, Lane TW, Lippmeier JC, Martinez D, Parker MS, Pazour GJ, Saito MA, Rokhsar DS, Armbrust EV and Bowler C. 2007. Identification and comparative genomic analysis of signaling and regulatory components in the diatom Thalassiosira pseudonana. *J Phycol* 43(3): 585-604.

Moore SK, Trainer VL, Mantua NJ, Parker MS, Laws EA, Backer LC and Fleming LE. 2008. Impacts of climate variability and future climate change on harmful algal blooms and human health. *Environ Health* 7 Suppl 2: S4.

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP, Henriksen JR, King GM, Belas R, Fuqua C, Brinkac L, Lewis M, Johri S, Weaver B, Pai G, Eisen JA, Rahe E, Sheldon WM, Ye W, Miller TR, Carlton J, Rasko DA, Paulsen IT, Ren Q, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Rosovitz MJ, Haft DH, Selengut J and Ward N. 2004. Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment. *Nature* 432(7019): 910-913.

Moriyama EN, Strope PK, Opiyo SO, Chen ZY and Jones AM. 2006. Mining the Arabidopsis thaliana genome for highly-divergent seven transmembrane receptors. *Genome Biology* 7(10): R96.

Munir M, Wong K and Xagoraraki I. 2011. Release of antibiotic resistant bacteria and genes in the effluent and biosolids of five wastewater utilities in Michigan. *Water Res* 45(2): 681-693.

Newton RJ, Vandewalle JL, Borchardt MA, Gorelick MH and McLellan SL. 2011. Lachnospiraceae and Bacteroidales alternative fecal indicators reveal chronic human sewage contamination in an urban harbor. *Appl Environ Microbiol* 77(19): 6972-6981.

Nogales B, Aguilo-Ferretjans MM, Martin-Cardona C, Lalucat J and Bosch R. 2007. Bacterial diversity, composition and dynamics in and around recreational coastal areas. *Environ Microbiol* 9(8): 1913-1929.

Nogales B, Lanfranconi MP, Pina-Villalonga JM and Bosch R. 2011. Anthropogenic perturbations in marine microbial communities. *FEMS Microbiol Rev* 35(2): 275-298.

Nordstrom KJV, Almen MS, Edstam MM, Fredriksson R and Schioth HB. 2011. Independent HHsearch, Needleman-Wunsch-Based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. *Mol Biol Evol* 28(9): 2471-2480.

NRC (National Research Council). 1999. From monsoons to microbes: Understanding the oceans role in human health. Washington, D.C.: National Academy Press. 166 p.

NRC (National Research Council). 2007. The new science of metagenomics. Washington D.C.: National Academy Press. 170 p.

Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R and Konstantinidis KT. 2011. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* 77(17): 6000-6011.

Ohmori Y, Hirouchi M, Taguchi J and Kuriyama K. 1990. Functional coupling of the γ-aminobutyric $acid_B$ receptor with calcium ion channel and GTP-binding protein and its alteration following solubilization of the γ-aminobutyric $Acid_B$ receptor. *J Neurochem* 54(1): 80-85.

Ottesen EA, Marin R, Preston CM, Young CR, Ryan JP, Scholin CA and DeLong EF. 2011. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J* 5(12): 1881-1895.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O and Vonstein V. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33(17): 5691-5702.

Parker MS, Mock T and Armbrust EV. 2008. Genomic insights into marine microalgae. *Annu Rev Genet* 42: 619-645.

Parks DH and Beiko RG. 2010. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26(6): 715-721.

Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, Brooks S and Beiko RG. 2009. GenGIS: A geospatial information system for genomic data. *Genome Res* 19(10): 1896-1904.

Partridge SR, Tsafnat G, Coiera E and Iredell JR. 2009. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev* 33(4): 757-784.

Patel PV, Gianoulis TA, Bjornson RD, Yip KY, Engelman DM and Gerstein MB. 2010. Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families. *Genome Res* 20(7): 960-971.

Perkins ND. 2007. Integrating cell-signalling pathways with NF-kappaB and IKK function. *Nat Rev Mol Cell Biol* 8(1): 49-62.

Perovic S, Krasko A, Prokic I, Muller IM and Muller WE. 1999. Origin of neuronal-like receptors in Metazoa: cloning of a metabotropic glutamate/GABA-like receptor from the marine sponge Geodia cydonium. *Cell Tissue Res* 296(2): 395-404.

Petranovic D and Nielsen J. 2008. Can yeast systems biology contribute to the understanding of human disease? *Trends Biotechnol* 26(11): 584-590.

Pettersson E, Lundeberg J and Ahmadian A. 2009. Generations of sequencing technologies. *Genomics* 93(2): 105-111.

Pin JP, Galvez T and Prezeau L. 2003. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol Ther* 98(3): 325-354.
Port JA, Wallace JC, Griffith WC and Faustman EM. 2012. Metagenomic profiling of microbial composition and antibiotic resistance determinants in Puget Sound. *PLoS One* 7(10): e48000.
Pulido OM. 2008. Domoic acid toxicologic pathology: a review. *Mar Drugs* 6(2): 180-219.
Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A and Finn RD. 2012. The Pfam protein families database. *Nucleic Acids Res* 40(Database issue): D290-301.
Qian B, Soyer OS, Neubig RR and Goldstein RA. 2003. Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMMs. *FEBS Lett* 554(1-2): 95-99.
Raes J, Korbel JO, Lercher MJ, von Mering C and Bork P. 2007. Prediction of effective genome size in metagenomic samples. *Genome Biol* 8: R10.
Raiffa H. 1968. Decision analysis. Reading, MA. Addison-Wesley. 309 p.
Riesenfeld CS, Goodman RM and Handelsman J. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* 6(9): 981-989.
Rogers YH and Venter JC. 2005. Genomics: massively parallel sequencing. *Nature* 437(7057): 326-327.
Rosenbaum DM, Rasmussen SG and Kobilka BK. 2009. The structure and function of G-protein-coupled receptors. *Nature* 459(7245): 356-363.
Rosenkilde MM, Waldhoer M, Luttichau HR and Schwartz TW. 2001. Virally encoded 7TM receptors. *Oncogene* 20(13): 1582-1593.
Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M and Venter JC. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3): e77.
Sajjaphan K, Shapir N, Wackett LP, Palmer M, Blackmon B, Tomkins J and Sadowsky MJ. 2004. Arthrobacter aurescens TC1 atrazine catabolism genes trzN, atzB, and atzC are linked on a 160-kilobase region and are functional in Escherichia coli. *Appl Environ Microbiol* 70(7): 4402-4407.
Salyers AA and Amabile-Cuevas CF. 1997. Why are antibiotic resistance genes so resistant to elimination? *Antimicrob Agents Chemother* 41(11): 2321-2325.
Salyers AA, Gupta A and Wang Y. 2004. Human intestinal bacteria as reservoirs for antibiotic resistance genes. *Trends Microbiol* 12(9): 412-416.
Salyers AA, Shoemaker NB, Stevens AM and Li LY. 1995. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiol Rev* 59(4): 579-590.
Sanapareddy N, Hamp TJ, Gonzalez LC, Hilger HA, Fodor AA and Clinton SM. 2009. Molecular diversity of a North Carolina wastewater treatment plant as revealed by pyrosequencing. *Appl Environ Microbiol* 75(6): 1688-1696.
Schluter A, Krahn I, Kollin F, Bonemann G, Stiens M, Szczepanowski R, Schneiker S and Puhler A. 2007. IncP-1-beta plasmid pGNB1 isolated from a bacterial community from a

wastewater treatment plant mediates decolorization of triphenylmethane dyes. *Appl Environ Microbiol* 73(20): 6345-6350.
Schmieder R and Edwards R. 2012. Insights into antibiotic resistance through metagenomic approaches. *Future Microbiol* 7(1): 73-89.
Scholin CA, Gulland F, Doucette GJ, Benson S, Busman M, Chavez FP, Cordaro J, DeLong R, De Vogelaere A, Harvey J, Haulena M, Lefebvre K, Lipscomb T, Loscutoff S, Lowenstine LJ, Marin R, Miller PE, McLellan WA, Moeller PDR, Powell CL, Rowles T, Silvagni P, Silver M, Spraker T, Trainer V and Van Dolah FM. 2000. Mortality of sea lions along the central California coast linked to a toxic diatom bloom. *Nature* 403(6765): 80-84.
Schoneberg T, Schulz A, Biebermann H, Hermsdorf T, Rompler H and Sangkuhl K. 2004. Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacol Ther* 104(3): 173-206.
Seifert R and Wenzel-Seifert K. 2002. Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. *Naunyn Schmiedebergs Arch Pharmacol* 366(5): 381-416.
Shelp BJ, Bown AW and McLean MD. 1999. Metabolism and functions of gamma-aminobutyric acid. *Trends Plant Sci* 4(11): 446-452.
Shokralla S, Spall JL, Gibson JF and Hajibabaei M. 2012. Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 21(8): 1794-1805.
Shrestha RP, Tesson B, Norden-Krichmar T, Federowicz S, Hildebrand M and Allen AE. 2012. Whole transcriptome analysis of the silicon response of the diatom Thalassiosira pseudonana. *BMC Genomics* 13(1): 499.
Soge OO, Meschke JS, No DB and Roberts MC. 2009. Characterization of methicillin-resistant Staphylococcus aureus and methicillin-resistant coagulase-negative Staphylococcus spp. isolated from US West Coast public marine beaches. *J Antimicrob Chemother* 64(6): 1148-1155.
Sommer MO, Dantas G and Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325(5944): 1128-1131.
Spiegel AM and Weinstein LS. 2004. Inherited diseases involving G proteins and G protein-coupled receptors. *Annu Rev Med* 55: 27-39.
Staley C, Gordon KV, Schoen ME and Harwood VJ. 2012. Performance of two quantitative PCR methods for microbial source tracking of human sewage and implications for microbial risk assessment in recreational waters. *Appl Environ Microbiol* 78(20): 7317-7326.
Storteboom H, Arabi M, Davis JG, Crimi B and Pruden A. 2010. Identification of antibiotic-resistance-gene molecular signatures suitable as tracers of pristine river, urban, and agricultural sources. *Environ Sci Technol* 44(6): 1947-1953.
Suenaga H, Koyama Y, Miyakoshi M, Miyazaki R, Yano H, Sota M, Ohtsubo Y, Tsuda M and Miyazaki K. 2009. Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis. *ISME J* 3(12): 1335-1348.
Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J and Wooley J. 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* 39(Database issue): D546-551.
Szczepanowski R, Bekel T, Goesmann A, Krause L, Kromeke H, Kaiser O, Eichler W, Puhler A and Schluter A. 2008. Insight into the plasmid metagenome of wastewater treatment plant

bacteria showing reduced susceptibility to antimicrobial drugs analysed by the 454-pyrosequencing technology. *J Biotechnol* 136(1-2): 54-64.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M and Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10): 2731-2739.
Taniura H, Sanada N, Kuramoto N and Yoneda Y. 2006. A metabotropic glutamate receptor family gene in Dictyostelium discoideum. *J Biol Chem* 281(18): 12336-12343.
Taylor NG, Verner-Jeffreys DW and Baker-Austin C. 2011. Aquatic systems: maintaining, mixing and mobilising antimicrobial resistance? *Trends Ecol Evol* 26(6): 278-284.
Torres-Cortes G, Millan V, Ramirez-Saad HC, Nisa-Martinez R, Toro N and Martinez-Abarca F. 2011. Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. *Environ Microbiol* 13: 1101-1114.
Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P and Rubin EM. 2005. Comparative metagenomics of microbial communities. *Science* 308(5721): 554-557.
Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, Yap J, Yao F, Suan ST, Ing SK, Haynes M, Rohwer F, Wei CL, Tan P, Bristow J, Rubin EM and Ruan Y. 2008. The airborne metagenome in an indoor urban environment. *PLoS One* 3(4): e1862.
Turano FJ, Panta GR, Allard MW and van Berkum P. 2001. The putative glutamate receptors from plants are related to two superfamilies of animal neurotransmitter receptors via distinct evolutionary mechanisms. *Mol Biol Evol* 18(7): 1417-1420.
Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R and Gordon JI. 2007. The human microbiome project. *Nature* 449(7164): 804-810.
Tusnady GE and Simon I. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17(9): 849-850.
UNEP (United Nations Environment Programme). 2006. Marine and coastal ecosystems and human well-being: a synthesis report based on the findings of the Millennium Ecosystem Assessment. UNEP. 76 p.
Uyaguari MI, Fichot EB, Scott GI and Norman RS. 2011. Characterization and quantitation of a novel beta-lactamase gene found in a wastewater treatment facility and the surrounding coastal ecosystem. *Appl Environ Microbiol* 77(23): 8226-8233.
Vardi A, Formiggini F, Casotti R, De Martino A, Ribalet F, Miralto A and Bowler C. 2006. A stress surveillance system based on calcium and nitric oxide in marine diatoms. *Plos Biology* 4(3): 411-419.
von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N and Bork P. 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 315(5815): 1126-1130.
Wang Q, Garrity GM, Tiedje JM and Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73(16): 5261-5267.
Washington State Department of Ecology and Herrera Environmental Consultants I (2010). Phase 3: Loadings of Toxic Chemicals to Puget Sound from POTW discharge of treated wastewater. Olympia, WA. 241 p.
Washington State Department of Health. 2012. 2011 Annual Report: Commercial and recreational shellfish areas in Washington State. Olympia, WA. 22 p.

Whelan S and Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18(5): 691-699.
Whitman WB, Coleman DC and Wiebe WJ. 1998. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 95(12): 6578-6583.
Wright GD. 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol* 5(3): 175-186.
Wright GD. 2010. Antibiotic resistance in the environment: a link to the clinic? *Curr Opin Microbiol* 13(5): 589-594.
Wu CH, Sercu B, Van de Werfhorst LC, Wong J, DeSantis TZ, Brodie EL, Hazen TC, Holden PA and Andersen GL. 2010. Characterization of coastal urban watershed bacterial communities leads to alternative community-based indicators. *PLoS One* 5(6): e11285.
Yang Z, Kumar S and Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141(4): 1641-1650.
Ye L and Zhang T. 2011. Pathogenic bacteria in sewage treatment plants as revealed by 454 pyrosequencing. *Environ Sci Technol* 45(17): 7173-7179.
Zeigler Allen L, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LD, Thiagarajan M, Rusch DB, Nealson KH, Williamson SJ, Venter JC and Allen AE. 2012. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. *ISME J* 6(7): 1403-1414.
Zhang T, Zhang XX and Ye L. 2011. Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. *PLoS One* 6(10): e26041.
Zhang Y, Marrs CF, Simon C and Xi C. 2009. Wastewater treatment contributes to selective increase of antibiotic resistance among Acinetobacter spp. *Sci Total Environ* 407(12): 3702-3706.
Zhou X, Ren L, Meng Q, Li Y, Yu Y and Yu J. 2010. The next-generation sequencing technology and application. *Protein Cell* 1(6): 520-536.
Zhu W, Lomsadze A and Borodovsky M. 2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 38(12): e132.
Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DB, Martiny JB, Sogin M, Boetius A and Ramette A. 2011. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One* 6(9): e24570.
Zou S, Xu W, Zhang R, Tang J, Chen Y and Zhang G. 2011. Occurrence and distribution of antibiotics in coastal water of the Bohai Bay, China: impacts of river discharge and aquaculture activities. *Environ Pollut* 159(10): 2913-2920.